Biophysical studies of RNA:DNA:DNA triplexes and Characterization of riboswitches in cell-free transcription-translation systems

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Biochemie, Chemie und Pharmazie
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von

Jasleen Kaur Bains

aus Dhandowal (Indien)

Frankfurt am Main, 2023

(D30)

Vom Fachbereich Biochemie, Chemie und Pharmazie

der Johann Wolfgang Goethe Universität-Frankfurt am Main als Dissertation angenommen.

Dekan:	Prof. Dr. Clemens Glaubitz
Erster Gutachter:	Prof. Dr. Harald Schwalbe
Zweiter Gutachter:	Prof. Dr. Martin Grininger
Datum der Disputation:	05.05.2023

This thesis was prepared under the supervision of Prof. Dr. Harald Schwalbe between May 2017 and February 2023 at the Institute for Organic Chemistry and Chemical Biology of the Goethe University in Frankfurt am Main, Germany.

To my grandmothers

Table of Content

Summary

RNA research is very important since RNA molecules are involved in various gene regulatory mechanisms as well as pathways of cell physiology and disease development.[1] RNAs have evolved from being considered as carriers of genetic information from DNA to proteins, with the three major types of RNA involved in protein synthesis, including messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA).[2] In addition to the RNAs involved in protein synthesis numerous regulatory non-coding RNAs (ncRNAs) have been discovered in the transcriptome. The regulatory ncRNAs are classified into small ncRNAs (sncRNAs) with transcripts less than 200 nucleotides (nt) and long non-coding RNAs (lncRNAs) with more than 200 nt.[3]

LncRNAs represent the most diverse and versatile class of ncRNAs that can regulate cellular functions of chromatin modification, transcription, and post-transcription through multiple mechanisms.[4] They are involved in the formation of RNA:protein, RNA:RNA and RNA:DNA complexes as part of their gene regulatory mechanism.[4,5] The RNA:DNA interactions can be divided into RNA:DNA heteroduplex formation, also called R-loops, and RNA:DNA:DNA triplex formation. In triplex formation, RNA binds to the major groove of double-stranded DNA through Hoogsteen or reverse Hoogsteen hydrogen bonding, resulting in parallel or anti-parallel triplexes, respectively. *In vitro* studies have confirmed the formation of RNA:DNA:DNA triplexes.[6] However, the extent to which these interactions occur in cells and their effects on cellular function are still not understood, which is why these structures are so exciting to study (**Chapter I RNA:DNA:DNA Triplexes**).

This cumulative thesis investigates several functional and regulatory important RNAs. The first project involves the improved biochemical and biophysical characterization of RNA:DNA:DNA triplex formation between lncRNAs of interest and their target genes. Triplex formation was confirmed by a series of experiments including electromobility shift assays (EMSA), thermal melting assays, circular dichroism (CD), and liquid state nuclear magnetic resonance (NMR) spectroscopy. The following is a summary of the main findings of these publications.

In **research article 5.1**, the oxygen-sensitive *HIF1α-AS1* was identified as a functionally important triplex-forming lncRNA in human endothelial cells using a combination of bioinformatics techniques, RNA/DNA pulldown, and biophysical experiments. Through RNA:DNA:DNA triplex formation, endogenous *HIF1α-AS1* decreases the expression of several genes, including EPH receptor A2 (*EPHA2*) and adrenomedullin (*ADM*), by acting as an adaptor for the repressive human silencing hub (HUSH) complex, which has been studied by our collaborators in the groups of Leisegang and Brandes.

Triplex formation between *HIF1α-AS1* and the target genes *EPHA2* and *ADM* was investigated in biochemical and biophysical studies. The EMSA results indicated that *HIF1α-AS1* forms a low mobility RNA:DNA:DNA triplex complex with the *EPHA2* DNA target sequence. The CD spectrum of the triplex showed distinct features compared to the *EPHA2* DNA duplex and the RNA:DNA heteroduplex. Melting curve analysis revealed a biphasic melting transition for triplexes, with a first melting point corresponding to the dissociation of the RNA strand with melting of the Hoogsteen hydrogen bonds. The second, higher melting temperature corresponds to the melting of stronger Watson-Crick base pairing. Stabilized triplexes were formed using an intramolecular *EPHA2* DNA duplex hairpin construct in which both DNA strands were attached to a 5 nucleotide (nt) thymidine linker. This approach allowed improved triplex formation with lower RNA equivalents and higher melting temperatures. By NMR spectroscopy, the triplex characteristic signals were observed in the ^{1}H NMR spectrum, the imino signals in a spectral region between 9 and 12 ppm resulting from the Hoogsteen base pairing. To elucidate the structural and sequence specific Hoogsteen base pairs 2D ^{1}H,^{1}H-NOESY measurements of the *EPHA2* DNA duplex and the *HIF1α-AS1:EPHA2* triplex were performed. The ^{1}H,^{1}H-NOESY spectrum of the *HIF1α-AS1:EPHA2* triplex with a 10-fold excess of RNA was semi-quantitatively analyzed for changes in the DNA duplex spectrum. We discovered, strong and moderate attenuation of cross peak intensities in the imino region of the NOESY spectrum. This attenuation was proposed to result from weakening of Watson-Crick base pairing by Hoogsteen hydrogen bonding induced by RNA binding. The Hoogsteen interactions can be mapped based on the analysis of the cross peak attenuation in the NOESY spectra, which we used to generate a structural model of the RNA:DNA:DNA triplex. These biophysical results support the physiological function of *HIF1α* as a triplex-forming lncRNA that recruits the HUSH-epigenetic silencing complex to specific target genes such as *EPHA2* and *ADM,* thereby silencing their gene expression through RNA:DNA:DNA triplex formation.

The triplex-forming lncRNA *Fendrr* (fetal-lethal non-coding developmental regulatory RNA) was investigated, which plays a critical role in a variety of developmental processes such as heart and lung development. *Fendrr* has been shown by our collaborators in the Grote group to repress its target genes *Foxf1* and *Pitx2* through interactions with histone modifying complexes, including polycomb repressive complex 2 (PRC2) and the trithorax group/mixed lineage leukemia (TrxG/MLL) protein complexes, by inserting repressive marks at the target genes.[7,8] The triplex formation of the lncRNA *Fendrr* with its previously identified target promoters was confirmed by CD spectroscopy and thermal melting assays (**Research article Ali and Rogala *et al.*, in revisions**).

A novel triplex prediction tool, TriplexAligner, has been developed in the groups of our collaborators Leisegang and Brandes. This tool uses probabilistic nucleotide pairing models based on sequencing data of triplex-forming DNA and RNA and not only on Hoogsteen base pairing rules as used by the previously reported tools *Triplexator* or *Triplex Domain Finder* (**Research article 5.2**).[9] The RNA:DNA:DNA triplexes predicted by *TriplexAligner* were verified *in vitro* by EMSA, CD and melting curve analysis.

The biochemical and biophysical *in vitro* data provided a reliable read-out of RNA:DNA:DNA triplex formation, as triplexes exhibit distinct features compared to their single-stranded DNA and RNA components as well as RNA:DNA heteroduplex motifs. Thus, this set of experiments can be further applied to validate computationally predicted triplex formation.

The second project focuses on the development and further improvement of a cell-free transcription-translation system to investigate transcriptional and translational riboswitches. Cell-free protein expression systems (CFPS) offer several advantages for the study of riboswitches and other regulatory RNAs. These include the decoupling of transcription and translation processes as well as the avoidance of cellular interactions (**Chapter II Cell-free protein synthesis**).[10,11] Riboswitches are RNA-based regulatory elements that use ligand-induced structural changes encoded in the 5'-untranslated region of mRNA to control gene expression. They adapt the expression level of their associated genes in response to the intracellular concentration of specific low-molecular weight ligands. Gene expression is activated or inhibited by the conformational switching of the riboswitch between the ligand-bound and ligand-free states. By regulating fundamental physiological pathways, natural bacterial riboswitches are potential drug targets against bacterial infections (**Chapter III Riboswitches**).[12–14]

First, the cell-free transcription-translation assay was established to investigate the influence of ribosomal S1 (rS1) protein domains on protein expression (**Research article 5.3**). The expression levels were monitored by the shifted green fluorescent protein (sGFP) using cell-extracts containing rS1-depleted ribosomes. The *Vibrio vulnificus* rS1 protein is composed of six homologous oligonucleotide binding domains, of which the first two domains (D1 and D2) are responsible for ribosome binding and domains D3 to D6 bind mRNA, while domain D5 enhances protein chaperone activity. Therefore, the full-length rS1 construct and its truncated mutants with a reduced number of mRNA binding domains were examined. Deletion of the D6 domain did not significantly affect protein expression, whereas deletion of the D5 domain reduced expression levels to 77%. Deletion of the D4 domain restored protein expression to basal levels, and deletion of the entire mRNA binding domain resulted in translation inhibition. Thus, the cell-free experiments demonstrated the importance of the D3-D5 mRNA-binding domains for efficient translation.

In a follow-up project, the dynamic network between adenine-sensing riboswitch (Asw), rS1 and ribosome was investigated by *in vitro* transcription-translation assays as well as NMR spectroscopy and *in vivo* assays (**Research article 5.4**). As a translational on-switch of the adenosine deaminase (*add*) gene, Asw regulates gene expression in an adenine-dependent manner by adopting three distinct conformations. The two ligand-free conformations (apoA, apoB) are in a temperature-dependent pre-equilibrium. Upon adenine binding, the Asw converts to its functional on-state (holo) with an accessible ribosome-binding site, which is required for translation initiation.[15–17]

The coupled transcription-translation assays were performed with the Asw sequence inserted before the sGFP gene to examine adenine- and rS1-dependent expression levels using cell-extracts containing 70S ribosomes and rS1-depleted 70S ribosomes. In the presence of 5 eq rS1, the addition of adenine resulted in a 1.6-fold increase in sGFP expression. However, in the absence of rS1, the switching behavior of Asw remained at basal levels, which could be restored by the addition of rS1. This study demonstrated that the addition of adenine is not sufficient to unwind the secondary structure of Asw to its translational on-state. The chaperone activity of rS1, which is bound to the 30S ribosomal subunit, is important for unwinding the riboswitch expression platform and releasing the ribosome binding site for translation initiation.

We further adapted the fluorescence-based dual read-out *in vitro* transcription-translation system to simultaneously monitor mRNA and protein levels of transcriptional and translational riboswitches, whereas previous riboswitch studies have focused on protein expression (**Research article Bains *et al.*, in revisions**). The thiamine pyrophosphate (TPP)-sensing riboswitches thiM and thiC from *E. coli* and the Asw from *V. vulnificus* were examined for the influence of the ligands TPP and adenine, respectively (Chapter III Riboswitches). The time-dependent mRNA and protein levels were determined using an orthogonal reporter system of the fluorescent Mango-(IV) RNA aptamer and sGFP chromophore. The Mango-(IV) RNA aptamer becomes fluorescent upon binding to the thiazole orange (TO) fluorophore (**Chapter IV Fluorescent RNA aptamers**).[18] The mRNA read-out was improved by designing the Mango-(IV) RNA aptamer as an array construct with four aptamer sequences in a row (M-IVx4). The M-IVx4 array construct showed a 2.1-fold increase in fluorescence intensity compared to the Mango-(IV) RNA aptamer monomer (M-IV). Cell-free results revealed transcriptional and translational off-switch character for both TPP-sensing riboswitches thiM and thiC in the presence of 100 µM TPP. Asw showed a 1.2-fold increase in sGFP expression in the presence of 1 mM adenine. The stabilized apoA mutant showed the highest sGFP expression levels, while the apoB mutant remained in the off-state. The translational on-switch activity of Asw was confirmed as the mRNA levels for Asw, apoA and apoB were not affected by the addition of adenine, only the sGFP expression.

Moreover, the cell-free transcription-translation system with sGFP as a read-out can be used to investigate the translation inhibitory effect of a modified puromycin derivative with two photocleavable protecting groups (**Research article 5.5**). The tyrosyl side chain-like moiety of puromycin was modified to *ortho*-nitrophenylalanine, resulting in deactivation of puromycin by a cyclization reaction upon near-UV irradiation. A t*hio*-coumarylethyl group was inserted as the second photoprotective group, resulting in the modified *thio*-DEACM-nitro-puromycin with controllable activation and deactivation properties of puromycin by wavelength selective irradiation. The active *o*-nitro-puromycin was released upon irradiation of *thio*-DEACM-nitro-puromycin at 488 nm, leading to 80% inhibition of the initial sGFP expression. In addition, the puromycin derivative was deactivated upon irradiation of *thio*-DEACM-nitro-puromycin and *o*-nitro-puromycin at 365 nm and 488/365 nm, respectively, as the photocycled cinnoline puromycin derivative was released. The deactivation of these puromycin derivatives was confirmed by increasing sGFP protein levels that were comparable to those without the addition of inhibitor.

In summary, the author of this thesis developed an *in vitro* transcription-translation assay to enable robust studies of riboswitches and other regulatory RNAs. The dual read-out assay monitors the dynamics of mRNA and protein production in a time-dependent manner. It allows direct monitoring of protein expression kinetics that may be affected by mRNA structure, riboswitch ligand availability and/or mRNA codon usage. The assay allows easy modulation of conditions such as temperature, addition of metabolites or ligands, which can be adjusted to their relative concentrations.

Furthermore, the author of this thesis was part of the COVID19 NMR laboratory team that prepared and purified numerous isotope-labeled RNA and protein samples from SARS-CoV-2 for structural studies by NMR spectroscopy (**Research articles 5.6-5.13**). In addition to secondary structure determination, the druggability of the conserved RNA regulatory elements and proteins was experimentally validated using an NMR-based fragment screening (**Research articles 5.14, 5.15**). Potential binders targeting proteins and the RNA genome of SARS-CoV-2 were identified with binding affinities ranging from micromolar to millimolar. The screening results will provide insights to the medicinal chemistry community to develop new drugs using the fragment-based approaches.

German summary

Die Erforschung von RNAs ist von großer Bedeutung, da RNA-Moleküle an verschiedenen Mechanismen der Genregulation sowie an Signalwegen in der Zellphysiologie und bei der Entstehung von Krankheiten beteiligt sind.[1] RNAs haben sich von Trägern der genetischen Information von der DNA zu Proteinen entwickelt, wobei die drei wichtigsten RNA-Typen an der Proteinsynthese beteiligt sind: Boten-RNA (mRNA), Transfer-RNA (tRNA) und ribosomale RNA (rRNA).[2] Neben den an der Proteinsynthese beteiligten RNAs wurden im Transkriptom zahlreiche regulatorische nicht-kodierende RNAs (ncRNAs) entdeckt. Die regulatorischen ncRNAs werden unterteilt in kurze ncRNAs (sncRNAs) mit Transkripten kürzer als 200 Nukleotide (nt) und lange ncRNAs (lncRNAs) mit Transkripten länger als 200 nt.[3]

LncRNAs stellen die vielfältigste und vielseitigste Klasse von ncRNAs dar, die durch verschiedene Mechanismen zelluläre Funktionen der Chromatinmodifikation, Transkription und Posttranskription regulieren können.[4] Sie sind an der Bildung von RNA:Protein-, RNA:RNA- und RNA:DNA-Komplexen als Teil ihrer Genregulationsmechanismen beteiligt.[4,5] Die RNA:DNA-Interaktionen können in die Bildung von RNA:DNA Heteroduplexen, auch R-Schleifen genannt und die Bildung von RNA:DNA:DNA Triplexen unterteilt werden. Bei der Triplexbildung bindet die RNA über Hoogsteen oder inverse Hoogsteen-Wasserstoffbrücken an die große Furche der doppelsträngigen DNA, was zu parallelen oder anti-parallelen Triplexen führt. *In vitro* Studien haben die Bildung von RNA:DNA:DNA Triplexen bestätigt.[6] Inwieweit diese Wechselwirkungen in Zellen auftreten und sich auf die Zellfunktionen auswirken, ist noch nicht geklärt, weshalb diese Strukturen für die Forschung von großem Interesse sind (**Kapitel I RNA:DNA:DNA Triplexe**).

In dieser kumulativen Dissertation werden verschiedene funktionell und regulatorisch wichtige RNAs untersucht. Das erste Projekt umfasst die verbesserte biochemische und biophysikalische Charakterisierung von RNA:DNA:DNA Triplexen. Die Triplexbildung wurde durch eine Reihe von Experimenten bestätigt, darunter Gelelektrophorese Untersuchungen (electromobility shift assay, EMSA), thermische Schmelzanalyse, Circulardichroismus (CD) und Kernspinresonanzspektroskopie (NMR). In den folgenden Abschnitten werden die wichtigsten Ergebnisse dieser Veröffentlichungen zusammengefasst.

In der **Publikation 5.1** wurde das sauerstoffsensitive *HIF1α-AS1* als funktionell wichtige Triplex-bildende lncRNA in humanen Endothelzellen durch eine Kombination von bioinformatischen Techniken, RNA/DNA-Pulldown und biophysikalischen Experimenten identifiziert. Durch RNA:DNA:DNA Triplexbildung reduziert endogenes *HIF1α-AS1* die Expression mehrerer Gene, einschließlich des EPH- Rezeptors 2 (*EPHA2*) und Adrenomedullin (*ADM*), indem es als Adaptor für den repressiven human Silencing Hub (HUSH)-Komplex fungiert, der von den Arbeitsgruppen Leisegang und Brandes untersucht wurde.

Die Triplexbildung zwischen *HIF1α-AS1* und den Zielgenen *EPHA2* und *ADM* wurde biochemisch und biophysikalisch untersucht. Die EMSA-Ergebnisse zeigten, dass *HIF1α-AS1* einen RNA:DNA:DNA Triplex-Komplex mit reduzierter Mobilität mit der *EPHA2* DNA-Zielsequenz bildet. Das CD-Spektrum des Triplexes zeigte deutliche Unterschiede im Vergleich zur *EPHA2* Duplex DNA und zum RNA:DNA Heteroduplex. Die Schmelzkurvenanalyse ergab einen zweiphasigen Schmelzübergang für Triplexe, wobei der erste Schmelzpunkt der Dissoziation des RNA-Stranges mit dem Schmelzen der Hoogsteen-Wasserstoffbrücken entspricht. Der zweite, höhere Schmelzpunkt entspricht der stärkeren Watson-Crick Basenpaarung. Stabilisierte Triplexe wurden mit Hilfe eines *EPHA2* intramolekularen Haarnadel DNA Duplex Konstrukts gebildet, bei dem beide DNA-Stränge an einen 5 Nukleotide langen Thymidin-Linker gebunden waren. Dieser Ansatz ermöglichte eine verbesserte Triplexbildung mit geringeren RNA-Äquivalenten und die höheren Schmelztemperaturen. Mittels NMR-Spektroskopie wurden charakteristische Triplex Signale im ^{1}H-NMR Spektrum beobachtet, wobei die Imino Signale im Spektralbereich zwischen 9 und 12 ppm lagen, was auf die Hoogsteen Basenpaarung zurückzuführen ist. Zur Aufklärung der strukturellen und sequenzspezifischen Hoogsteen-Basenpaare wurden 2D-1H,1H-NOESY Messungen des EPHA2 DNA Duplexes und des *HIF1α-AS1:EPHA2* Triplexes im 10-fachen Überschuss durchgeführt. Im Imino-Bereich des NOESY Spektrums wurde eine starke und eine mäßige Abschwächung der Intensitäten der Kreuzpeaks beobachtet. Als Ursache wurde die Abschwächung der Watson-Crick Basenpaarung durch die Hoogsteen Bindung vermutet, die durch die RNA-Bindung induziert wird. Die Hoogsteen Wechselwirkungen können durch die Analyse der Abschwächung der Kreuzpeaks in den NOESY Spektren dargestellt werden, was wiederum zur Erstellung eines Strukturmodells des RNA:DNA:DNA Triplexes verwendet wurde. Diese biophysikalischen Ergebnisse unterstützen die physiologische Funktion von *HIF1α-AS1* als Triplex-bildende lncRNA, die den HUSH-epigenetischen Suppressor-Komplex an spezifische Zielgene wie *EPHA2* und *ADM* rekrutiert und dadurch deren Genexpression durch RNA:DNA:DNA Triplexbildung stilllegt.

Die zweite untersuchte lncRNA, *Fendrr* (fetal-lethal non-coding developmental regulatory RNA), spielt eine entscheidende Rolle bei einer Vielzahl von Entwicklungsprozessen wie der Herz- und Lungenentwicklung. Es konnte gezeigt werden, dass *Fendrr* seine Zielgene *Foxf1* und *Pitx2* durch Interaktion mit Histon-modifizierten Komplexen wie dem Polycomb Repressive Complex 2 (PRC2) und dem Trithorax Group/Mixed Lineage Leukemia (TrxG/MLL) Proteinkomplex unterdrückt, indem es repressive Markierungen in die Zielgene einbaut.[7,8] Die Triplexbildung der lncRNA *Fendrr* mit ihren zuvor identifizierten Zielpromotoren wurde durch CD-Spektroskopie und thermische Schmelzkurven bestätigt (**Publikation Ali und Rogala *et al.*, in Revision**).

Ein neuartiges Tool zur Triplex-Vorhersage, *TriplexAligner*, wurde von den Arbeitsgruppen Leisegang und Brandes entwickelt. Dieses Tool verwendet probabilistische Nukleotidpaarungsmodelle, die auf Sequenzierungsdaten von DNA und RNA, die Triplexe bilden, basieren, und nicht nur auf Hoogsteen

Basenpaarungsregeln, wie sie von den zuvor vorgestellten Tools wie *Triplexator* und *Triplex Domain Finder* verwendet werden (**Publikation 5.2**).[9] Die von *TriplexAligner* vorhergesagten RNA:DNA:DNA Triplexe wurden *in vitro* durch EMSA, CD und Schmelzkurvenanalyse verifiziert.

Die biochemischen und biophysikalischen *in vitro* Daten lieferten ein zuverlässiges Maß für die Bildung von RNA:DNA:DNA Triplexen, da Triplexe im Vergleich zu ihren einzelsträngigen DNA- und RNA-Komponenten sowie zu RNA:DNA Heteroduplexmotiven unterschiedliche Eigenschaften aufweisen. Daher können diese Experimente zur Validierung von bioinformatisch vorhergesagten Triplexen verwendet werden.

Das zweite Projekt zielt auf die Entwicklung und weitere Verbesserung eines zellfreien Transkriptions-Translations-Systems zur Untersuchung von transkriptionellen und translationellen Riboschaltern. Zellfreie Protein-Expressionsysteme (cell-free protein expression systems – CFPS) bieten mehrere Vorteile für die Untersuchung von Riboschaltern und anderen regulatorischen RNAs. Dazu gehören die Entkopplung von Transkriptions- und Translationsprozessen sowie die Vermeidung zellulärer Interaktionen (**Kapitel II Cell-free protein synthesis**).[10,11] Riboschalter sind RNA-regulatorische Elemente, die ligandeninduzierte strukturelle Veränderungen in der 5'-untranslatierten Region der mRNA nutzen, um die Genexpression zu steuern. Sie passen die Expression ihrer assoziierten Gene als Reaktion auf die intrazelluläre Konzentration spezifischer niedermolekularer Liganden an. Die Genexpression wird durch den Konformationswechsel des Riboschalters zwischen ligandengebundenem und ligandenfreiem Zustand aktiviert oder inhibiert. Natürliche bakterielle Riboschalter sind potenzielle Angriffspunkte für Medikamente gegen bakterielle Infektionen, da sie grundlegende physiologische Prozesse regulieren (**Kapitel III Riboswitches**).[12–14]

Um den Einfluss der ribosomalen S1 (rS1)-Proteindomänen auf die Proteinexpression zu untersuchen, wurde zunächst der zellfreie Transkriptions-Translations-Assay etabliert (**Publikation 5.3**). Die Expressionslevel wurden mit Hilfe eines verschobenen grün fluoreszierenden Proteins (sGFP) unter Verwendung von Zellextrakten mit rS1-depletierten Ribosomen verfolgt. Das rS1-Protein von *Vibrio vulnificus* besteht aus homologen Oligonukleotid-Bindungsdomänen. Die ersten beiden Domänen (D1 und D2) sind für die Ribosomenbindung verantwortlich, die Domänen D3 bis D6 binden an die mRNA, während die Domäne D5 die Chaperonaktivität des Proteins verstärkt. Daher wurden sowohl das rS1-Konstrukt in voller Länge als auch seine verkürzten Mutanten, denen die mRNA-Bindungsdomänen fehlen, untersucht. Die Expression des Proteins wurde durch die Deletion der D6-Domäne nicht signifikant beeinflusst, während die Deletion der D5-Domäne die Expression auf 77% reduzierte. Die Deletion der D4-Domäne stellte die basale Expression wieder her, und die Deletion der gesamten mRNA-Bindungsdomäne führte zu einer Translationsinhibierung. Die Bedeutung der mRNA-Bindungsdomänen D3-D5 für eine effiziente Translation konnte somit in zellfreien Experimenten nachgewiesen werden.

In einem Folgeprojekt wurde das dynamische Netzwerk zwischen Asw, rS1 und Ribosom mit Hilfe von *in vitro* Transkriptions-Translations-Assays sowie durch NMR-Spektroskopie und *in vivo* Assays untersucht (**Publikation 5.4**). Als Translationsschalter des Adenosindeaminase (*add*)-Gens reguliert Asw die Genexpression adeninabhängig, indem es drei verschiedene Konformationen annimmt. Die beiden ligandenfreien Konformationen (apoA, apoB) befinden sich in einem temperaturabhängigen Präequilibrium. Nach der Adeninbindung wechselt Asw in den funktionellen on-Zustand (holo) mit einer zugänglichen Ribosomen-Bindungsstelle, die für die Translationsinitiierung notwendig ist.[15–17] Gekoppelte Transkriptions-Translations-Assays wurden mit der vor dem sGFP-Gen eingefügten Asw-Sequenz durchgeführt, um die Adenin- und rS1-abhängigen Expressionsniveaus zu untersuchen. Es wurden Zellextrakte verwendet, die 70S-Ribosomen und rS1-depletierte 70S-Ribosomen enthielten. Die Zugabe von Adenin führte in Gegenwart von 5 Äquivalenten rS1 zu einer 1,6-fachen Steigerung der sGFP-Expression. In Abwesenheit von rS1 blieb das Asw-Schaltverhalten jedoch auf dem Ausgangsniveau, das durch die Zugabe von rS1 wiederhergestellt wurde. Die Studie zeigt, dass die Zugabe von Adenin nicht ausreicht, um die Sekundärstruktur von Asw in den translatorischen An-Zustand zu überführen. Die Chaperonaktivität von rS1, das an die 30S ribosomale Untereinheit gebunden ist, ist wichtig für die Auslösung der Riboschalter-Expressionsplattform und die Freisetzung der ribosomalen Bindungsstelle für die Translationsinitiierung.

Das duale Fluoreszenz-basierte *in vitro* Transkriptions-Translations-System wurde weiter angepasst, um gleichzeitig die mRNA- und Proteinlevel von transkriptionellen und translationellen Riboschaltern zu überwachen, während sich frühere Studien über Riboschalter auf die Proteinexpression konzentrierten (**Publikation Bains *et al.*, in Revision**). Der Einfluss der Liganden Thiaminpyrophosphat (TPP) und Adenin auf die TPP-sensitiven Riboschalter thiM und thiC aus *E. coli* und Asw aus *V. vulnificus* wurde untersucht (Kapitel III Riboswitches). Die zeitabhängigen mRNA- und Proteinlevels wurden mit einem orthogonalen Reportersystem untersucht. Das System besteht aus einem fluoreszierenden Mango-(IV) RNA-Aptamer und einem sGFP-Chromophor. Durch die Bindung an das Fluorophor Thiazolorange (TO) wird das RNA-Aptamer fluoreszent (Kapitel IV Fluorescent RNA Aptamers).[18] Das Auslesen der mRNA wurde durch das Design des Mango-(IV) RNA-Aptamers als Array-Konstrukt mit vier in Reihe geschalteten Aptamersequenzen (M-IVx4) verbessert. Das M-IVx4 Array Konstrukt zeigte eine 2,1-fache Steigerung der Fluoreszenzintensität im Vergleich zum Mango-(IV) RNA-Aptamer Monomer Konstrukt (M-IV). Die Ergebnisse des zellfreien Transkriptions-Translations-Assays zeigten, dass die beiden TPP-sensitiven Riboschalter thiM und thiC in Gegenwart von 100 µM TPP sowohl auf transkriptioneller als auch auf translationaler Ebene inaktiviert werden können. Die sGFP-Expression von Asw wurde in Gegenwart von

1 mM Adenin um das 1,2-fache erhöht. Dabei zeigte die stabilisierte apoA-Mutante die höchste sGFP-Expression, während die apoB-Mutante in einem „off"-Zustand blieb. Die translatorische „On-Switch" Aktivität von Asw konnte bestätigt werden, da die mRNA Levels für ASW, apoA und apoB durch die Zugabe von Adenin nicht beeinflusst wurden, sondern nur die sGFP-Expression.

Des Weiteren kann das zellfreie Transkriptions-Translations-System mit sGFP als Indikator verwendet werden, um die translations-inhibierende Wirkung eines modifizierten Puromycin-Derivats mit zwei photospaltbaren Schutzgruppen zu untersuchen (**Publikation 5.5**). Die Tyrosyl-Seitenkette von Puromycin wurde zu *ortho*-Nitrophenylalanin modifiziert. Dies führte zu einer Deaktivierung des Puromycins durch eine Zyklisierungsreaktion bei Bestrahlung mit Licht im nahen UV-Bereich. Als zweite photoreaktive Gruppe wurde eine Thio-Coumarylethyl-Gruppe eingeführt. Dadurch entstand das modifizierte Thio-DEACM-Nitro-Puromycin mit kontrollierbaren Aktivierungs- und Deaktivierungseigenschaften des Puromycins durch wellenlängenselektive Bestrahlung. Das aktive *o*-Nitro-Puromycin wurde durch Bestrahlung des Thio-DEACM-Nitro-Puromycins mit 488 nm freigesetzt und inhibierte 80% der ursprünglichen sGFP-Expression. Zusätzlich wurde das Puromycin-Derivat bei Bestrahlung von Thio-DEACM-Nitro-Puromycin und *o*-Nitro-Puromycin bei 365 nm bzw. 488/365 nm durch Freisetzung des photocyclischen Cinnolin-Puromycin-Derivats inaktiviert. Die Inaktivierung dieser Puromycin-Derivate wurde durch hohe sGFP-Werte bestätigt, die mit denen ohne Inhibitorzusatz vergleichbar waren.

Zusammenfassend ist festzuhalten, dass die Autorin dieser Dissertation einen *in vitro* Transkriptions-Translations-Assay entwickelt hat, der eine robuste Untersuchung von Riboschaltern und anderen regulatorischen RNAs ermöglicht. Mit dem dualen Fluoreszenz-basierten Assay kann die Dynamik der mRNA- und Proteinproduktion als Funktion der Zeit verfolgt werden. Dies ermöglicht eine direkte Überwachung der Proteinexpressionskinetik, die durch die mRNA-Struktur, die Verfügbarkeit von Riboschalter-Liganden und/oder die mRNA-Codon-Nutzung beeinflusst werden kann. Der Assay erlaubt eine einfache Anpassung der Bedingungen wie Temperatur, Zugabe von Metaboliten oder Liganden, deren relative Konzentration eingestellt werden kann.

Zudem war die Autorin dieser Dissertation Teil des COVID19 NMR-Laborteams, das zahlreiche isotopenmarkierte RNA- und Proteinproben von SARS-CoV-2 hergestellt und aufgereinigt hat für Strukturuntersuchungen mittels NMR-Spektroskopie (**Publikationen 5.6-5.13**). Zusätzlich wurde die Wirksamkeit der konservierten RNA-Regulationselemente und Proteine durch NMR-basiertes Fragment-Screening experimentell validiert (**Publikationen 5.14, 5.15**). Potenzielle Binder für Proteine und das RNA-Genom von SARS-CoV-2 wurden mit Bindungsaffinitäten im mikromolaren bis millimolaren Bereich identifiziert. Die Ergebnisse des Screenings werden der medizinischen Chemie Hinweise für die Entwicklung neuer Medikamente mit Fragment-basierten Ansätzen liefern.

List of Abbreviations

ASW	adenine-sensing riboswitch
2’-AE	2’-O-aminoethyl
2’-OMe	2’-O-methyl
2-AP	2-aminopurine
6-thioG	2’-deoxy-6-thioguanosine
7-dzaX	7-deaza-2’-deoxyxanthosine
A	adenine
add	adenosine deaminase
ADM	adrenomedullin
ADP	adenosine-5'-diphosphate
Ago	Argonaute protein
B. subtillis	*Bacillus subtillis*
C	cytosine
CD	circular dichroism
c-di-AMP	cyclic di-adenosine monophosphate
c-di-GMP	cyclic di-guanosine monophosphate
cDNA	complementary DNA
CECF	continuous exchange cell-free
ceRNA	competing endogenous RNA
CFCF	continuous flow cell-free
CFPS	cell-free protein synthesis
CHIP	Chromatin immunoprecipitation
DFHBI	3,5-difluoro-4-hydroxybenzylidene imidazolinone
DHFR	dihydrofolate reductase
DMHBI	3,5-dimethoxy-4-hydroxybenzylidene imidazolinone
DNA	deoxyribonucleic acid
DNMT	DNA methyltransferase
dsDNA	double-stranded DNA
dsRNA	double-stranded RNA
E. coli	*Escherichia coli*
EF	elongation factor
eGFP	enhanced green fluorescent protein
EMSA	electromobility shift assay
ENA	2’-O,4’-C-ethylene-linked nucleic acid
EPHA2	ephrin receptor A2
ESI-MS	electrospray ionization mass spectrometry
FACS	fluorescence-activated cell sorting
FANTOM	functional annotation of the mammalian genome
Fendrr	fetal-lethal non-coding developmental regulatory RNA
G	guanine
GFP	green fluorescent protein
GSW	guanine-sensing riboswitch
GTP	guanosine-5'-triphosphate
H3K27me3	trimethylation of lysine 27 on histone 3
H3K4me3	trimethylation of lysine 4 on histone H2
HIF1A	hypoxia-inducible factor 1-alpha
HIF1α-AS1	hypoxia-inducible factor 1α antisense 1
HMP	4-amino-5-hydroxymethyl-2-methylpyrimidine
HUSH	human silencing hub
IF	initiation factor
IRES	internal ribosomal entry site
IVEC	*in vitro* expression cloning
L	liter
LNA	locked nucleic acid
lncRNA	long non-coding RNA
LSD1	lysine-specific demethylase 1
MALAT1	metastasis-associated lung adenocarcinoma transcript 1
MAT2A	methionine adenosyltransferase 2A
mC	5-methylcytosine
MEG3	maternally expressed gene 3
miRNA	microRNA
MLL	mixed lineage leukemia
MP	membrane protein
MPP8	M-phase phosphoprotein 8
mRNA	messenger ribonucleic acid
N^6-8-oxo-dA	N^6-methyl-8-oxo-2’-deoxyadenosine
ncRNA	non-coding RNA
NMR	nuclear magnetic resonance
PARTICLE	promoter of MAT2A-antisense radiation-induced circulating ncRNA
PCR	polymerase chain reaction
piRNA	piwi-associated RNA
PNA	peptide nucleic acid
PRC2	polycomb repressive complex
preQ$_1$	7-aminomethyl-7-deazaguanine riboswitch
PURE	protein synthesis using recombinant elements
RBS	ribosome-binding site
RF	release factor
RIP	RNA immunoprecipitation
RISC	RNA-induced silencing complex
RNA	ribonucleic acid
RNAP II	RNA polymerase II
RNase T1	ribonuclease T1
RNP	ribonucleoprotein

RP-HPLC	reversed-phase high-performance liquid chromatography
RRF	ribosome recycling factor
rRNA	ribosomal RNA
SAM	S-adenosylmethionine
SAXS	small-angle X-ray scattering
SD	Shine-Dalgarno
SELEX	systematic evolution of ligands by exponential enrichment
SETDB1	SET domain bifurcated histone lysine methyltransferase 1
sGFP	shifted green fluorescent protein
SHAPE	2'-hydroxyl acylation using primer extension
siRNA	small interfering RNA
smFRET	single-molecule Förster resonance energy transfer
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
SPHK1	sphingosine kinase 1
SPHK1-B	*SPHK1* isoform B
ssDNA	single-stranded DNA
ssRNA	single-stranded RNA
T	thymine
TER	telomerase RNA subunit
TERT	telomerase reverse-transcriptase
TFIIB	transcription factor IIB
TGF-β	transforming growth factor-β
TO	thiazole orange
TO1-biotin	thiazole orange 1-biotin
TPP	thiamine pyrophosphate
TPP	thiamine pyrophosphate
tRNA	transfer ribonucleic acid
TrxG	trithorax group
TTS	triplex-target site
U	uracil
UAA	unnatural amino acid
UTR	untranslated region
UV	ultraviolet
V. vulnificus	*Vibrio vulnificus*
Xist	X-inactive-specific transcript

CHAPTER I

RNA:DNA:DNA Triplexes

1.1 RNA overview

The cellular identity, development, proper function, and response to environmental changes of complex organisms are based on the regulation of gene expression because each cell contains identical copies of the genome. The genome is transcribed into numerous ribonucleic acids (RNAs) that vary in size, abundance, and protein-coding ability.[19] According to the central dogma of molecular biology, RNA is considered as an intermediate molecule in the transcription of deoxyribonucleic acid (DNA) into messenger RNA (mRNA), which is further translated into protein.[20] RNA is not only involved in the transfer of genetic information, but is also known to have various regulatory, structural, and enzymatic functions. Sequence mutations and dysfunctions of RNAs or ribonucleoproteins (RNPs) complexes are the cause of various diseases such as autoimmune diseases,[21,22] cancer,[23–25] cardiovascular diseases,[23,26] diabetes,[27,28] obesity,[27,28] neuromuscular and neurodegenerative disorders.[1,29,30] Therefore, understanding the role of RNAs involved in gene regulatory mechanisms is crucial for the development of RNA-based therapeutics.[1]

RNA and DNA are composed of four basic building blocks, including two purine nucleobases, adenine (A) and guanine (G), and the pyrimidine nucleobases, cytosine (C), uracil (U) in RNA, and thymine (T) in DNA, which possess a sugar moiety and are linked by phosphodiester bonds (Figure 1A).[31] In 1953, Watson and Crick, along with Rosalind Franklin's X-ray diffraction results, revealed the structure of the DNA double helix. Interactions between complementary nucleobases within the two anti-parallel DNA strands are mediated by hydrogen bonds, which have been characterized as the canonical G-C and A-T Watson-Crick base pairs, and as A-U in RNA strands (Figure 1B).[32]

Structural diversity of RNA is much greater, mainly due to the fact that RNA is transcribed as a single stand, allowing RNA motifs to form through intra- or intermolecular interactions, whereas DNA strands form double helical structures.[33] Leontis and Westhof categorized RNA base pairs based on the interaction sites for hydrogen bonding of the nucleobases, which are divided into three edges, namely the Watson-Crick, sugar, and Hoogsteen or C-H edge of purines and pyrimidines, respectively. A total of 12 geometric families have been reported with respect to the orientation of the glycosidic bonds attached to the bases in either *cis* or *trans* configuration.[34,35] The G·U wobble base pair has been identified in numerous functional RNAs such as ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and ribozymes (Figure 1C).[36] Ladner *et al.* solved the G·U wobble base pair structure at atomic resolution in yeast tRNA$^{\text{Phe}}$.[36,37] Another example of *cis* Watson-Crick/Watson-Crick base pairs is the hemi-protonated C·C^{+} base pair that was observed in the catalytic pocket of the hammerhead ribozyme (Figure 1C).[38] G-rich sequences can interact as G·G noncanonical base pairs to form G-quadruplex structures in the presence of monovalent cations (Figure 1D).[39,40] In addition, the *cis* Watson-Crick/Hoogsteen interaction is important for base triplet formation (shown in Figure 3).[35] The metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) lncRNA can form an intramolecular RNA triple helix using U·A non-canonical base pairs (Figure 1D).[41,42]

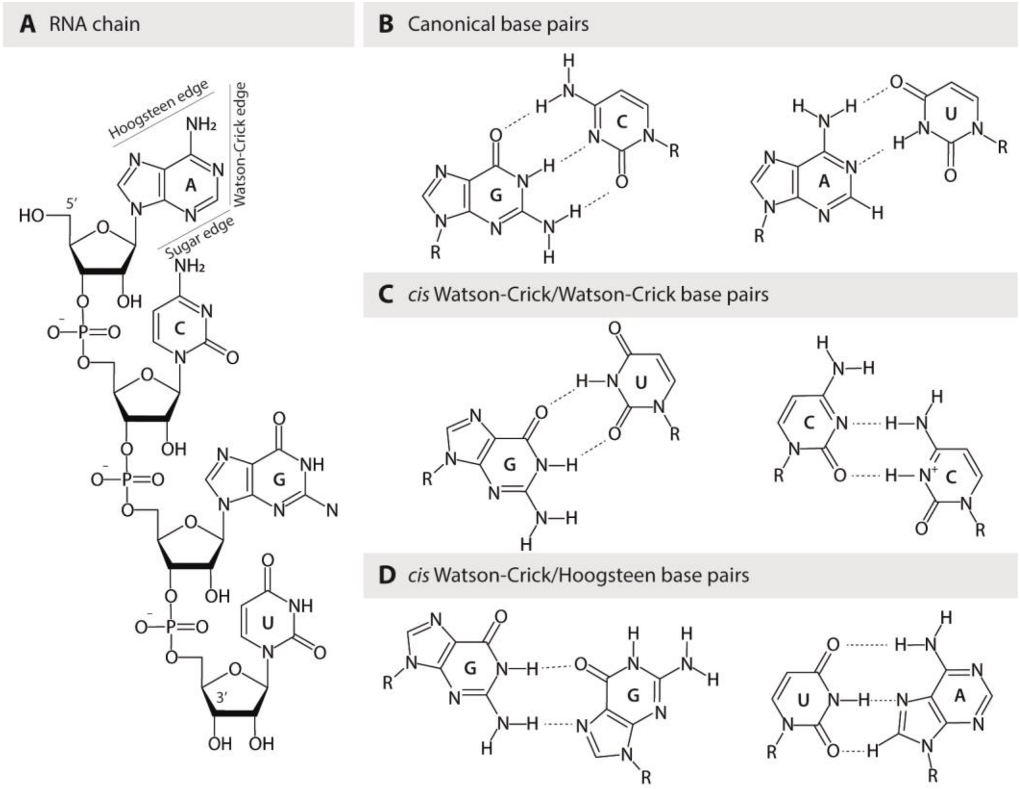

A RNA chain
Hoogsteen edge
Watson-Crick edge
Sugar edge
5′
3′
B Canonical base pairs
C cis Watson-Crick/Watson-Crick base pairs
D cis Watson-Crick/Hoogsteen base pairs

Figure 1. Overview of base pairs that are present in RNA structures. (A) Chemical structure of an RNA chain showing the nucleobases that are attached to the sugar moiety and the phosphodiester bonds. (B) Chemical structures of the canonical G-C and A-U base pairs. (C) Chemical structures of the *cis* Watson-Crick/Watson-Crick base pairs including the G·U and C·C$^+$ base pairs. (D) Chemical structures of the *cis* Watson-Crick/Hoogsteen base pairs including the G·G and U·A base pairs.

The interaction between various molecules and RNA occurs due to its ability to form secondary structures by base pairing between complementary nucleobases, resulting in shielding or exposure of the recognized regions within the RNA.[2] Double-helical RNA adopts the A-form helix, other non-helical RNA secondary structures are based on the different base pairing and base stacking patterns, resulting in structural motifs such as double helices, hairpins, bulges, internal loops, and junctions. In particular, the three-dimensional structural organization of RNA elements specifies the regulatory functions described by base pairing, including canonical Watson-Crick, wobble, and non-canonical base pairs, as well as long-range intramolecular interactions.[43]

1.2 Small non-coding RNAs

Non-coding RNAs (ncRNAs) are key regulators of gene expression in humans and many other organisms at the levels of transcription, RNA processing, and translation. Analysis of the human genome revealed that 73% of the transcribed genes are non-coding and only 2% of the genome is translated into functional proteins.[44,45]

Small ncRNAs are divided into housekeeping and regulatory ncRNAs. Housekeeping RNAs include small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), tRNAs, and rRNAs.[19] The snRNA class has been identified in the nucleus of eukaryotic cells and controls processes such as mRNA splicing, telomere maintenance, and RNA polymerase II transcription factor regulation.[43,46] Chemical modifications such as pseudo-uridylation and 2’-*O*-methylation of rRNAs and tRNAs are introduced by snoRNAs in a complex with RNA-binding proteins as ribonucleoproteins (snoRNPs).[47–49] Both tRNAs and rRNAs play a central role in the translation of mRNA into peptides and proteins.[50] During translation initiation, the translation start site is determined by the binding of the 3’-end of the 16S rRNA in the small 30S ribosomal subunit to the initiator codons in the 5’-end of the mRNA.[51]

MicroRNAs (miRNAs), small interfering RNAs (siRNAs) and piwi-associated RNAs (piRNAs) have been identified as small regulatory ncRNAs. Approximately 22 nucleotides (nt) long single-stranded miRNAs control gene expression at the post-transcriptional level by binding to the complementary mRNA, resulting in translation inhibition or mRNA degradation and gene silencing.[7,52] Specifically, miRNAs have been described as regulators of endogenous genes, whereas siRNAs maintain genome integrity in response to viral infection by invading exogenous genes. Both miRNAs and siRNAs are processed by the enzyme Dicer from their double-stranded RNA (dsRNA) precursors into short, approximately 20-30 nt RNA fragments of incomplete dsRNA stem-loop precursors and perfectly base paired dsRNAs, respectively.[53] In addition, the processed RNA strand interacts with Argonaute (Ago) proteins to form the RNA-induced silencing complex (RISC), which allows the target RNA recognition by Watson-Crick base pairing to enable the RNA silencing mechanisms of miRNAs and siRNAs.[54] Another class of small ncRNAs are piRNAs. These are approximately 27 nt long, bind to Ago-related piwi proteins, and are generated in the germline of many animals. They have been extensively studied in *Drosophila,* where regulation of gene expression by transcriptional silencing of transposons and the control of somatic development have been described.[55–57] In particular, these processes have been described in plants and animals to confer transcriptional resistance to parasitic invasion.[2]

1.3 Long non-coding RNAs

The non-coding genome can be divided into short non-coding RNAs (such as miRNAs) with a length of less than 200 nt and long non-coding RNAs (lncRNAs) with a length of more than 200 nt. In the early 2000s, the large-scale sequencing project based on complementary DNA (cDNA) sequencing of the mouse genome by the Functional Annotation of the Mammalian Genome (FANTOM) consortium discovered

11,665 non-coding RNAs as a major component of the mouse transcriptome compared to the 4,258 small amount of the mouse genome transcribed into protein-coding mRNAs.[58] The subsequent FANTOM3 consortium project published 34,030 lncRNAs using cDNA sequencing.[59] More recently, the GENCODE consortium annotated 19,379 protein-coding and 19,933 lncRNA genes of the human genome (GENCODE V42, January 2023).[60]

LncRNAs are classified based on their genomic proximity to protein-coding genes, and are described as overlapping, sense, *cis*-antisense, *trans*-spliced, intronic, bidirectional, and intergenic lncRNAs.[61] Similar to mRNAs, many lncRNAs are transcribed by the RNA polymerase II, including 5′-end capping, and 3′-end polyadenylation. Due to low sequence conservation, lncRNAs were thought to be non-functional. In fact, lncRNAs are expressed in a cell type-specific manner and contain shorter conserved sequences with functional structures and domains that regulate processes of chromatin modification, transcription, and post-transcription. LncRNAs have been shown to bind to DNA, RNA, and proteins. Interactions occur between single-stranded DNA (ssDNA), single-stranded RNA (ssRNA), and double-stranded DNA (dsDNA) to form motifs such as RNA:DNA heteroduplexes, also called R-loops, and RNA:DNA:DNA triplexes (Figure 2A).[4]

The molecular functions of lncRNAs have been classified into four archetypes as molecular signals, decoys, guides, and scaffolds (Figure 2B).[62] LncRNA expression regulates transcription as molecular signals at the levels of initiation, elongation, and termination (Figure 2B). Many processes are regulated by the lncRNAs at the post-transcriptional level, including mRNA processing, splicing, editing, transport, translation, and degradation (Figure 2C).[4] Epigenetic regulation of lncRNAs occurs through recruitment of chromatin-modifying complexes to DNA, which can methylate histones and inhibit transcription.[62] One of the first lncRNAs discovered was the approximately 17 kilobase (kb) long *cis*-acting lncRNA X-inactive-specific transcript (*Xist*), which covers the X-chromosome from which it is transcribed, resulting in X-chromosome inactivation through recruitment of silencing factors such as the polycomb repressive complex (PRC2) (Figure 2C).[61] Moreover, lncRNAs can serve as molecular decoys, in which the transcribed lncRNA binds to RNA-binding proteins, such as transcription factors, regulatory factors, and chromatin modifiers, and titrates them away (Figure 2B).[62] An example of a decoy is the lncRNA transcribed from the human dihydrofolate reductase (*DHFR*) gene. The lncRNA prevents preinitiation complex formation by forming a stable RNA:DNA complex in the promoter region and interacting with transcription factor IIB (TFIIB) (Figure 2C).[63] Additionally, lncRNAs can act as guides by recruiting chromatin-modifying enzymes to form ribonucleoprotein complexes at the target gene in a *cis*- or *trans*-specific manner. Another transcriptional modulation mechanism involves the co-activation and interaction with transcriptional cofactors that modulate RNA polymerase II (RNAP II) activity (Figure 2C). Furthermore, lncRNAs indirectly interact with mRNAs and upregulate target mRNAs. Post-transcriptional regulation occurs as competing endogenous RNAs (ceRNAs) that can bind to miRNA as a sponge (Figure 2C).[24,64] As scaffolds, lncRNAs control gene expression at the transcriptional level by recruiting multiple RNA-binding proteins such as chromatin modifiers, or species-specific accessory proteins to the gene promoter regions, which may have transcriptional activating or repressing functions (Figure 2B,C).[62]

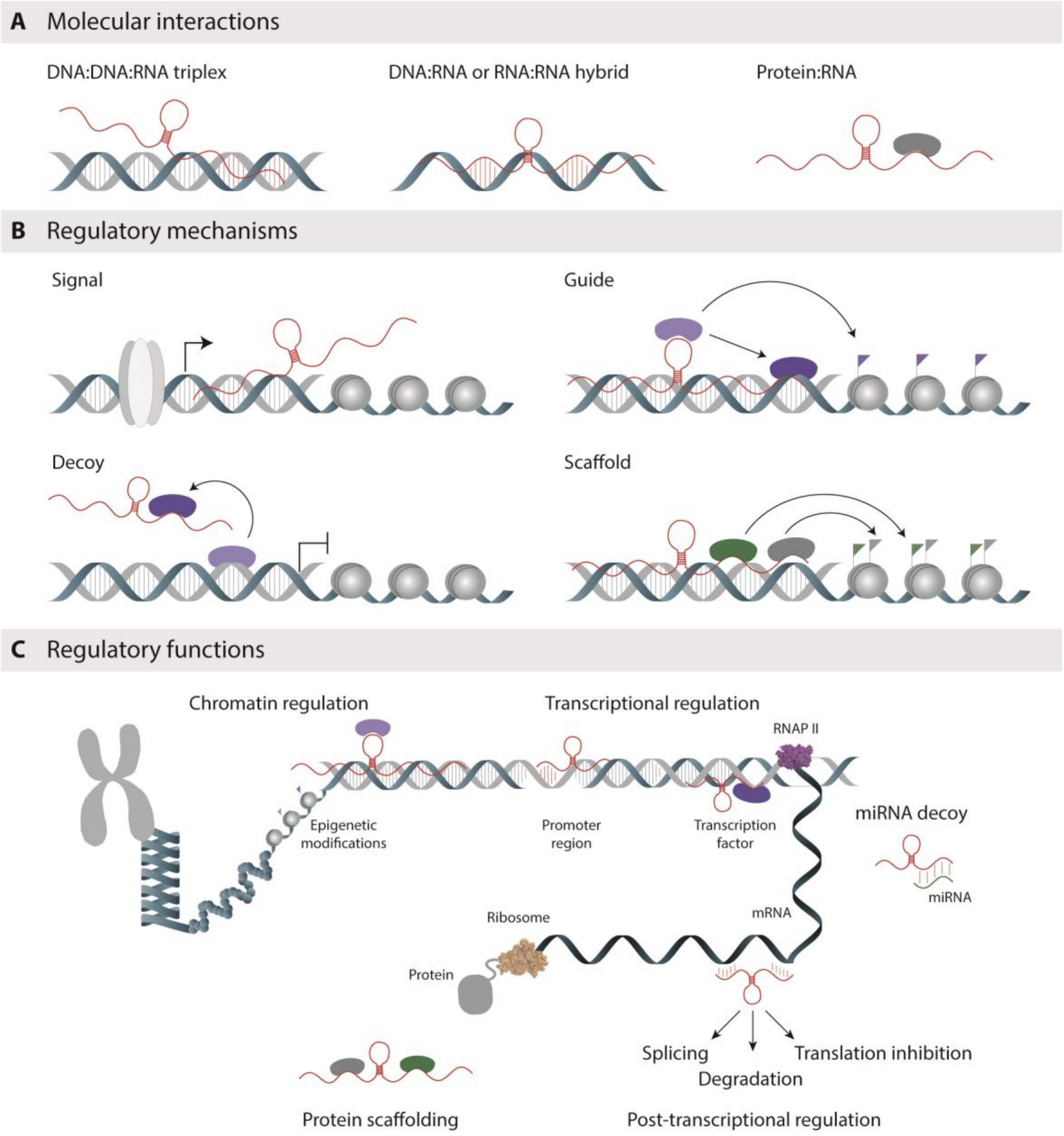


Figure 2. Overview of lncRNA interactions, regulatory mechanisms and functions. (A) Schematic representation of the lncRNA interaction partners. LncRNAs regulate gene expression at various levels through DNA, RNA and protein interactions. (B) Schematic representation of the four archetypes of regulatory mechanisms. LncRNAs can act as signals, guides, decoys, and scaffolds. (Adapted from Wang and Chang[62]) (C) LncRNAs are involved in numerous biological processes such as chromatin and transcriptional regulation that occur in the nucleus. The miRNA interactions and post-transcriptional regulation take place in the cytoplasm. (Adapted from Cheng *et al.*[65])

1.4 Triplexes

Gene expression by lncRNAs is regulated by molecular mechanisms, including intermolecular RNA-protein, RNA-DNA, and RNA-RNA interactions. A possible regulatory mechanism of lncRNAs in targeting specific DNA sequences includes RNA:DNA:DNA triplexes and RNA:DNA hybrid formation. The cellular functions of RNA:DNA:DNA triplexes include chromatin modifications, DNA repair, transcriptional regulation and post-transcriptional RNA processing.[66] The first triple helix, also known as triplex formation, was demonstrated by Felsenfeld *et al.* in 1957 using a 2-fold excess of poly-U strand over poly-A strand in the presence of $MgCl_2$.[67] Moser *et al.* formed triplexes using oligonucleotide sequences (15mers) as triplex-forming oligonucleotides (TFOs) that bind to the major groove of the DNA via Hoogsteen hydrogen bonds.[68]

Single-stranded RNAs (ssRNAs) or TFOs bind to the major groove of double-stranded DNA (dsDNA) or the triplex-target site (TTS) through weaker Hoogsteen or reverse Hoogsteen hydrogen bonds compared to the strong Watson-Crick base-pairing of the dsDNA. Depending on the sequence specificity and the resulting orientation of the RNA strand to the DNA strand, two different classes of intermolecular triplexes can be formed, namely parallel and anti-parallel triplexes (Figure 3A).[6]

Anti-parallel triplexes are formed by reverse Hoogsteen base pairing of purine-rich RNA to double-stranded DNA, resulting in the following triplets: U·AT, G·GC, and A·AT (Figure 3B). They are formed under physiologically relevant pH conditions in the presence of multivalent cations such as Mg^{2+}, but triplex formation with G-rich sequences can compete with G-quadruplex formation, especially in the presence of K^+.[6,69] Parallel triplexes are formed when pyrimidine-rich RNA binds to dsDNA through Hoogsteen hydrogen bonds, resulting in U·AT, G·GC, and C^+·GC triplets. The formation of C^+·GC triplexes requires acidic pH for cytosine protonation, whereas anti-parallel triplexes are formed at neutral pH (Figure 3C).[70] The orientation of the third strand is determined by its sequence. Parallel triplexes are preferentially formed by TC- or UC-rich sequences and anti-parallel triplexes by AG-rich sequences, whereas TG- or UG-rich sequences can form both parallel and anti-parallel triplexes (Figure 3).[6]

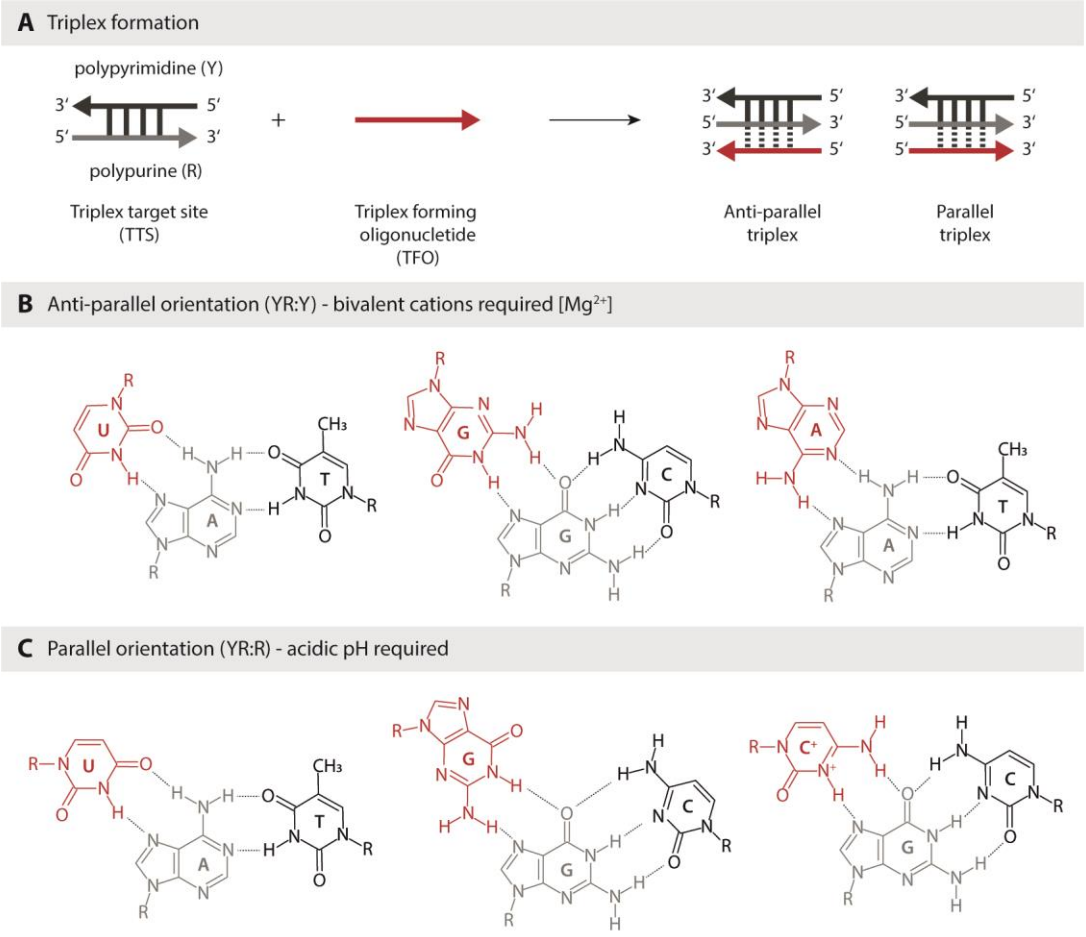

A Triplex formation
polypyrimidine (Y)
polypurine (R)
Triplex target site (TTS)
Triplex forming oligonucleotide (TFO)
Anti-parallel triplex
Parallel triplex
B Anti-parallel orientation (YR:Y) - bivalent cations required [Mg2+]
C Parallel orientation (YR:R) - acidic pH required

Figure 3. Overview of triplex formation. (A) Schematic representation of anti-parallel and parallel triplex formation between the triplex target site and the triplex-forming oligonucleotide. (B) Chemical structures of T·AT, G·GC, and A·AT anti-parallel triplexes. (C) Chemical structures of U·AT, G·GC, and C$^+$·GC parallel triplexes. (Adapted from Jain *et al.*[71])

Triplex stability is determined by the steric properties and the availability of hydrogen acceptor and donor groups from the RNA (or DNA) third strand.[66] Sandström *et al.* used circular dichroism (CD) and nuclear magnetic resonance (NMR) spectroscopy to demonstrate that A-tract DNA TTS are unfavorable for triplex formation due to their intrinsically rigid and highly propeller-twisted structure.[72] Thermodynamic studies have shown improved relative stability of triplex formation when the third strand consists of RNA rather than DNA.[73,74] Kunkler *et al.* determined the stabilities of 16 parallel U·AT-rich RNA:DNA:DNA triplexes as a function of base triplet composition and RNA strand length using native gel-shift assays and thermal denaturation assays. The most stable triplexes were formed with canonical U·AT and C·GC base triplets. However, triplex stability is significantly more dependent on the identity of the Watson-Crick base pairs than on the Hoogsteen base pairs. Triplexes can tolerate a limited number of consecutive non-canonical base triplets except for C·AT, A·AT, and G·AT. The optimal RNA strand length has been determined to be between 19 and 27 nucleotides with apparent binding constants of approximately 220 nM.[75]

Biophysical studies of triplexes destabilized by non-perfect complementarity using CD and NMR spectroscopy showed that the insertion of mismatches, bulges, loops, or other structures was tolerated.[76,77] However, destabilization effects depended on the number of proximal mismatches and their location in the triplex sequences. Larger destabilizing effects were obtained for mismatches in the center of the triplex compared to those located at the terminal positions.[78,79] Kinetic studies of triplexes revealed three orders of magnitude slower triplex association rates (~ 10^3 M^{-1} s^{-2}) compared to duplex association, and half-lives ranging from several minutes to days, depending on the triplex sequences.[80,81]

Triple helices have been identified in several naturally occurring RNAs, such as the S-adenosylmethionine (SAM)-bound riboswitch,[82,83] the U2/U6 snRNA triplex,[84] the human telomerase RNA pseudoknot,[85] and the metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) lncRNA[42]. In addition, miRNAs have been shown to regulate transcription through miRNA:DNA:DNA triplex formation stabilized by Ago proteins.[86]

1.4.1 Intramolecular triplexes

Intramolecular DNA triplexes, also called H-DNA, were first identified as homopyrimidine-homopurine mirror repeat sequences within plasmid DNA. In general, these mirror repeat sequences are located near the regulatory promoter regions of eukaryotic genes.[71] Bacolla *et al.* demonstrated the presence of long homopyrimidine and homopurine sequences of more than 100 base pairs in introns, promoters, and 5'- and 3'-untranslated regions of genes using human genome-wide studies.[87] The intramolecular triplex structure is formed when half of the DNA double helix is disrupted into its single strands, followed by base-pairing with a complementary single strand, resulting in a triple helix conformation.[88] The single-stranded DNA not involved in the H-DNA triplex formation remains single-stranded.[71]

In 1985, Lyamichev *et al.* demonstrated the H-DNA structural transition of supercoiled DNA plasmids containing oligopyrimidine and oligopurine repeat sequences using two-dimensional gel electrophoresis. Similar to the parallel-motif triplexes, the intramolecular H-DNA triplexes are formed under acidic pH conditions.[89] In contrast, the intramolecular homopurine-homopyrimidine sequences are formed into H-DNA-like structures called as *H-DNA triplexes, which are formed at neutral pH and in the presence of Mg^{2+}, analogous to the anti-parallel triplex formation in reverse Hoogsteen configuration.[90]

Intramolecular triplex formation in living cells at single-base resolution has been confirmed in *E. coli* using various chemical probes such as osmium tetroxide[91], trimethyl psoralen[92], and chloroacetaldehyde[93], although the experiments were performed under non-physiological conditions. To investigate the regulatory effect of intramolecular triplexes on gene expression, the triplex sequences were inserted into the coding region of the β-lactamase promoter or *lacZ* gene and the expression of the reporter proteins β-lactamase or β-galactosidase was monitored.[94,95] Furthermore, *in vitro* experiments showed that intramolecular triplexes can interrupt DNA polymerase while acting as arrest signals or acting as pause sites during DNA replication processes *in vivo*.[96]

Intramolecular RNA triplexes play a key role in stabilizing the tertiary structure and folding of RNA.[97,98] The pseudoknot RNA-folding motif is described by pseudoknot structural elements that are stabilized by triplex formation for optimal catalytic activation of human and yeast telomerases. The ends of linear eukaryotic chromosomes are protected from degradation and chromosome fusion by telomeres, which are DNA-protein complexes. The linear ends of eukaryotic chromosomes are replicated by telomerase, an approximately 1000 kDa ribonucleoprotein (RNP) complex consisting of an RNA subunit (TER) that provides the RNA template for the telomeric DNA synthesis, by telomerase catalytic reverse-transcriptase (TERT), and by binding of other protein factors.[99,100]

Some RNA pseudoknots can form triplex structures that are formed between RNA hairpin loop sequences and their complementary bases located elsewhere in the RNA sequence. They can regulate protein synthesis by regulating translation initiation when pseudoknots are located at the internal ribosomal entry site (IRES), or -1 ribosomal frameshifting has been found in many viruses.[101]

1.4.2 Biochemical and biophysical characterization of triplexes

To fully understand the biological functions and molecular mechanisms of triplexes, it is important to gain insight into their structural features using various biochemical and biophysical techniques such as triplex-specific antibodies, electromobility shift assays (EMSAs), CD spectroscopy, melting curve analysis and solution NMR spectroscopy, together with molecular modeling approaches.[102]

Bioinformatics tools such as *TriplexAligner*, *Triplexator*, and *Triplex Domain Finder* have revealed the existence of many triplex-forming motifs in the gene-regulatory regions of the genome, such as the promoter region, which strongly suggesting a regulatory mechanism by triplex formation *in vivo*.[9,103,104] The *in vivo* existence of triplexes has been demonstrated using various approaches, such as triplex-binding monoclonal antibodies (Jel318, Jel466)[105,106], triplex-recognizing fluorescent molecules[107], and immunofluorescence staining of chromosomes with triplex-specific monoclonal antibodies[106]. Lubitz *et al*. demonstrated a strong binding affinity of the fluorescent dye thiazole orange (TO) for DNA triplexes and G-quadruplex DNA compared to dsDNA.[107] Moreover, Dixon *et al.* demonstrated the helicase activity of RecQ, which cleaves triplex structures in the 3'- to 5'-direction of DNA triplexes.[66,108] EMSAs provide a simple and popular approach to validate triplex formation by shifting to a higher molecular weight for the triplex sample compared to ssDNA and dsDNA. Additionally, RNase H digestion allows the differentiation between triplexes and DNA:RNA heteroduplexes.[109–111]

Biophysical methods such as CD spectroscopy, ultraviolet (UV) and thermal melting assays showed distinct features for triplexes compared to dsDNA and RNA:DNA heteroduplexes. For anti-parallel triplexes, changes in ellipticity are prominent, such as the two negative peaks at ~210 and 240 nm, a small positive peak at ~220 nm, and a blue shift of the peak at ~280 nm to ~270 nm in the CD spectrum.[112,113] The CD spectrum of a parallel triplex includes two negative peaks at ~210 and 240 nm, a transition at ~260 nm, and a large positive ellipticity peak at ~280 nm (Figure 4A).[114] Moreover, the biphasic melting transition is a characteristic feature of triplex formation, which is described by the first melting temperature resulting from the dissociation of the RNA from dsDNA by melting of the weaker Hoogsteen hydrogen bonds. The second transition at higher temperatures corresponds to the melting of the strong Watson-Crick base pairing of the dsDNA (Figure 4B).[115]

In addition to the X-ray crystal structures of the RNA:DNA:DNA triplexes, solution state NMR studies of various triplexes were performed to gain insight into their structural properties without the need to crystallize the complex. The proton NMR spectra of parallel and anti-parallel triplexes formed inter- or intramolecularly show characteristic resonances in the imino region. The imino protons involved in Watson-Crick base pairing resonate between 13.7 and 14.3 ppm and for Hoogsteen interactions between 12.9 and 13.6 ppm (Figure 4C). In particular, for parallel triplexes such as C^+·GC, the imino protons of the guanosine resonate between 13.3 and 12.6 ppm and those of protonated cytidine between 14.5 and 16.0 ppm at acidic pH.[70,116,117] NMR studies of anti-parallel triplexes were performed in a Li^+ buffer system to inhibit G-quadruplex formation of the guanosine-rich sequences in the presence of Na^+ and K^+.[118] Furthermore, Sørensen *et al.* determined the NMR solution structure of an intramolecular dsDNA:locked nucleic acid (LNA) triplex as a spiral-like hydrogen bonding pattern of the nucleobases.[119]

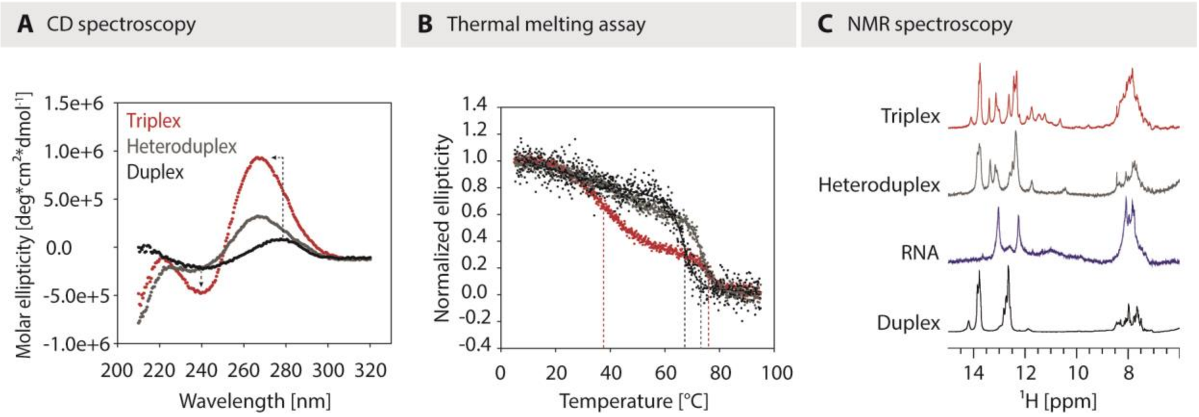


Figure 4. Biophysical characterization of triplex-forming lncRNA *MEG3* and *TGFB2* target DNA. All samples were prepared in triplex buffer (25 mM HEPES, 50 mM NaCl, 10 mM $MgCl_2$ (pH 7.4)). (A) CD spectra of the RNA:DNA:DNA triplex (red), RNA:DNA heteroduplex (gray), and DNA duplex (black). (B) Thermal melting assay of the RNA:DNA:DNA triplex (red), RNA:DNA heteroduplex (gray), and DNA duplex (black). (C) NMR spectra of the RNA:DNA:DNA triplex (red), RNA:DNA heteroduplex (gray), RNA alone (blue), and DNA duplex (black). NMR spectra were measured in 5% D_2O at 298K.

1.4.3 Chemical modifications of TFOs

Triplex-forming oligonucleotides are promising tools for gene targeting and modification. Their modification strategies are being studied to investigate their therapeutic relevance. In general, oligonucleotides can be modified to improve target binding affinity, specificity, and stability to enhance the delivery and uptake of TFOs into cells for *in vivo* applications.[120,121] To modify these TFOs, the ease of synthesis of starting reagents and the introduction of chemical modifications to the bases, e.g., the sugar-phosphate backbone or at the 5′- and 3′-ends of the oligonucleotides, are advantageous.[71]

The nucleobases of TFOs can be modified, where the cytosines used for parallel triplex formation can be substituted by 5-methylcytosine (mC)[122] or N^6-methyl-8-oxo-2′-deoxyadenosine (N^6-8-oxo-dA)[123] to reduce the pH dependence in the parallel triplex motif. The TFOs used for anti-parallel motif triplexes can be modified to prevent intramolecular secondary structure formation, which was achieved by 2′-deoxy-6-thioguanosine (6-thioG)[124] and 7-deaza-2′-deoxyxanthosine (7-dzaX)[125] substitutions (Figure 5).

Modifications of the deoxyribose sugar moiety have been developed to improve TFO stability, as the RNA third strands form more stable triplexes compared to DNA.[73] The 2′-O-methyl ribose (2′-OMe) modification has been described to stabilize triplex formation due to the C3′-endo sugar conformation and reduced sensitivity to RNase H.[126–129] The 2′-O-aminoethyl ribose (2′-AE) modification stabilizes triplex formation by reducing phosphate charge repulsion at physiological pH due to the positively charged amino group (Figure 5).[130–132]

Other sugar modifications were introduced using locked nucleic acids (LNAs) and 2′-O,4′-C-ethylene-linked nucleic acids (ENAs). The LNA molecules provided a C3′-endo locked conformation resulting in stabilized triplexes that were initiated by restricted flexibility in the furanose ring due to the methylene bridge between the 2′-O, 4′-O and 4′-C.[133–135] Frieden *et al.* demonstrated increased resistance to 3′-exonuclease and S1-endonuclease digestion of LNA oligonucleotides characterized by reversed-phase high-performance liquid chromatography (RP-HPLC) and electrospray ionization mass spectrometry (ESI-MS).[136] Partially modified LNA triplexes were shown to have increased thermal stability and binding affinity, although formation of fully modified LNA triplexes was inhibited.[137–139] Triplexes formed with partially modified ENA oligonucleotides have been reported to have increased stability due to the C3′-endo conformation, similar to LNA TFOs, but with the advantage of triplex formation with fully modified ENA TFOs at physiological pH (Figure 5).[138]

Several phosphodiester backbone modifications to increase triplex stability and resistance to enzymatic degradation have been introduced by acridine-linked phosphorothioates[140], N3′-P5′ phosphoramidates[141], and phosphorodiamidate morpholino oligonucleotides[142] where their uncharged backbones reduce the electrostatic repulsion of the negatively charged dsDNA phosphodiester backbone. Furthermore, peptide nucleic acids (PNAs) were designed to overcome the electrostatic repulsion of the dsDNA phosphodiester backbone with a substituted sugar-phosphate backbone, using an uncharged N-(2-aminoethyl)glycine polyamide. Sequence-specific binding of PNAs to dsDNA targets was proposed to occur by strand displacement, resulting in very stable triplexes with high resistance to nucleases and proteases.[143,144] Hu *et al.* showed that PNAs can recognize chromosomal target sequences and regulate gene expression by inhibiting transcription in human cancer cells (Figure 5).[145]

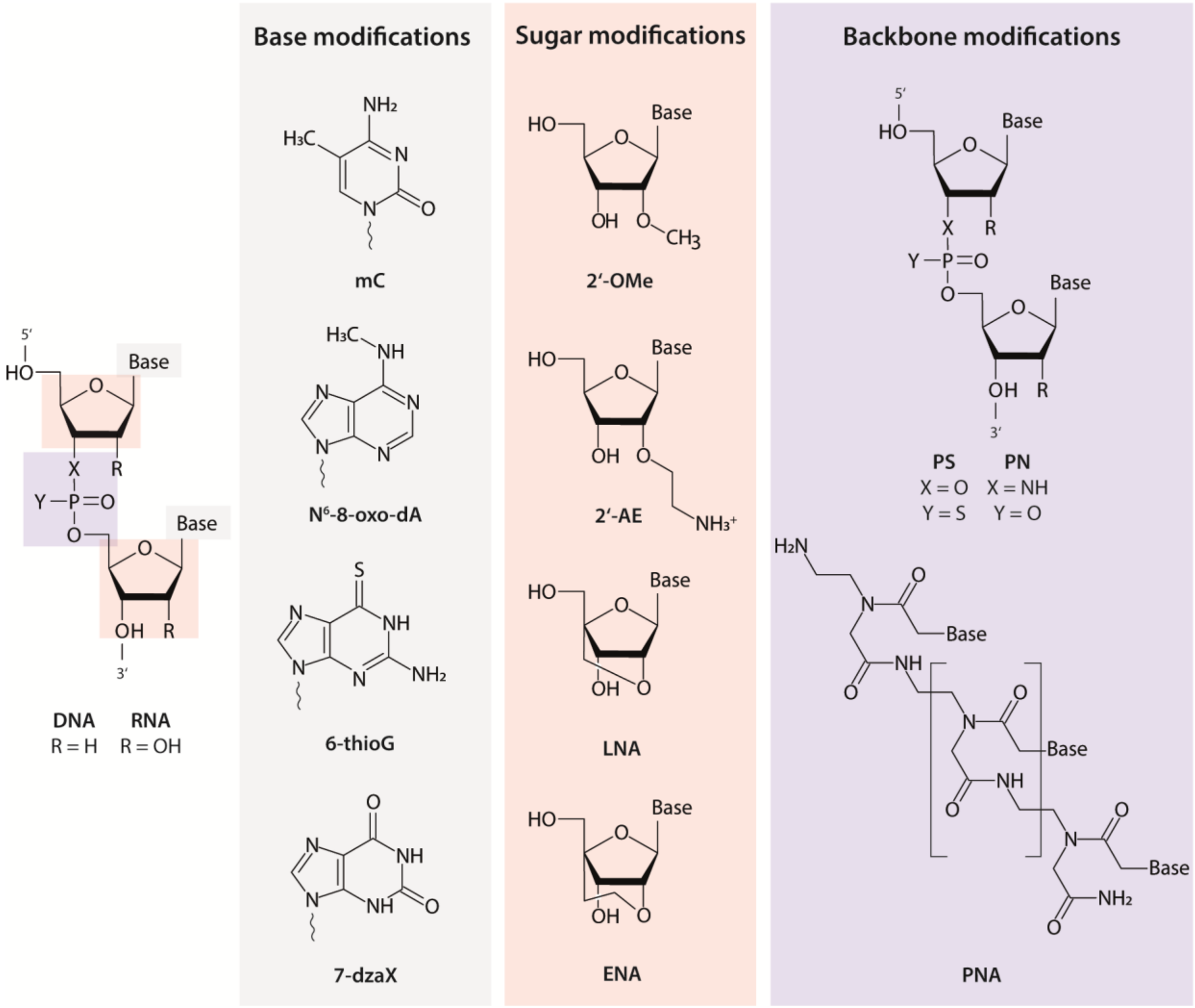


Figure 5. Overview of chemical modifications introduced in TFOs including base, sugar, and backbone modifications. (Adapted from Duca *et al.*[102])

1.5 R-loops

R-loops, also known as RNA:DNA heteroduplexes, are commonly formed as intermediates during DNA replication, DNA repair, and transcription. In 1960, the first RNA:DNA hybrids were synthesized using polyriboadenylic acid and polydeoxyribothymidylic acid and characterized by spectrophotometric and ultra-centrifugation experiments.[146] Hall and Spiegelman reported the formation of RNA:DNA hybrids by annealing T2-specific RNA molecules produced in bacteriophage-infected cells and denatured viral DNA.[147] The crystal structure of RNA polymerase (RNAP) demonstrated that RNA:DNA hybrids of eight base pairs in length were formed within the active site of RNAP II. Inside the transcription bubble, the interaction between single-stranded RNA and double-stranded DNA occurs, where the RNA:DNA hybrid is formed and results in a displacement of the second DNA strand.[148]

Lesnik and Freier reported that the thermodynamic stability of RNA:DNA heteroduplexes, as compared to that of DNA and RNA duplexes, depends on the length and composition of the oligonucleotide strands, with the percentage of deoxypyrimidine in the DNA strand to the corresponding DNA or RNA strand $(A)_n$:$(T/U)_n$. They demonstrated decreased thermal stability for shorter oligonucleotides and for hybrids composed of high $(A)_n$:$(T/U)_n$ content when present in continuous stretches within the sequence.[149,150] Roberts and Crothers demonstrated higher stability of RNA:DNA hybrids containing a purine-rich RNA strand and a pyrimidine-rich DNA strand compared to DNA duplexes.[73] Structural analysis of RNA:DNA heteroduplexes by solid-state NMR and X-ray diffraction studies suggested changing conformations for the RNA:DNA heteroduplex depending on its hydration state between the B-form as compared to dsDNA and the A-form as dsRNA.[150]

RNase H enzymes specifically recognize and cleave the RNA moiety of RNA:DNA heteroduplexes during replication to remove the RNA primers. In addition, in bacteria, R-loops are unwound by the RecG DNA helicase and the Rho RNA helicase, as well as by the Cas3 protein, which cleaves R-loops in the presence of ATP, whereas its RNA:DNA annealing process occurs in the absence of ATP. In eukaryotes, R-loops are cleaved by the Pif1 DNA helicase, the human DHX9 (RHA) RNA helicase, and the human senataxin helicase.[111]

1.6 Triplex-forming lncRNAs

While RNA:DNA:DNA triplex formation has been studied *in vitro* for many years, the identification of triplex-forming lncRNAs *in vivo* and their physiological relevance was demonstrated 50 years later. In 2007, Martianov *et al.* demonstrated triplex formation between a major promoter and an interfering non-coding RNA transcribed upstream of the minor promoter of the human gene encoding dihydrofolate reductase (DHFR). The formation of anti-parallel triplexes was analyzed using *in vivo* and *in vitro* experiments such as chromatin immunoprecipitation (ChIP), RNA immunoprecipitation (RIP), and electromobility shift assays. Transcriptional repression of the GC-rich major promoter occurred through interaction between the AG-rich non-coding RNA and transcription factor IIB (TFIIB), followed by the dissociation of the pre-initiation complex from the major promoter (Figure 6A).[63]

Schmitz *et al.* reported transcriptional regulation of ribosomal RNA genes by promoter-associated RNAs (pRNAs) through RNA:DNA:DNA triplex formation with the transcription factor (TTF-1) target site within the rRNA promoter recognized by DNA methyltransferases 3a and b (DNMT3a and b). They proposed an ncRNA-dependent DNA methylation mechanism as pRNAs can target the enzymes DNMT3a and DNMT3b to specific genomic sites through heterochromatin formation in the rDNA promoter, resulting in DNA methylation and transcriptional silencing (Figure 6B).[109]

Furthermore, fetal-lethal non-coding developmental regulatory RNA (*Fendrr*) is critical for mammalian heart and body development by *cis*- and *trans*-regulating gene expression of the target promoters *Foxf1* and *Pitx2*, respectively. Grote *et al.* proposed an epigenetic control mechanism for *Fendrr* to recruit the polycomb repressive complex 2 (PRC2) or trithorax group/mixed lineage leukemia (TrxG/MLL) protein complexes to the target promoter, which can modify its histone methylation status by repressing or activating histone marks, respectively. They identified a 40 nt long UC-rich *Fendrr* RNA region that can form parallel-oriented DNA:RNA triplexes with its target promoters *Foxf1* and *Pitx2*.[8] The epigenetic modifier PRC2 is recruited by *Fendrr* to the target promoters, resulting in repressive marks as trimethylation of lysine 27 on histone 3 (H3K27me3), whereas the TrxG/M:*Fendrr* complex provides activating marks through trimethylation of lysine 4 on histone H2 (H3K4me3) (Figure 6C).[7]

Postepska-Igielska *et al.* functionally characterized the UC-rich lncRNA *Khps1*, which is an antisense transcript of the proto-oncogene sphingosine kinase 1 (*SPHK1*) gene, which induces proliferation and protects cells from apoptosis.[151,152] Upstream of the transcription start site of *SPHK1* isoform B (*SPHK1-B*), several triplex target sites or triplex-forming regions have been proposed to interact with the lncRNA *Khps1.* Thus, *SPHK1* transcription is induced by *Khps1* through the recruitment of the histone acetyltransferase p300/CBP to the promoter region, resulting in an open chromatin structure with modified histones, which further ensures the binding of the transcription factor E2F1.[151] Recently, a follow-up study revealed a “feed-forward” transcriptional regulation mechanism involving a RNA:DNA:DNA triplex formation between the *SPHK1* enhancer (*eSPHK1*) and *Khps1* with the recruitment of E2F1 and p300, which induced eRNA-*SPHK1* and *SPHK1* mRNA transcription. *SPHK1* expression was enabled by inhibited binding of the transcription factor CTCF (Figure 6D).[153,154]

O’Leary *et al.* showed that the GA-rich antisense lncRNA promoter of the MAT2A-antisense radiation-induced circulating ncRNA (*PARTICLE*) is transcribed from the methionine adenosyltransferase 2A (*MAT2A*) promoter region by low-dose irradiation. Triplex formation occurs between *PARTICLE* and the *MAT2A* promoter, which inhibits the expression of the tumor suppressor *MAT2A*. The triplex formation induces the recruitment of transcriptional repressor complex subunit proteins of the PRC2 complex, namely G9a and SUZ12 that inhibits the *MAT2A* expression by methylating its CpG region (Figure 6E).[155]

Mondal *et al.* described the regulatory mechanism of the triplex-forming and chromatin-interacting lncRNA transcribed from human maternally expressed gene 3 (*MEG3*), where its GA-rich sequences direct *MEG3* to transforming growth factor-β (TGF-β) chromatin target genes in a *trans*-regulatory manner via anti-parallel RNA:DNA:DNA triplex formation. Triplex formation was confirmed by *in vitro* assays such as EMSA and CD spectroscopy, as well as *in vivo* immunostaining experiments using anti-triplex monoclonal antibodies. *MEG3* contains PRC2-interacting sequences that can scaffold PRC2 components to the chromatin by introducing H3K27me3 marks and inhibit transcription of the TGF-β (*TGFBR1*) target genes through triplex formation (Figure 6F).[156]

Another *trans*-acting and GA-rich lncRNA *HOTAIR*, transcribed from the *HOXC* locus, was shown by Kalwa *et al.* to regulate gene expression of cellular aging processes in mesenchymal stem cells at the level of DNA methylation via triplex formation. The transcription of the *HOTAIR* target genes *PCDH7* and *HOXB2* is inhibited by triplex formation and subsequent scaffolding of the histone modification complexes such as PRC2 and lysine-specific demethylase 1 (LSD1) (Figure 6G).[157]

The triplex-forming and *trans*-acting lncRNA *HIF1α-AS1* is an antisense transcript of the hypoxia-inducible factor 1-alpha (*HIF1A*) gene, which is an oxygen-dependent, angiogenic and lung disease-related lncRNA. Anti-parallel triplex formation was identified between *HIF1α-AS1* and the DNA target genes of ephrin receptor A2 (*EPHA2*) and adrenomedullin (*ADM*). Chromatin modifying proteins such as the M-phase phosphoprotein 8 (MPP8) and SET domain bifurcated histone lysine methyltransferase 1 (SETDB1) from the human silencing hub (HUSH) complex are recruited by the lncRNA to the target genes mediating gene silencing (Figure 6H).[158]

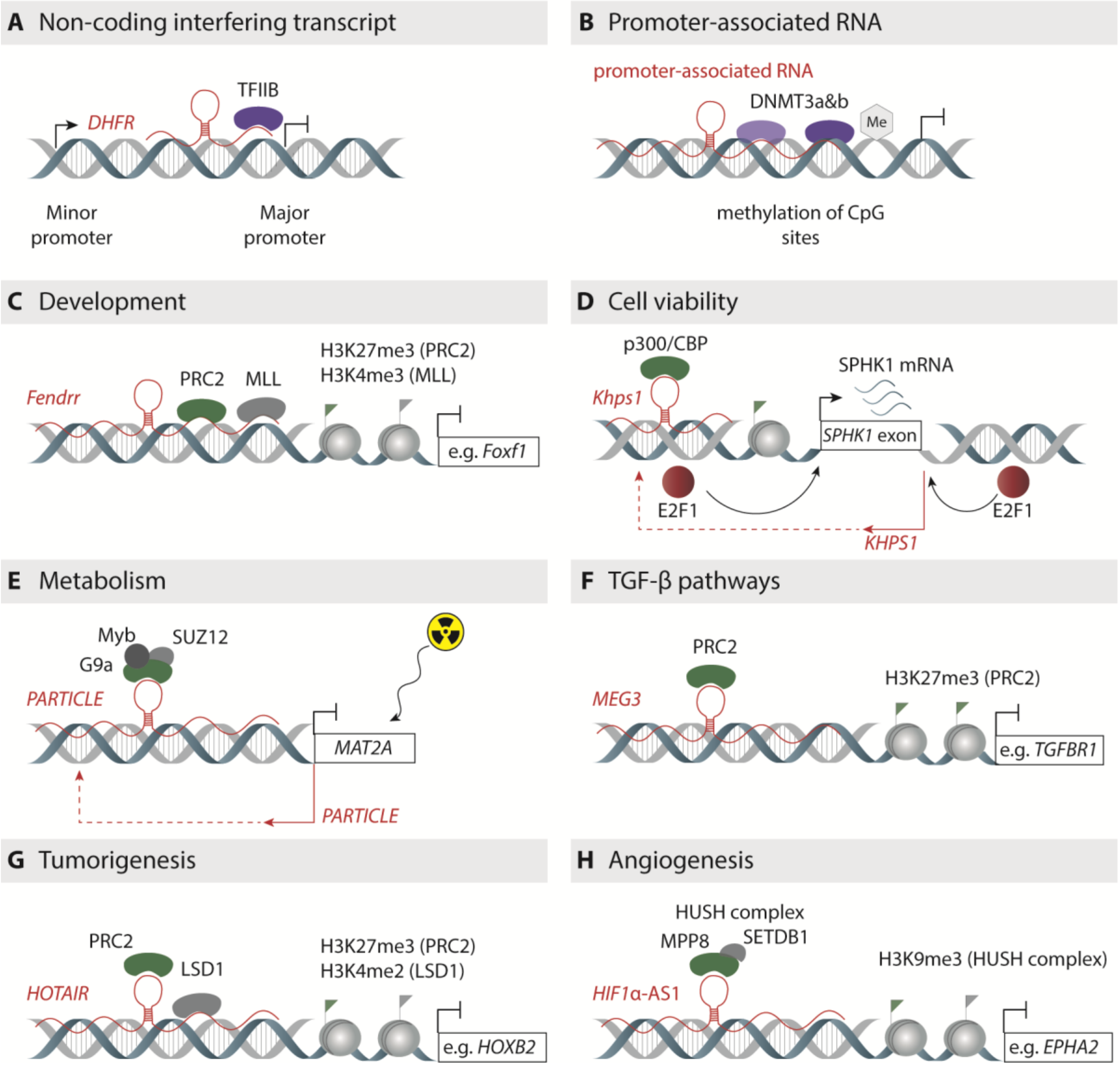


Figure 6. Overview of triplex-forming lncRNAs and their proposed regulatory mechanisms. (A) The interfering ncRNA *DHFR* interacts with the major promoter through triplex formation and binds TFIIB, leading to dissociation of the pre-initiation complex and transcriptional repression. (B) Promoter-associated RNA methylates the CpG sites by recruiting DNA methyltransferase enzymes, resulting in transcriptional silencing. (C) The triplex-forming lncRNA *Fendrr* recruits the chromatin modifiers PRC2 and MLL, which can set the repressive marks H3K27me3 and H3K4me3 at the target genes such as *Foxf1,* leading to repressed gene expression. (D) The *SPHK1* mRNA transcription is enhanced by the interaction of the triplex-forming lncRNA *Khps1* with the transcription factors p300/CBP and E2F1. (E) The anti-sense triplex-forming lncRNA *PARTICLE* is transcribed from *MAT2A* upon irradiation. Gene expression is downregulated by binding of a transcription-repressive complex to the triplex. (F) The trans-acting lncRNA *MEG3* inhibits the gene expression of its target genes through anti-parallel triplex formation and recruitment of PRC2. (G) *HOTAIR* represses transcription of its target genes through triplex formation and recruitment of histone-modifying complexes such as PRC2 and LSD1. (H) *HIF1α-AS1* mediates gene silencing of its target genes such as *EPHA2* and *ADM* through triplex formation and recruitment of the HUSH complex, resulting in repressive chromatin at the target loci. (Adapted from Warwick *et al.*[9])

CHAPTER II

Cell-free protein synthesis

2.1 Cell-free protein synthesis

The cell-free protein synthesis (CFPS) approach is based on the idea of performing *in vitro* translation reactions using the ribosomal translation machinery extracted from living cells, as an alternative to the classically used *in vivo* protein expression.[159] In the 1960s, Nirenberg and Matthaei pioneered cell-free experiments using cell-extracts from *Escherichia coli* (*E. coli*) to decipher the genetic code and they discovered the link between mRNA and protein synthesis.[160,161] Cell-free experiments have also been used to elucidate the regulatory mechanisms of the *E. coli* lactose- and tryptophan operons.[162,163] An advantage of the CFPS approach is that the open nature of the system enables many transcription and translation parameters to be adjusted.[10]

Traditional approaches for protein production include solid-phase chemical synthesis and traditional cell-based expression systems such as *E. coli* or *Saccharomyces cerevisiae*.[164,165] A disadvantage of traditional solid-phase chemical synthesis is the size limitation of approximately 40 residues, which allows only the production of small peptides.[165] The cell-based expression system is very time- and resource-consuming, as overexpression of recombinant proteins is difficult for large-scale protein production or high-throughput screening.[11]

CFPS reactions can be performed in coupled and uncoupled formats. Coupled CFPS reactions are mediated by the addition of a DNA template where simultaneous mRNA and protein synthesis occurs. The coupled CFPS reaction setup is easy to handle because the reaction is performed in a single tube. The mRNA initiates uncoupled CFPS reactions, which are then translated into proteins. Direct addition of mRNA to the reaction mixture allows precise manipulation of transcription and translation conditions. In addition, the uncoupled reaction setup is advantageous for eukaryotic CFPS systems because it allows for mRNA modifications, such as pseudouridine incorporation, to increase translation efficiency.[11,166]

The main component of the CFPS system is the crude cell-extract or lysate prepared from *E. coli*, rabbit reticulocytes, wheat germ, or insect cells.[167–169] For CFPS reactions, the crude cell-extract is depleted of endogenous DNA and mRNA but retains the translation machinery. The cell-extract is supplemented with tRNA and tRNA enzymes, protein translation factors, and an energy regeneration system mixed with necessary cofactors, salts, amino acids, and nucleotides (Figure 7).[170] CFPS reactions are performed using a DNA template as a linear polymerase chain reaction (PCR) product, a linearized DNA plasmid, or a circular vector consisting of a strong T7 promoter sequence in the 5'-untranslated region (5'-UTR), which facilitates high transcription rates.[171] The promoter is followed by a ribosome-binding site (RBS), the Shine-Dalgarno (SD) sequence in prokaryotes and the Kozak sequence in eukaryotes, the coding sequence as well as a T7 terminator in the 3'-UTR to enable efficient release and recycling of the ribosome.[172,173]

Depending on the application, CFPS reactions can be performed in a variety of setups, including batch, continuous exchange, and flow setups. The batch setup is easy to use with a short reaction time and low reaction cost. The reaction mixture is incubated in a single tube, which limits the reaction time because the amount of substrate available is limited and the amount of inhibitory byproducts increases over time (Figure 7).[170] The continuous setup is advantageous because it utilizes a two-compartment system where new substrate is added to the reaction mixture and the inhibitory byproducts are removed, resulting in higher protein yields and longer reaction times. The continuous exchange cell-free (CECF) setup consists of two compartments separated by a semi-permeable membrane that allows exchange of low-molecular weight compounds. The reaction mixture, which contains nucleic acids and proteins, is separated from the feeding mixture, which consists of low-molecular weight components such as nucleotides and amino acids (Figure 7).[169] In the continuous flow cell-free (CFCF) setup, the substrate-enriched feeding mixture is pumped into the reaction compartment, while the synthesized protein and inhibitory byproducts such as pyrophosphate are pushed out through an ultrafiltration membrane (Figure 7).[174] However, the batch setup is suitable for simple and rapid protein synthesis, which can be performed for upscaled reactions and high-throughput screening in *E. coli* or wheat germ systems. For large-scale protein production, especially in organisms with low protein yields, the continuous setups are advantageous as they can be applied in industrial applications.[169] Zawada and coworkers reported an industrial-scale 100 L CFPS reaction performed in a bioreactor resulting in a protein yield of 700 mg/L over a 10 h synthesis time.[175]

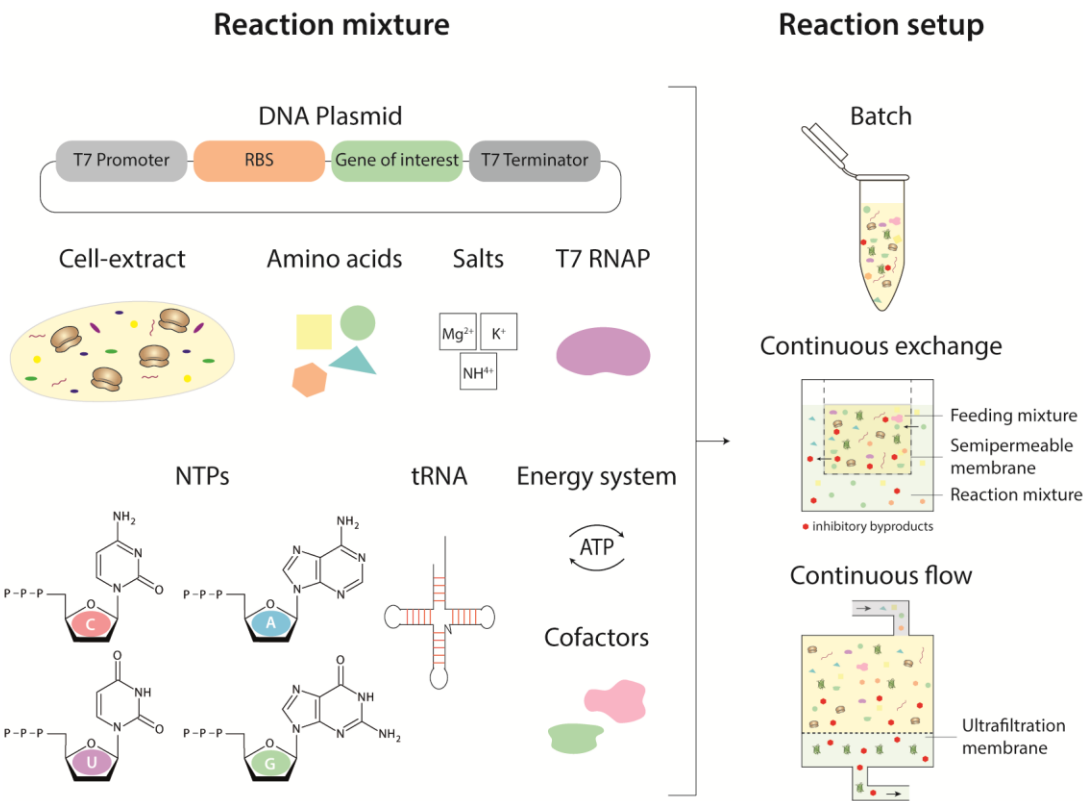


Figure 7. Schematic representation of the CFPS system. The reaction mixture components of the CFPS reaction are shown on the left. The right side shows a comparison of the batch, continuous exchange, and continuous flow reaction setups. (Adapted from Gregorio *et al.*[11])

In general, cell-extract-based CFPS kits are commercially available with ready-to-use mixtures that are quite expensive compared to in house-prepared cell-extracts.[176] In addition, the protein synthesis using recombinant elements (PURE) system is a reconstituted CFPS system with highly purified protein components of the translation machinery that is also commercially available (e.g., PURExpress® by New England BioLabs). The initiation factors (IF1, IF2, and IF3), elongation factors (EF-G, EF-Tu, EF-Ts), release factors (RF1, RF2, RF3), ribosome recycling factors (RRF), 20 aminoacyl tRNA synthetases, creatine kinase, myokinase, nucleoside diphosphate kinase, methionyl tRNA formyltransferase, and pyrophosphatase have been His-tagged, purified individually, and reconstituted in the PURE system. The PURE system is completed by the addition of ribosomes prepared by sucrose density gradient centrifugation, tRNA mixtures, buffer, and NTPs.[177,178] An advantage of the PURE system is the absence of biomolecules and metabolites irrelevant to protein synthesis, as well as proteases and nucleases, which results in increased transcription and translation levels. However, the preparation and purification of all the components used in the PURE system is very time-consuming and the commercially available kits are expensive.[159,171]

2.2 Applications of CFPS

In vitro protein synthesis with the CFPS system offers several advantages in terms of production and scalability. The reaction conditions of CFPS systems can be easily modified, which is a major advantage for difficult-to-express proteins, such as membrane proteins, due to the improved protein folding in the CFPS system. This approach was initially used as an analytical tool but has gained popularity as a preparative technique with a wide range of applications, especially for protein expression, therapeutic development, high-throughput approaches, and educational purposes, because the reactions are performed in a simple manner with reduced reaction volume and process time.[164,179]

Traditional *in vivo* protein expression is not suitable for membrane proteins (MPs) because they can be overexpressed in misfolded, aggregated, or cytotoxic forms. However, the structural and functional study of MPs is crucial because they are involved in many processes, including cell recognition, signal transduction, immune response, and molecular transport.[11] For proper folding and solubilization, MPs require a hydrophobic environment, which is provided by the addition of detergents, nanodiscs, or liposomes to the CFPS system.[180] The uncoupled CFPS system using *in vitro* transcribed mRNA is preferred for the expression of potential drugs, such as translation inhibitory proteins (Figure 8).[171]

Protein engineering using the CFPS approach is promising because it allows the parallel expression and rapid screening of large and diverse protein sequence libraries.[181] This has been applied in particular to '*in vitro* expression cloning' (IVEC), where a plasmid pool containing 50 to 100 clones is expressed and screened for the desired biochemical activity in a coupled transcription-translation system. This process is repeated with active pools until individual plasmid clones with the desired activity are obtained.[173,182] In addition, potential vaccine candidates have been screened and synthesized in the CFPS system, such as virus-like particles (VLPs), which are genome-free and non-infectious but contain structural proteins of the virus. The production of VLPs in CFPS systems has been performed on a large scale, resulting in high yields. Moreover, toxic VLPs could be synthesized because the CFPS approach allows disulfide bond formation by adjusting the redox potentials within the reaction mixture.[183,184] This approach has been used in particular for novel malaria vaccine candidates produced in wheat germ cell-extracts (Figure 8).[185]

Protein modifications are easily incorporated into the CFPS using non-natural or chemically modified tRNAs. N-terminal fluorescently labeled proteins are produced using fluorescently labeled initiator, methionine tRNA (fmet-tRNA).[186] While, C-terminal fluorescent labeling is possible by incorporating fluorescently labeled puromycin analogs such as 'Cy5-dC-puromycin', a fluorophore with a deoxycytidine linker attached to puromycin.[187] Incorporation of unnatural amino acids (UAA) is performed using an orthogonal pair of tRNA and aminoacyl tRNA synthetase to suppress mutations by frameshifting. This approach can be used to synthesize proteins with novel properties, structural elements, and functions, as well as to incorporate post-translational modifications, including ubiquitination, glycosylation, and phosphorylation.[164,188] As an example, *E. coli* cell-extracts were prepared from a strain lacking the release factors that normally recognize the ochre (TAA), opal (TGA), and amber (TAG) stop codons in the mRNA sequence, thereby terminating translation.[188,189] The amber stop codon approach allows the synthesis of a selectively modified protein by incorporating UAAs using a specifically designed orthogonal pair of complementary amber tRNA^{CUA} and aminoacyl-tRNA synthetase (Figure 8).[188,190]

Selectively isotope-labeled (^{15}N-labeled) proteins can be prepared for structure determination by NMR spectroscopy.[181] Compared to heterologous protein expression in *E. coli,* the cell-free system allows for rapid reactions in small reaction volumes and introduces the isotope-labeled amino acids in a resource-efficient manner with reduced isotope scrambling effects.[191–193] Furthermore, the CFPS system allows for high-throughput production of amino-acid specific isotope-labeled proteins in high yields. Proteins obtained from crude reaction mixtures can be measured by NMR spectroscopy after simple dialysis to remove low-molecular weight compounds and without chromatographic purification (Figure 8).[191,194]

Freeze-dried CFPS systems are important tools as educational kits for teaching the basics of molecular and synthetic biology, particularly protein synthesis, fluorescence, and enzyme kinetics. BioBits™, for example, is a commercially available lyophilized CFPS system that contains the complete transcription and translation machinery from the cell lysate. Users can simply start reactions by adding water and the DNA templates for various fluorescent proteins (Figure 8).[195,196]

Living cells respond to environmental changes by controlling gene expression through various DNA-modifying regulators, RNAs, and proteins. To study these complex regulatory processes, *in vitro* genetic circuits, also called biosensors, have been developed *in vitro* using the cell-free expression system and further verified in *in vivo* systems (Figure 8).[197,198] Shin and Noireaux demonstrated synthetic genetic circuits of transcription rate regulation using various polymerases, including various *E. coli* sigma factors, T7 and T3 bacteriophage RNA polymerases.[199] In addition, Chizzolini *et al.* characterized a strong influence of different T7 transcriptional promoters at the transcriptional level compared to the translational level. They simultaneously quantified the mRNA and protein levels of one-, two- and three-gene operons using Spinach RNA aptamers and fluorescent proteins encoded on the same plasmid.[200] Additionally, genetic circuits with RNA-regulatory elements that control gene expression via transcriptional processes (RNA transcriptional attenuators), catalysis (ribozymes) and translation (sRNAs, RNA riboswitches, RNA aptamers, RNA thermometers) have been investigated.[198,201] Currently, cell-free biosensor engineering is important for environmental and clinical applications to develop low-cost biosensors that can detect chemicals, nucleic acids, and pathogens with high sensitivity and selectivity (Figure 8).[202] Thavarajah *et al.*

demonstrated a fluoride-sensing riboswitch-based biosensor that can monitor fluoride-contaminated groundwater with a detection limit of 50 µM fluoride concentration.[203] Recently, Hunt *et al.* developed a SARS-CoV-2 CFPS paper-based toehold switch biosensor. In the unbound state of viral RNA, the ribosome binding site (RBS) and start codon are sequestered, resulting in translation inhibition. Upon binding of SARS-CoV-2 RNA to the toehold switch, the RBS and start codon are released, inducing translation of the NanoLuc bioluminescent reporter protein.[204] Moreover, these toehold switch biosensors have been used to detect Zika and Ebola virus sequences.[205,206]

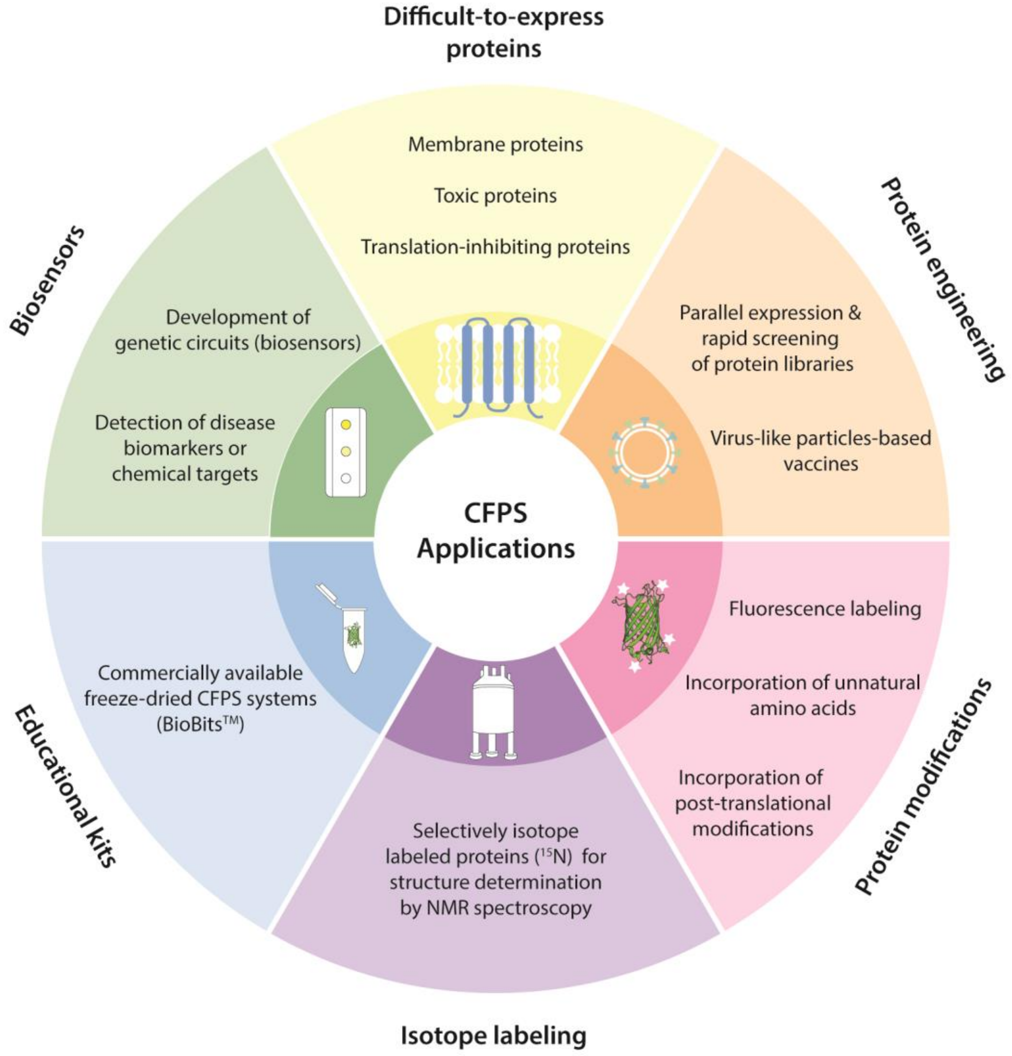

Difficult-to-express proteins
Membrane proteins
Toxic proteins
Translation-inhibiting proteins
Protein engineering
Parallel expression & rapid screening of protein libraries
Virus-like particles-based vaccines
Protein modifications
Fluorescence labeling
Incorporation of unnatural amino acids
Incorporation of post-translational modifications
Isotope labeling
Selectively isotope labeled proteins (^{15}N) for structure determination by NMR spectroscopy
Educational kits
Commercially available freeze-dried CFPS systems (BioBits™)
Biosensors
Development of genetic circuits (biosensors)
Detection of disease biomarkers or chemical targets
CFPS Applications

Figure 8. Various applications of CFPS systems. CFPS has gained popularity as a preparative technique with a wide range of applications, including expression of difficult-to-synthesize proteins, protein engineering, incorporation of various protein modifications, and isotopic labeling. In addition, this technique can be used for educational purposes and as a biosensor.

CHAPTER III

Riboswitches

3.1 Riboswitches

Riboswitches are conserved regulatory elements that have been identified mostly in the 5'-untranslated region of mRNAs from bacteria, but also in eukaryotes, including fungi, plants, and algae.[12] These organisms sense and respond to external changes, such as temperature and metabolite concentration, through structural changes in the non-coding RNA elements. Gene expression of the metabolite biosynthetic pathway is usually controlled in a *cis*-acting manner, but an additional *trans*-acting mechanism has been demonstrated for two S-adenosylmethionine (SAM) riboswitches from *Listeria monocytogenes*.[207,208] The metabolites, including coenzymes (adenosylcobalamin, flavin mononucleotide, S-adenosylmethionine), nucleotide derivatives (adenine, guanine, 2-deoxyguanosine), ions (Mg^{2+}, Mn^{2+}, F^-), amino acids (lysine, glycine, glutamine), and signaling molecules (cyclic di-guanosine monophosphate (c-di-GMP), cyclic di-adenosine monophosphate (c-di-AMP)) are recognized by the known riboswitch classes.[12,209,210] Riboswitches are crucial for gene regulation processes in many bacteria and these RNA regulatory elements are being investigated as novel antimicrobial drug targets.[211,212]

Riboswitches are divided into two structurally relevant parts: the aptamer and the expression platform. The aptamer contains the conserved RNA sensor that binds its cognate ligand with high specificity and micro- to nanomolar affinity and induces folding of the downstream expression platform that controls gene expression.[13,14,213] In some riboswitches, tandem aptamers containing two aptamer domains have been discovered, where both aptamers bind either the same or different ligands.[214] Upon ligand binding, riboswitches can control gene expression at the level of transcription, translation, as well as mRNA splicing and mRNA decay, resulting in either gene activation (on-switch) or gene repression (off-switch) of the associated gene (Figure 9).[13]

The transcriptional regulatory riboswitches control gene expression by altering the secondary structure of the expression platform induced by ligand binding to the aptamer domain.[215] The formation of a stabilized transcription-terminator stem with subsequent uridine residues at the 3'-end of the riboswitch induces RNA polymerase stalling, resulting in transcription termination with release of the DNA template and transcribed RNA (Figure 9A, B).[13,216,217] Most *Bacillus subtilis* (*B. subtilis*) riboswitches regulate gene expression through premature transcription termination induced by the formation of Rho-independent terminators.[13,213]

Many *E. coli* riboswitches regulate gene expression at the translational level by adjusting the accessibility of the ribosome-binding site and the start codon (AUG) to the ribosome. Translation control of riboswitches is regulated by the sequestered or released form of the RBS, including the Shine-Dalgarno sequence present in the expression platform. The AUG start codon is located approximately six nucleotides downstream of the RBS sequence.[218] During translation initiation, the purine-rich RBS sequence interacts with 16S ribosomal RNA by Watson-Crick base pairing.[219] Upon ligand binding, a translational on-switch alters its expression platform to allow the ribosome to bind to the RBS sequence of the mRNA. In the case of translational off-switches, ribosome binding to the mRNA is inhibited by sequestering the RBS (Figure 9C, D).[220] For many translation-regulating *E. coli* riboswitches, another mRNA regulation mechanism has been demonstrated as a consequence of translation inhibition. The *E. coli ribB* FMN-sensing riboswitch regulates its mRNA levels by Rho-dependent transcription termination and the *E. coli lysC* lysine-sensing riboswitch regulates mRNA decay by Rho-dependent transcription termination and RNase E-dependent degradation.[221–223]

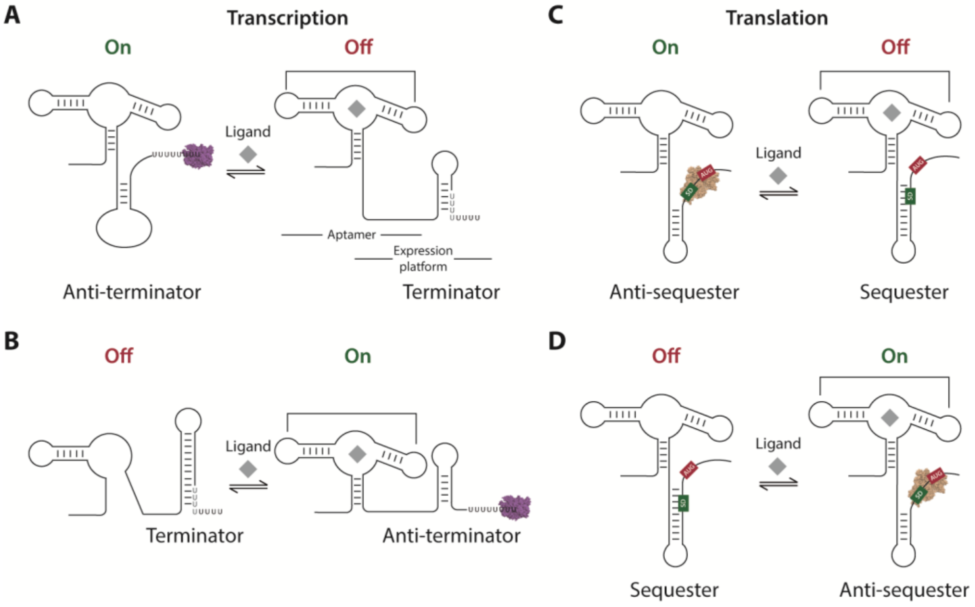


Figure 9. Overview of riboswitch mechanisms that control gene expression at the transcriptional and translational levels. (A) Transcriptional off-switches inhibit gene expression through a transcriptional termination mechanism. The ligand-unbound state forms an anti-terminator hairpin that allows the transcription of the entire gene by RNA polymerase (purple, PDB: 1MSW). Upon ligand binding, the intrinsic terminator hairpin is formed, which inhibits the transcription. (B) Transcriptional on-switches activate gene expression through a transcriptional elongation mechanism. The ligand-unbound state forms a terminator hairpin that terminates transcription. Ligand binding to the aptamer induces anti-terminator formation, allowing transcription to continue. (C) Translational off-switches inhibit translation initiation by sequestering the SD sequence and start codon within the expression platform upon ligand binding. (D) Translational on-switches activate gene expression by initiating translation. Upon ligand binding, the SD sequence and start codon are accessible to the ribosome (brown, PDB: 6O7K), which initiates translation. (Adapted from Ariza-Mateos *et al.*[224])

3.2 Thiamine pyrophosphate-sensing riboswitches

The thiamine pyrophosphate (TPP)-sensing riboswitches (THI-box) regulate genes involved in the biosynthesis and transport of thiamine and its phosphorylated derivatives, as the coenzyme TPP is crucial for enzymes in carbohydrate metabolism.[225,226] The conserved sequences of TPP-sensing riboswitches have been discovered in many organisms, including bacteria and archaea, and it is the only class of riboswitches found in eukaryotes.[227] Specifically, these sequences have been found in 3'-UTRs of plants and in introns of fungi.[228,229] Bacteria can produce thiamine (vitamin B_1) on their own, whereas animals and fungi require the uptake of thiamine or its analogues.[226]

The regulatory mechanisms of TPP-sensing riboswitches are diverse depending on the regulating organism. In bacteria, they control gene expression by terminating transcription and inhibiting translational initiation, and in eukaryotes, they regulate splicing.[230] Specifically, in *E. coli*, TPP-sensing riboswitches regulate thiamine biosynthesis and transport by controlling gene expression of *the thiMD* operon at the translational level and *the thiCEFSGH* operon at the transcriptional and translational levels.[231] The *E. coli* thiM riboswitch controls gene expression of the enzyme hydroxyethylthiazole kinase as a translational off-switch.[225,230] In the presence of TPP, the thiM expression platform undergoes conformational changes that inhibit translation initiation by the sequestered SD sequence (Figure 10A). Bastet *et al.* demonstrated a continuous regulation mechanism of thiM by a Rho-dependent transcription termination that occurs after translation regulation.[223] Furthermore, Winkler *et al.* reported a 1000-fold discrimination of TPP over its derivatives thiamine and thiamine monophosphate for the thiC riboswitch. They determined a TPP affinity of 100 µM for thiM (165 nucleotide (nt) fragment) and 600 nM for thiC (240 nt fragment),[225] whereas Kulshina *et al.* demonstrated Mg^{2+}-dependent TPP binding affinities for thiM (0.5 mM Mg^{2+}: K_D ~ 200 nM; 2.5 mM Mg^{2+}: K_D ~ 9 nM).[232]

Crystal structures have been determined for the TPP-bound conformations of the *E. coli* thiM riboswitch.[233,234] The structures showed that TPP-riboswitches are composed of five helices (P1 to P5) connected by terminal loop regions (L3 and L5) and junction bulges (J3-2, J2-4, and J4-5), whereas the P2/J3-2/P3/L3 and P4/J4-5/P5/L5 regions form two parallel helices connected to helix P1 by a three-way junction (Figure 10B).[230] TPP has been shown to bind with its pyrimidine ring to the J3-2 junction bulge within the aptamer domain, whereas the negatively charged pyrophosphate residue binds to the J4-5 junction bulge in the presence of divalent cations (Mg^{2+}, Ba^{2+}, Mn^{2+}). Specifically, the J3-2 junction recognizes the 4-amino-5-hydroxymethyl-2-methylpyrimidine (HMP) ring by intercalation of the pyrimidine residue between G42 and A43 and further interactions with G19 and G40 (Figure 10C, D).[234–236] Small-angle X-ray scattering (SAXS) data revealed a compact thiM structure in the presence of physiological Mg^{2+} conditions and folding of the aptamer domain upon metabolite binding.[237]

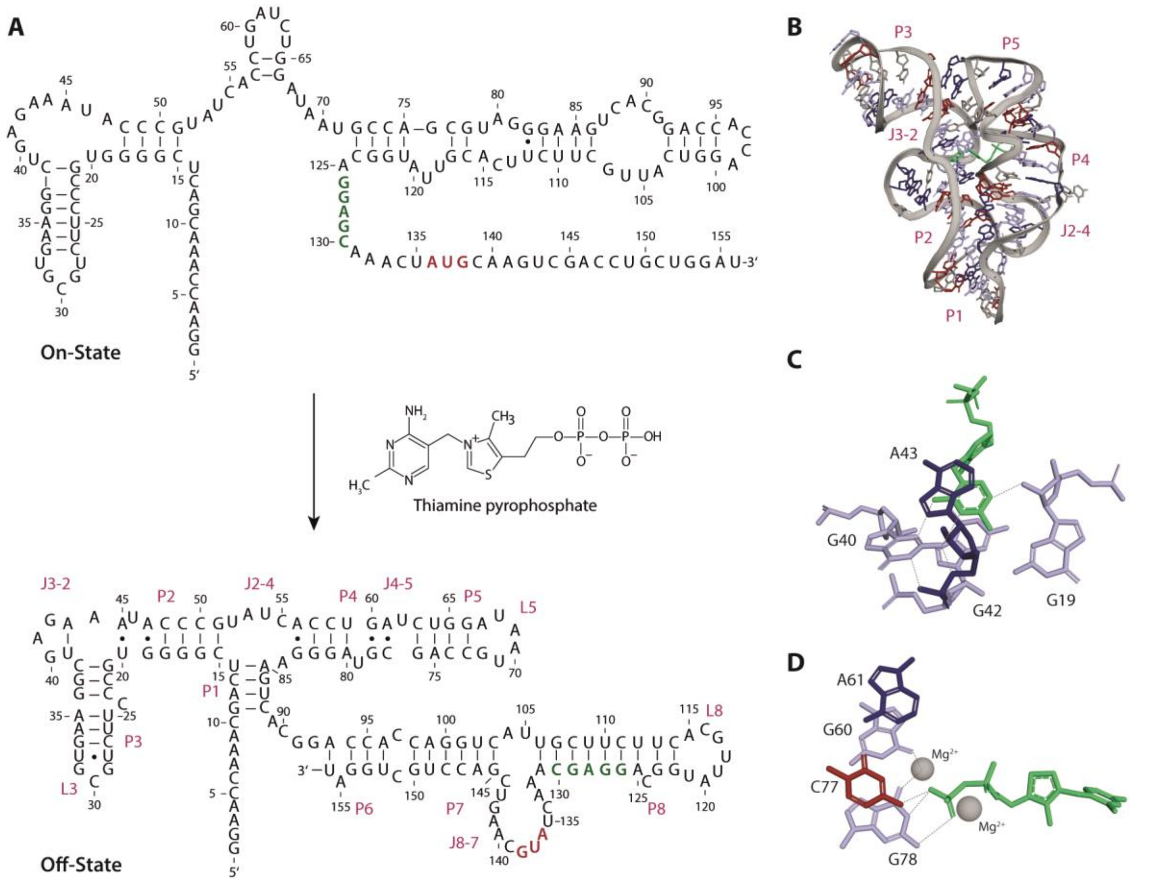


Figure 10. Thiamine pyrophosphate-sensing riboswitch thiM from *E. coli*. (A) Sequence and secondary structure of the thiM riboswitch representing the TPP-unbound and TPP-bound states. In the presence of TPP, the SD sequence (green) and the AUG start codon (red) are sequestered and are not accessible to the ribosome for translation initiation. The sequence was obtained from *E. coli* according to Bastet *et al.*[238] (B) Crystal structure of the aptamer domain of thiM (PDB: 7TDA) with bound TPP (green), Mg^{2+} and Mn^{2+}. (C) Interaction between the HMP ring of TPP and the J3-2 region of RNA. The HMP ring intercalates between G42 and A43 and further interacts with G19 and G40. (D) Interaction between the pyrophosphate group of TPP and the pyrophosphate-binding pocket of RNA. The negatively charged phosphate group is associated with divalent Mg^{2+}. The terminal phosphate of TPP is coordinated to Mg^{2+} by G60 and G78. Additionally, C77 and G78 interact with the terminal phosphate of TPP through hydrogen bonds. (Adapted from Serganov *et al.*[234])

The *E. coli* thiC riboswitch regulates gene expression by terminating transcription and inhibiting translation of phosphomethylpyrimidine synthase, which catalyzes the reaction of 5-aminoimidazole ribotide to hydroxymethyl pyrimidine phosphate.[225,239] The dual regulation of transcription and translation by TPP riboswitches occurs during mRNA synthesis of the terminator stem, resulting in the sequestration of the RBS sequence and AUG start codon located in this intrinsic terminator stem of the expression platform (Figure 11A).[240] Furthermore, Chauvier and coworkers demonstrated a regulatory pause site near the translation start codon of thiC. They reported a regulatory mechanism of thiC in which co-transcriptional binding of TPP induces RNA polymerase pausing, triggers Rho-dependent transcription termination, and inhibits TPP binding by transcription complexes located at the regulatory pause site.[241]

Thore *et al.* determined the crystal structure of the *Arabidopsis thaliana (A. thaliana)* thiC riboswitch, which showed the typical conserved secondary structure of TPP-sensing riboswitches, consisting of five helices P1 to P5 connected by junctions J3-2 and J2-4 (Figure 11B, C). Junction J3-2 recognizes the HMP ring by intercalation of the pyrimidine residue between G30 and A31 and further interactions with G11 and G28 (Figure 11D). The negatively charged pyrophosphate group is associated with Mg^{2+} and interacts through hydrogen bonding with G48, G64, G65, and G66 of the J4-5 junction (Figure 11E).[239,242]

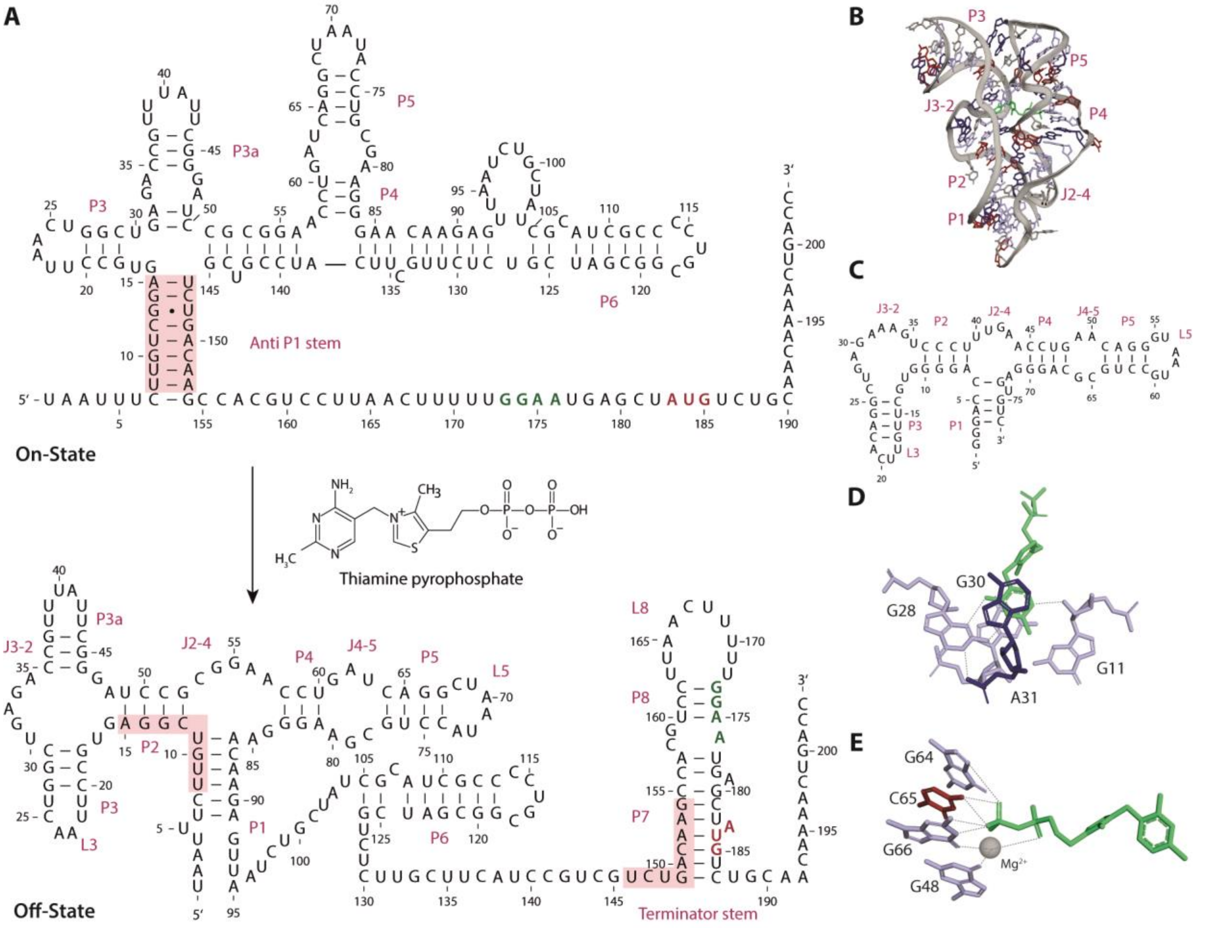


Figure 11. Thiamine pyrophosphate-sensing riboswitch thiC. (A) Sequence and secondary structure of the thiC riboswitch representing the TPP-unbound and TPP-bound states. In the presence of TPP, the SD sequence (green) and the AUG start codon (red) are sequestered and inaccessible to the ribosome for translation initiation. The sequence was obtained from *E. coli* according to Chauvier *et al.*[241] (B) Crystal structure of the TPP-sensing riboswitch thiC from *A. thaliana* (PDB: 3D2G) with bound TPP and Mg^{2+}. (C) Sequence and secondary structure of the TPP-bound aptamer domain of the *A. thaliana* thiC riboswitch. The sequence was used for crystallization. (D) Interaction between the HMP ring of TPP and the J3-2 region of RNA. The HMP ring intercalates between G30 and A31 and further interacts with G11 and G28. (E) Interaction between the pyrophosphate group of TPP and the pyrophosphate binding pocket of RNA. The negatively charged phosphate group is associated with the divalent Mg^{2+}. The terminal phosphate of TPP is coordinated to Mg^{2+} through hydrogen bonds with G48, G64, C65, and G66. (Adapted from Thore *et al.*[239])

3.3 Purine-sensing riboswitches

The classes of purine-sensing riboswitches include hypoxanthine, 2'-deoxyguanosine, 7-aminomethyl-7-deazaguanine (preQ$_1$), adenine- (Asw), and guanine-sensing (Gsw) riboswitches.[243] Gsw and Asw share similar and phylogenetically conserved aptamer domain structures. Both switches selectively recognize their cognate ligand through Watson-Crick base pair formation with a cytidine or uridine residue, respectively, resulting from a single mutation in the conserved aptamer sequence. Nevertheless, Gsw and Asw bind different cognate ligands with high specificity and regulate gene expression at the level of transcription and translation.[244]

The Gsw regulates gene expression of the *xpt-pbuX* operon of *B. subtillis* by selectively binding guanine with a binding affinity of 5 nM to the aptamer. Transcriptionally regulated gene expression by guanine binding is mediated by terminator hairpin formation, which inhibits transcription of the gene in the holo-state (Figure 12A).[243] Serganov *et al.* demonstrated the characteristic aptamer architecture as a three-way junction including the P1 stem, P2/L2, and P3/L3 hairpins (**Fehler! Verweisquelle konnte nicht gefunden werden.**Figure 12B).[245] The Watson-Crick interaction between the ligand guanine and the cytidine residue (C74) was confirmed by X-ray crystallization and NMR analysis.[245,246] Guanine also interacts with U22, U47, and U51 via hydrogen bonding (Figure 12C).

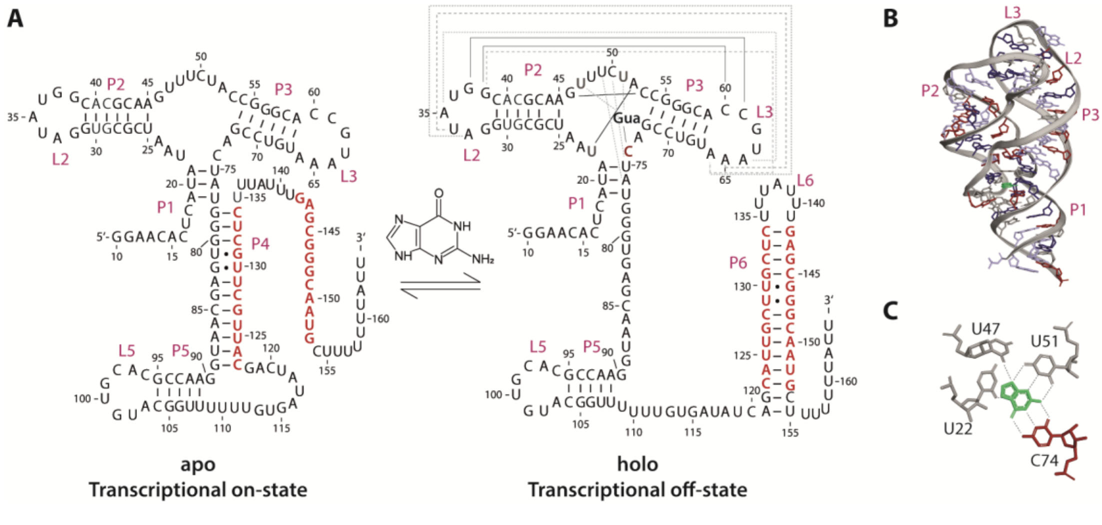


Figure 12. Guanine-sensing riboswitch from *B. subtilis*. (A) Sequence and secondary structure of *xpt-pbuX* Gsw representing guanine-unbound (apo) and guanine-bound (holo) states. In the absence of guanine, the terminator stem was repressed, resulting in a transcriptional on-state. Upon guanine binding, the transcriptional terminator stem is formed, which inhibits transcription. The tertiary Watson-Crick and non-canonical base pair interactions are indicated by solid and dashed lines, respectively. (Adapted from Steinert *et al.*[247]) (B) Crystal structure of the guanine-bound (green) Gsw-aptamer (PDB: 1Y27). (C) Guanine binds to C74 via Watson-Crick base pairing. Additionally, guanine interacts with U22, U47, and U51 of the Gsw-aptamer via hydrogen bonding. (Adapted from Serganov *et al.*[245])

While purine riboswitch aptamers are highly conserved, their regulatory mechanisms vary widely. In particular, the *B. subtilis* Asw *pbuE* regulates gene expression at the transcriptional level, whereas the *Vibrio vulnificus* (*V. vulnificus*) *add* riboswitch controls gene expression as a translational on-switch.[248,249]

The *B. subtilis pbuE* (also called *ydhL*) riboswitch encoding for the purine efflux pump regulates a transcriptional on-switch upon adenine binding by modulating a transcriptional anti-terminator sequence in its expression platform.[244] The binding affinity for adenine (K_D = 300 nM) is weaker than that of 2,6-diaminopurine to *pbuE* mRNA (K_D = 10 nM).[245] The *pbuE* riboswitch is regulated under kinetic control where adenine is co-transcriptionally bound while the mRNA is transcribed and folded (Figure 13A). Lemay *et al.* showed that the elongation factor NusA reduces the transcription rate and thus represents an additional factor in the regulation mechanism of *pbuE* in addition to transcription elongation rate and transcription pausing.[248] Delfosse *et al.* revealed the crystal structure for the U65C mutant of *pbuE* Asw (Figure 13B). They showed Watson-Crick base pairing between G48 and C74 and further hydrogen bonds to U22, U47, and U51, which is similar to the ligand binding of Gsw (Figure 13C).

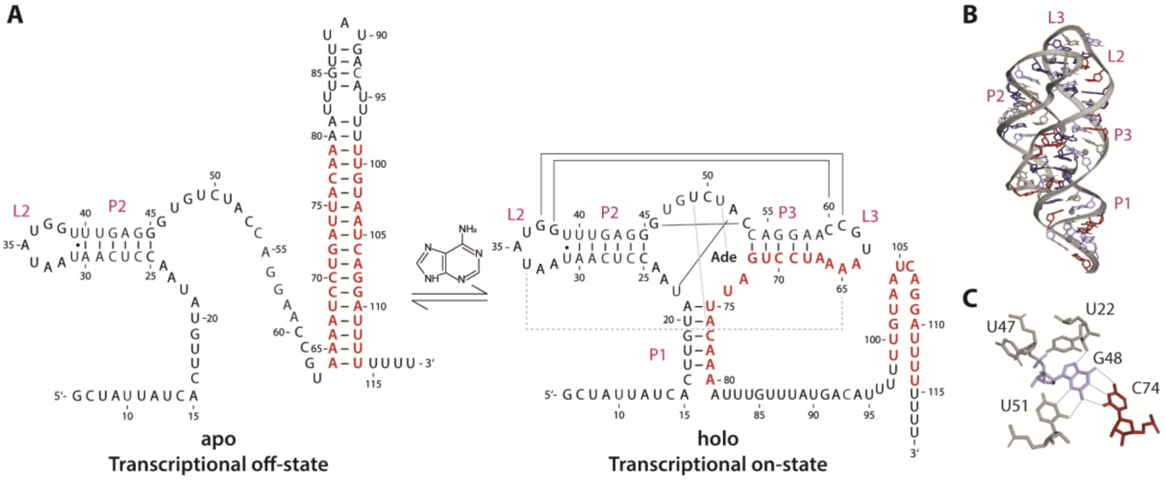


Figure 13. Adenine-sensing riboswitch from *B. subtilis*. (A) Sequence and secondary structure of *pbuE* Asw representing the adenine-unbound (apo) and adenine-bound (holo) states.[245] In the absence of adenine, the terminator hairpin is formed, resulting in a transcriptional off-state. Adenine binding to the aptamer forms the anti-terminator hairpin, allowing transcription. The Watson-Crick and non-canonical base pair tertiary interactions are indicated by solid and dashed lines, respectively. (B) Crystal structure of the U65C mutant (according to the sequence in **Fehler! Verweisquelle konnte nicht gefunden werden.**A: U74C) adenine-bound (green) Asw-aptamer (PDB: 3IVN). (C) In the U74C mutant (originally: U65C), Asw-aptamer G48 binds to C74 through Watson-Crick base pairing and further interacts with U22, U47, and U51. (Adapted from Delfosse *et al.*[250])

The adenine-sensing riboswitch of the human pathogenic Gram-negative bacterium *V. vulnificus* is located in the 5'-UTR of the *add* gene and regulates gene expression of the downstream adenosine deaminase (*add*) at the level of translational initiation.[245] *V. vulnificus* is found in marine environments, including brackish ponds, estuaries, and coastal regions, and is highly adaptable to changes in its environment with respect to the temperature, salinity, and nutrient availability.[251]

The *add* Asw is a translational on-switch in which adenine binds to the aptamer domain and induces unmasking of the RBS containing the SD sequence (GAAG) and start codon (AUG). This results in accessibility to the ribosome and initiation of translation. The 30S subunit of the ribosome associates with the SD sequence of the mRNA, and its anti-SD sequence is located in the 16S ribosomal RNA (rRNA).

Rieder *et al.* performed biophysical studies using stopped-flow fluorescence spectroscopy with a 2-aminopurine (2-AP)-modified 71 nt *add* Asw aptamer domain and with a 111 nt full-length Asw construct. They proposed a thermodynamic control mechanism based on translational activation, as the aptamer is conserved even in the presence of the expression domain.[249] The previously suggested thermodynamic control of the *add* Asw was further validated by Lemay and coworkers. Structural changes in the aptamer and expression platform of Asw upon adenine binding were detected by 2'-hydroxyl acylation using primer extension (SHAPE) analysis. Their partial nuclease digestion assay with ribonuclease T1 (RNase T1) demonstrated adenine- and Mg^{2+}-induced folding of the aptamer, resulting in an accessible SD sequence and start codon for translation initiation. In addition, cell-free *in vitro* and *in vivo* (*E. coli*) β-galactosidase assays were performed with riboswitch constructs fused to the lacZ reporter gene. The coupled cell-free data, where transcription and translation occur, showed a 2-fold increase in protein production in the presence of 500 µM adenine, whereas the uncoupled translation assay performed with *in vitro* transcribed *add* Asw RNA showed a 3-fold increase in protein expression. The *in vivo* assays confirmed the translational regulatory nature of *add* Asw, as protein expression was increased 3-fold in the presence of 500 µM adenine, while the transcriptional construct was unaffected upon adenine addition.[248] Further i*n vivo* fluorescence assays in *E. coli* were performed by Dixon *et al.* using *add* Asw constructs fused to enhanced GFP (eGFP) as a reporter protein and reported ~10-fold increase in eGFP expression for *add* Asw upon 500 µM adenine addition.[252]

Reining *et al.* demonstrated a temperature- and Mg^{2+}-dependent three-state switching mechanism for the *add* Asw (112 nt) using NMR spectroscopy. The two ligand-free (apo) conformations, apoA (adenine-binding-competent conformation) and apoB (adenine-binding-incompetent conformation), are linked by a temperature-dependent equilibrium (K_{Pre}) that allows the *add* Asw to function over a wide temperature range (10-37°C). An increase in temperature from 10°C to 30°C changed the apoA population from 12% to 40% in an Mg^{2+}-independent manner. In apoB, an alternative secondary structure was determined in the 5' region containing a helix-bulge-helix motif capped by a loop and with helices P3, P4, and P5 formed, but without the ligand-binding three-way junction, resulting in a ligand-binding-incompetent conformation. Both apo conformations were shown to be functional off-states, in apoA only the SD sequence was sequestered and in apoB both the SD sequence and the start codon were sequestered. In the presence of adenine, apoA switches to the ligand-bound on-state (holo), which is associated with a binding equilibrium (dissociation constant K_D) with stabilizing effects by Mg^{2+} at all temperatures, as validated by NMR, isothermal titration calorimetry (ITC), and stopped-flow fluorescence experiments (Figure 14A). Furthermore, the switching efficiencies were determined by comparing a two-state and a three-state Asw regulation mechanism, and resulted in a switching efficiency of 14% at 5°C for the two-state model, while higher switching efficiencies were obtained for the three-state model with, 67% switching efficiency at 5°C and 83% at 30°C.[15]

The combined NMR and single-molecule Förster resonance energy transfer (smFRET) spectroscopy analysis of full-length *add* Asw (112 nt), together with stabilized apoA and apoB mutants, revealed a functional on-state character of apoA with an open expression platform, which was previously reported as an off-state by Reining *et al*. In addition, smFRET data revealed heterogeneous aptamer kissing loop motifs of apoA and holo conformations.[16] The X-ray crystal structure was revealed by Serganov *et al.* who showed a similar X-ray crystal structure and adenine ligand-interaction within the aptamer of the *add* Asw from *V. vulnificus* compared to the *xpt-pbuX* Gsw from *B. subtilis* (Figure 14B, C).[245]

Recently, Qureshi *et al.* demonstrated by NMR spectroscopy and activity assays that the six-domain ribosomal protein S1 (rS1) from *V. vulnificus*, facilitates translation initiation of the structured *add* Asw mRNA by chaperoning and destabilizing its secondary structure. The two N-terminal domains (D1-D2) of rS1 are responsible for ribosome binding via protein-protein interactions with the ribosomal protein S2, whereas the C-terminal domains (D3-D6) interact with the RNA.[17] Furthermore, *in vivo* and *in vitro* biochemical assays as well as NMR spectroscopy studies demonstrated that the presence of adenine alone is not sufficient to induce structural changes to a translational on-switch. The combined incorporation of adenine and 30S ribosome including rS1 is essential for the accessibility of the translation initiation site of 127 nt *add* Asw.[253]

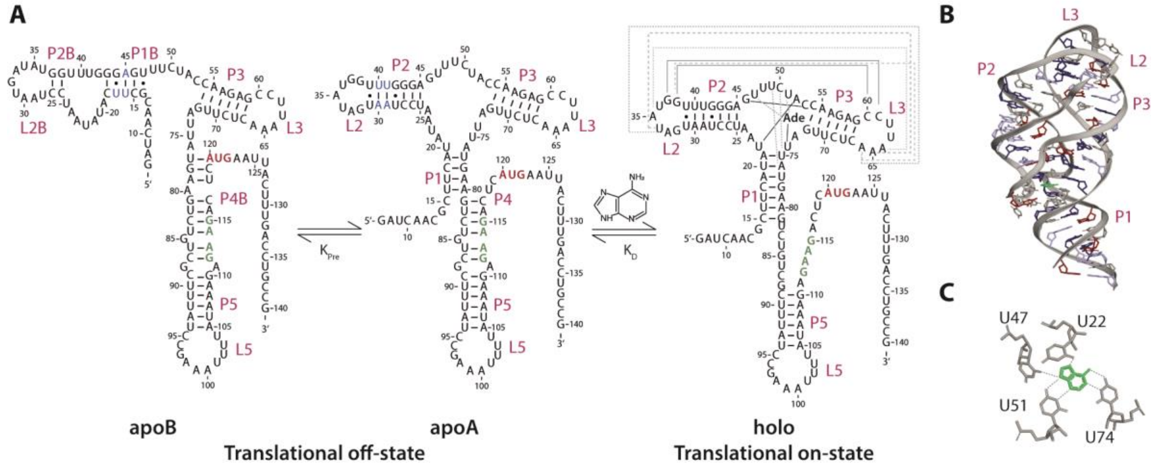


Figure 14. Adenine-sensing riboswitch from *V. vulnificus*. (A) Sequence and secondary structure of *add* Asw representing the three-state regulation mechanism with adenine-unbound (apo) and adenine-bound (holo) states. Upon adenine binding, the SD sequence and start codon are released leading to translation initiation. Watson-Crick and non-canonical tertiary base pair interactions are indicated by solid and dashed lines, respectively. (Adapted from Reining *et al*.[15]) (B) Crystal structure of adenine-bound (green) Asw in the presence of Mg^{2+} (PDB: 1Y26). (C) Adenine binds to U74 through Watson-Crick base pairing and interacts with U22, U47, and U51. (Adapted from Serganov *et al.*[245])

CHAPTER IV

Fluorescent RNA aptamers

4.1 Fluorescent light-up RNA aptamers

Fluorescent light-up aptamers are structured RNA elements that specifically bind to fluorophores and increase their fluorescence by several orders of magnitude. Both components are non-fluorescent in their unbound state. Binding of the fluorophore to the binding pocket of the RNA aptamer results in stabilization of the planar structure of the fluorophore with enhanced fluorescence activity. (Figure 15).[254]

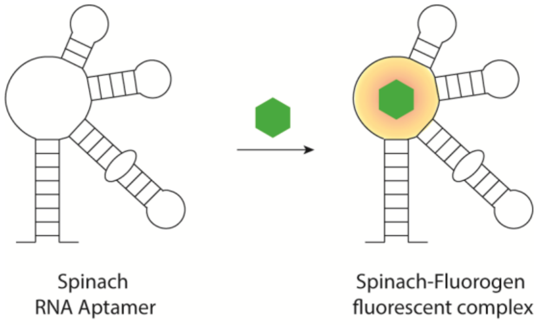


Figure 15. General principle of fluorescent light-up RNA aptamers. RNA aptamers and fluorophores are not fluorescent. In its unbound state, the fluorophore can dissipate excitation energy through non-radiative decay pathways, such as heat resulting from molecular motion. Upon binding of the fluorophore to the RNA aptamer, the planar structure of the fluorogen is stabilized, resulting in radiative decay pathways such as fluorescence and phosphorescence with increased fluorescence. (Adapted from Ouellet[18])

Since no naturally fluorescent RNAs are known, these fluorescent RNA aptamers were identified using a highly effective technique called systematic evolution of ligands by exponential enrichment (SELEX).[255] Here, the RNA sequences of interest are selected from a pool of random RNA sequences based on their affinity for the target molecules. This selection process is repeated until an aptamer with the sub-micromolar affinity to its target ligand is identified.[256] Ideal fluorophores are small, non-fluorescent ligands that are cell-permeable without cytotoxic and phototoxic properties. RNA aptamer-fluorophore complexes have been found for a variety of fluorogens with different excitation and emission characteristics (Figure 16).[257]

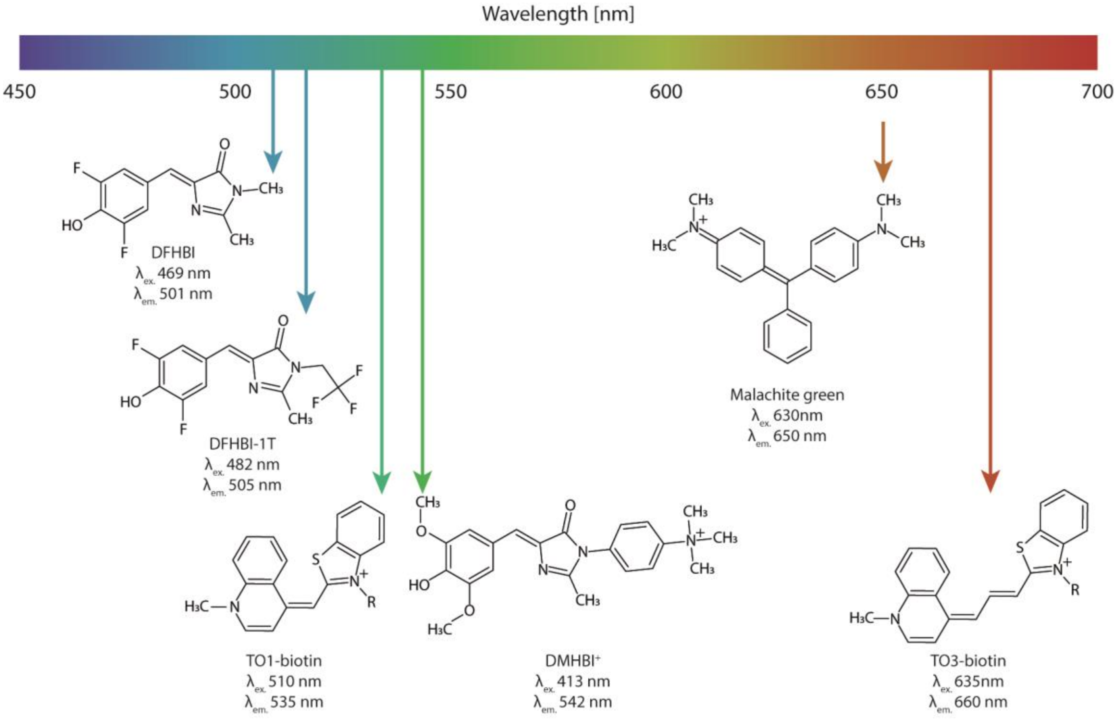


Figure 16. Overview of the fluorophores that bind to light-up RNA aptamers. The fluorophores are positioned according to their excitation and emission wavelengths for the RNA aptamer-fluorophore complexes. In the TO1-biotin and TO3-biotin chemical structures, R represents the biotin linker. (Adapted from Trachman and Ferré-D'Amaré[257])

The malachite green aptamer was the first light-up aptamer developed using SELEX.[258] However, a major drawback of the malachite green ligand is its high reactivity, which induces free radicals upon laser irradiation, resulting in degradation of labeled transcripts. Due to cytotoxicity, malachite green aptamers are not suitable for *in vivo* RNA imaging.[259] The structure of the malachite green aptamer, consisting of two helices (P1, P2), the binding pocket, and the tetraloop, was determined by solution NMR spectroscopy (Figure 17A, B). Flinders *et al.* showed that the malachite green fluorophore interacts with the RNA aptamer by intercalation of the aromatic rings of the ligand between the G8-C28 base pair and the RNA quadruple formed by the nucleotides C7, G24, G29, and A31 (Figure 17C).[260]

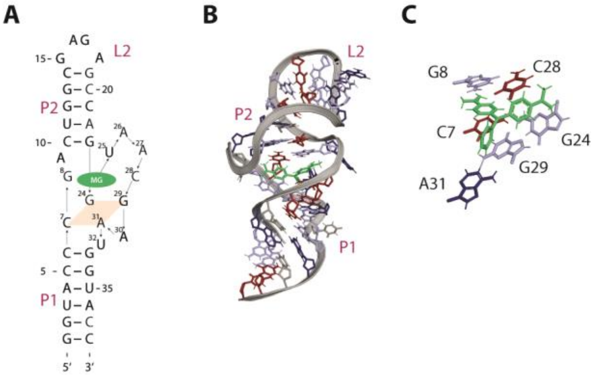


Figure 17. Malachite green RNA aptamer. (A) Sequence and secondary structure of the ligand-bound malachite green RNA aptamer (green sphere) in the binding pocket. (B) Solution NMR structure of the malachite green-bound (green) RNA aptamer (PDB: 1Q8N). (C) Malachite green intercalates between the base quadruple formed by nucleotides C7, G24, G29, and A31 as well as the base pairs G8 and C28. (Adapted from Flinders *et al.*[260])

In addition, fluorophores that are bound by RNA light-up aptamers are similar to the GFP protein-binding fluorophore 4-hydroxybenzylidene imidazolinone (HBI). The fluorophore is fluorescent when incorporated into the structure of GFP due to autocatalytic intramolecular cyclization of the amino acids Ser65-Tyr66-Gly67.[261] In particular, new aptamers have been identified, namely the Spinach RNA aptamer that binds the fluorophores 3,5-dimethoxy-4-hydroxybenzylidene imidazolinone (DMHBI) and 3,5-difluoro-4-hydroxybenzylidene imidazolinone (DFHBI). Paige *et al.* reported a 20-fold fluorescence enhancement of DFHBI upon binding to the 98 nt long Spinach aptamer (Figure 18A).[254] Warner and coworkers solved the crystal structure of the DFHBI-bound Spinach RNA aptamer, which is composed of three helices (P1, P2, P3) connected by two irregular junctions (J1-2, J2-3) (Figure 18B). The fluorophore DFHBI is immobilized in the J2-3 junction between a base triplet, a G-quadruplex, and an unpaired G in the presence of K^+ (Figure 18C).[262]

In further studies, modified and improved Spinach RNA aptamers, such as Spinach2, Baby-Spinach, and iSpinach variants have been developed. The 96 nt long Spinach2 aptamer binds to the cell-permeable DFHBI-1T, a trifluoroethyl derivative of DFHBI, which exhibits improved folding, photo- and thermal stability with the potential for *in vivo* and *in vitro* applications (Figure 18D).[263] Baby-Spinach (51 nt) is a compact version of the Spinach RNA aptamer with similar fluorescence properties, consisting of the fluorophore-binding region, specifically the junction bulge J2-3 (Figure 18E).[262] Autour *et al.* demonstrated the iSpinach RNA aptamer (69 nt) with improved folding, higher DFHBI binding affinity ($K_D^{iSpinach}$ ~ 100 nM determined at 25°C and in the presence of K^+) and 1.4-fold higher fluorescence in the presence of K^+ compared to Spinach2.[264]

The Broccoli RNA aptamer (49 nt) was designed by a combined SELEX and fluorescence-activated cell sorting (FACS) approach, which selected the aptamers that enhanced the fluorescence of the screened fluorophores and functioned in a cellular environment (Figure 18F).[265–267] They obtained improved thermostability and fluorescence properties for the Broccoli RNA aptamer compared to the Spinach2 RNA aptamer upon DFHBI and DFHBI-1T binding. The improved folding in the presence of low cytosolic Mg^{2+} concentrations enabled the Broccoli RNA aptamer for *in vivo* applications.[268]

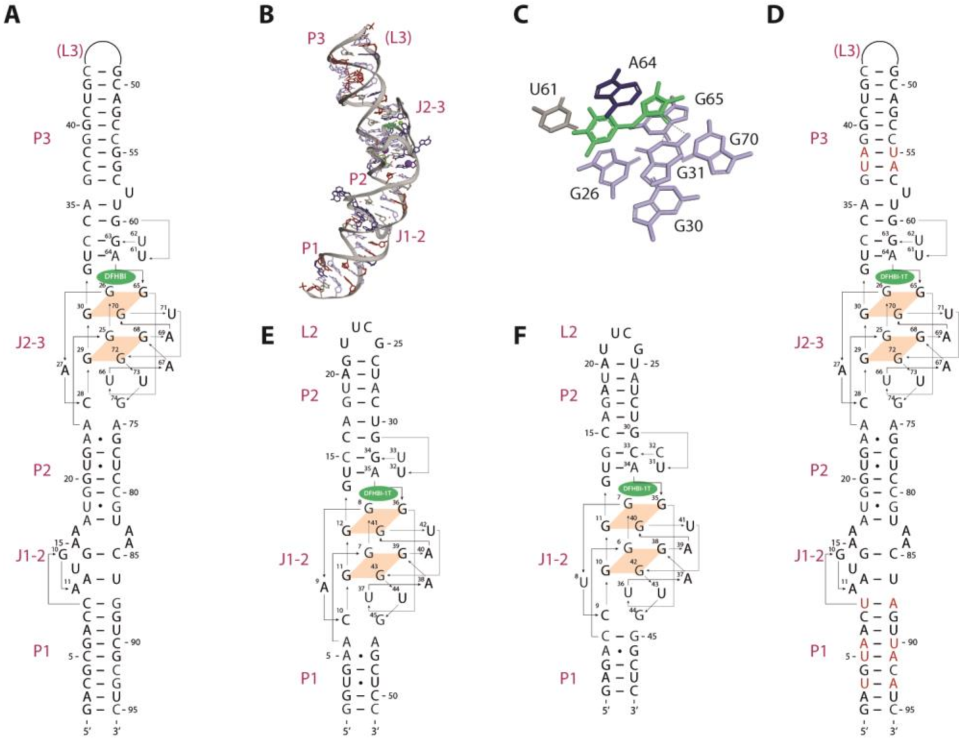


Figure 18. Overview of Spinach and Broccoli RNA aptamers. (A) Sequence and secondary structure of Spinach RNA aptamer with DFHBI ligand (green sphere). (B) Crystal structure of Spinach RNA aptamer in the presence of DFHBI ligand and K^+ (purple sphere) (PDB: 4TS2). (C) DFHBI intercalates between the base quadruple formed by nucleotides G26, G30, G65, and G70 as well as the base pairs U61 and A64. Additionally, the unpaired G31 interacts with the DFHBI carbonyl oxygen.[262] (D) Sequence and secondary structure of Spinach2 RNA aptamer with DFHBI-1T ligand (green sphere). The changed nucleotides compared to the Spinach RNA aptamer are marked in red. (E) Sequence and secondary structure of Baby Spinach RNA aptamer with DFHBI-1T ligand (green sphere). (F) Sequence and secondary structure of the Broccoli RNA aptamer with DFHBI-1T ligand (green sphere).[265] (Adapted from Ouellet[18])

4.2 Mango RNA aptamers

The Mango(I)-RNA aptamer was selected from a diverse and random sequence pool by Dolgosheina *et al.* to bind and enhance the fluorescence of thiazole orange 1-biotin (TO1-biotin) with nanomolar affinity (K_D ~ 3 nM).[269] The co-crystal structure of the Mango-(I) RNA aptamer and TO1-biotin complex revealed a three-tiered (T1, T2, and T3) G-quadruplex with T1 and T2 in parallel and T3 in anti-parallel connectivity. The ligand TO1-biotin intercalates between the ligand-binding core consisting of U15, A20, A25, and T3 of the G-quadruplex (Figure 19A, B). Novel Mango-RNA aptamers were designed to bind TO1-biotin and TO3-biotin using SELEX and droplet-based microfluidic fluorescence screening. TO3-biotin contains two additional carbons in the methine bridge of TO1-biotin, which consists of a benzothiazole ring covalently linked to a quinoline ring by a monomethine bridge.[269,270] Autour *et al.* developed Mango-(II), Mango-(III), and Mango-(IV) RNA aptamers with of 1.5-, 4-, and 3-fold improved fluorescence, respectively, compared to Mango(I) (Table 1). Binding affinities of TO1-biotin were demonstrated for Mango-(II) with sub-nanomolar affinity (K_D ~ 1 nM), whereas Mango-(III) (K_D ~ 6 nM) and Mango-(IV) (K_D ~ 11 nM) showed weaker binding affinities. The highest binding affinity and improved fluorescence properties in the presence of TO3-biotin were demonstrated for Mango-(II) ($K_D = 1.8 \pm 0.1$ nM) and Mango-(IV) ($K_D = 10.4 \pm 0.1$ nM). Mango-(III) binds to TO3-biotin with the lowest binding affinity ($K_D = 15.0 \pm 1.3$ nM) (Table 1).

The Mango-(II)-RNA aptamer forms a three-tiered G-quadruplex structure and provides enhanced thermal stability and high affinity for TO1-biotin binding due to the insertion of A15 between G14 and G16, resulting in a structural rearrangement of T3 together with the dinucleotide adenines A15 and A25 in this region (Figure 19C, D). Further, a resistance to formaldehyde fixation was demonstrated, which is useful for *in vivo* applications.[271] Circular dichroism (CD) spectroscopy and thermal melting assays demonstrated increased folding stability of Mango-(II) and Mango-(IV) compared to Mango-(I) in the presence of Mg^{2+}.[270]

The brightest Mango-(III) RNA aptamer forms a two-tiered G-quadruplex structure with an all-parallel connectivity next to the anti-parallel G18 in tier T2. The quadruplex stacks co-axially on a base triplet, which further stacks on the A-form duplex in the presence of K^+. TO1-biotin intercalates between the G-quadruplex and the Watson-Crick base pair of the Mango-(III) RNA aptamer (Figure 19E, F).[257,272]

Mango-(IV) RNA aptamer has been used in mammalian cells due to its high brightness *in vivo*.[270] Trachman *et al.* determined the crystal structure of the Mango-(IV) RNA aptamer in complex with TO1-biotin as a domain-swapped homodimer. Mango-(IV) RNA aptamer forms a three-tiered G-quadruplex with T1 and T2 in parallel and T3 in anti-parallel connectivity. Five unpaired nucleobases surround the fluorophore binding pocket of the G-quadruplex and do not stack on top of the fluorophore, resulting in an open binding pocket. The open binding pocket character allowed Mango-(II) and Mango-(IV) RNA aptamers to bind to fluorophores containing propenate (three-carbon) conjugated linkers such as TO3-biotin and TO3-acetate (**Fehler! Verweisquelle konnte nicht gefunden werden.**G, H).[257,273]

Table 1. Binding and fluorescence properties of the Mango RNA aptamer complexes.

RNA aptamer	**TO1-biotin**		**TO3-biotin**	
	K_D [nM] [(a)]	**ε ($M^{-1}*cm^{-1}$) [(b)]**	**K_D [nM] [(a)]**	**ε ($M^{-1}*cm^{-1}$)**
Mango-(I)	2.2 ± 0.3	10,850	5.1 ± 0.3	9,300 [(c)]
Mango-(II)	0.7 ± 0.3	17,000	1.8 ± 0.1	15,800 [(d)]
Mango-(III)	5.6 ± 0.2	43,000	15.0 ± 1.3	10,600 [(d)]
Mango-(IV)	11.1 ± 0.8	32,000	10.4 ± 0.1	17,400 [(d)]

(a) Binding coefficients from Autour *et al.*[270]
(b) Brightness of the TO1-biotin bound complexes from Trachman *et al.*[257]
(c) Brightness of the TO3-biotin bound Mango-(I) RNA aptamer from Dolgosheina *et al.*[269]
(d) Brightness of the TO3-biotin bound complexes determined from Autour *et al.*[270]

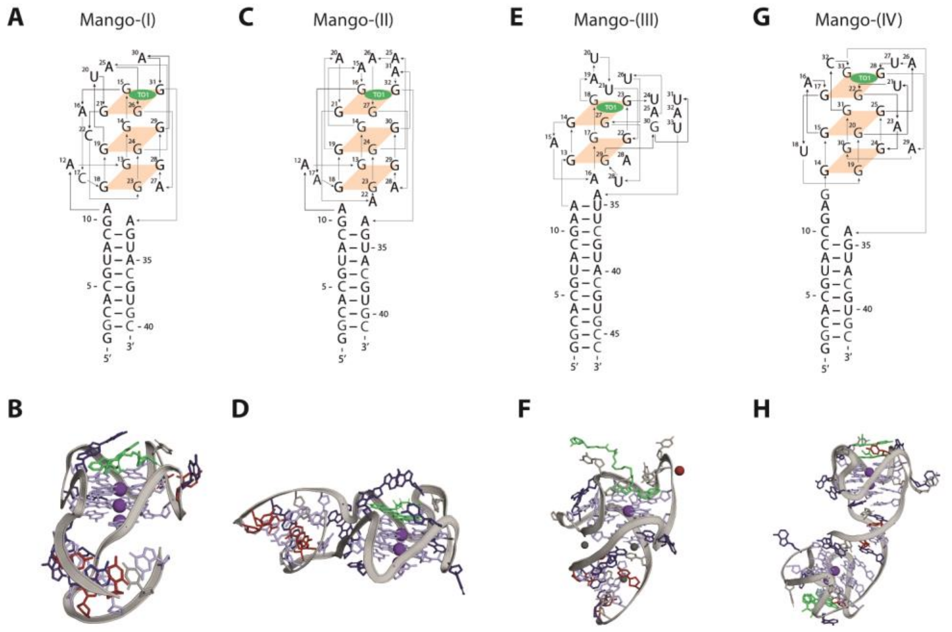


Figure 19. Overview of Mango RNA aptamers. (A) Sequence and secondary structure of Mango-(I) RNA aptamer with the TO1-biotin ligand (green sphere). (B) Crystal structure of the Mango-(I) RNA aptamer in the presence of the TO1-biotin ligand (green) and K^+ (purple sphere) (PDB: 5V3F).[274] (C) Sequence and secondary structure of Mango-(II) RNA aptamer with the TO1-biotin ligand (green sphere). (D) Crystal structure of the Mango-(II) RNA aptamer in the presence of the TO1-biotin ligand (green) and K^+ (purple sphere) (PDB: 6C65).[275] (E) Sequence and secondary structure of Mango-(III) RNA aptamer with the TO1-biotin ligand (green sphere). (F) Crystal structure of the Mango-(III) RNA aptamer in the presence of the TO1-biotin ligand (green), K^+ (purple sphere), Na^+ (red sphere), and Mg^{2+} (gray sphere) (PDB: 6E8U).[272] (G) Sequence and secondary structure of Mango-(IV) RNA aptamer with the TO1-biotin ligand (green sphere). (H) Crystal structure of the homodimer Mango-(IV) RNA aptamer in the presence of the ligand TO1-biotin (green) and K^+ (purple sphere) (PDB: 6V9B).[273]

4.3 Applications of fluorescent light-up aptamers

Fluorescent RNA aptamers are suitable for a wide range of *in vitro* and *in vivo* applications, especially for the monitoring of less abundant RNAs. The *in vitro* approaches with different fluorescent RNA aptamers include the RNA visualization in native or denaturing gels as a cost- and time-saving alternative to Northern blotting using the cognate ligand of the applied fluorescent RNA aptamer. For Spinach2, Broccoli or Mango RNA aptamers, the gel is incubated in the presence of DFHBI-1T or TO1-biotin and appropriate K^+ concentrations to achieve correct folding of the RNA aptamer.[276] With this approach, RNA concentrations of ~ 1 fmol could be detected, similar to the detection limit of SYBR Gold staining.[267]

The combination of riboswitch and fluorescent RNA aptamer sequences has allowed the development of *in vitro* assays to characterize riboswitch functions or to screen and identify ligands of orphan riboswitches.[18] To date, most riboswitches have been studied at the protein-level, where the binding of metabolites activates or inhibits the expression of fluorescent proteins.[18,255] High-throughput screening methods have been developed to identify new activators or inhibitors of RNA-modifying enzymes. Svensen and Jaffrey demonstrated that the m^{6}A-methylated Broccoli RNA aptamer remained non-fluorescent, although fluorescence activity was restored in the presence of ALKBH5 RNA demethylase, a protein associated with fat mass and obesity.[277]

The split-Spinach RNA aptamer probe approach has been used for fluorescence monitoring of ribozyme activity, RNA assembly, and even functional imaging of viral genome trafficking.[278] Real-time *in vitro* fluorescence monitoring of self-cleaving hammerhead ribozymes was performed using a ribozyme sequence linked to a split-Spinach RNA aptamer sequence, where the full-length Spinach RNA aptamer sequence is split into two non-functional halves. Upon ribozyme cleavage, the split-Spinach sequence is released and can bind to its complementary split-Spinach RNA aptamer sequence, and upon DHBI-1T binding, the RNA aptamer exhibits fluorescence.[279] Furthermore, the split-Spinach RNA aptamer approach has also been demonstrated to target RNA of interest using two vectors encoding the split-Spinach RNA aptamers, which were expressed and bound to the target RNA and formed a fluorescent complex after ligand binding.[255]

The metabolite-binding aptamer domain of riboswitches and fluorescent light-up aptamer fusions offered advantageous opportunities to establish protein-free metabolite sensors to investigate the dynamics, abundance, and flux of intracellular molecules *in vivo*. This technique allowed real-time monitoring of several metabolites, including adenosine, guanine, adenosine-5'-diphosphate (ADP), guanosine-5'-triphosphate (GTP), S-adenosylmethionine (SAM)[280], cyclic-di-GMP, and cyclic-AMP-GMP[281]. In addition to metabolite detection using Spinach RNA aptamers, the real-time imaging of bacterial proteins in *E. coli* was established by Song and coworkers. These sensors were designed with a protein-binding aptamer domain linked to the Spinach RNA aptamer domain by a transducer sequence. Upon protein binding, the secondary structure of the protein-binding aptamer domain changes to allow Watson-Crick base pairing of the transducer stem sequence, which induces Spinach RNA aptamer formation and DFHBI binding. The protein concentrations of thrombin, MS2 phage coat protein, and streptavidin were initially determined in *in vitro* experiments and later in *E. coli* cells.[282]

Single-molecule RNA imaging has been used to understand *in vivo* RNA dynamics during various processes such as transcription, post-transcriptional modification, translation, splicing and degradation. Cawte *et al.* tracked β-actin mRNA and the long non-coding RNA NEAT1 using single-molecule imaging with Mango (II)-RNA aptamers. The folding and fluorescence properties of Mango (II)-RNA aptamer were improved using tandem array constructs with multiple aptamers linked together in a row.[271]

References

1. Cooper, T. A., Wan, L. & Dreyfuss, G. RNA and Disease. *Cell* **136**, 777–793 (2009).

2. Sharp, P. A. The centrality of RNA. *Cell* **136**, 577–80 (2009).

3. Zhang, P., Wu, W., Chen, Q. & Chen, M. Non-Coding RNAs and their Integrated Networks. *J Integr Bioinform* **16**, (2019).

4. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**, 155–9 (2009).

5. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol* **22**, 96–118 (2021).

6. Li, Y., Syed, J. & Sugiyama, H. RNA-DNA Triplex Formation by Long Noncoding RNAs. *Cell Chem Biol* **23**, 1325–1333 (2016).

7. Grote, P. & Herrmann, B. G. The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. *RNA Biol* **10**, 1579–85 (2013).

8. Grote, P. *et al.* The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev Cell* **24**, 206–14 (2013).

9. Warwick, T., Brandes, R. P. & Leisegang, M. S. Computational Methods to Study DNA:DNA:RNA Triplex Formation by lncRNAs. *Non-Coding RNA 2023, Vol. 9, Page 10* **9**, 10 (2023).

10. Focke, P. J. *et al.* Combining in Vitro Folding with Cell Free Protein Synthesis for Membrane Protein Expression. *Biochemistry* **55**, 4212–4219 (2016).

11. Gregorio, N. E., Levine, M. Z. & Oza, J. P. A User's Guide to Cell-Free Protein Synthesis. *Methods Protoc* **2**, 24 (2019).

12. Mccown, P. J., Corbino, K. A., Stav, S., Sherlock, M. E. & Breaker, R. R. Riboswitch diversity and distribution. *RNA* **23**, 995–1011 (2017).

13. Breaker, R. R. Riboswitches and the RNA world. *Cold Spring Harb Perspect Biol* **4**, (2012).

14. Mandal, M. & Breaker, R. R. Gene regulation by riboswitches. *Nature Reviews Molecular Cell Biology 2004 5:6* **5**, 451–463 (2004).

15. Reining, A. *et al.* Three-state mechanism couples ligand and temperature sensing in riboswitches. *Nature* **499**, 355–9 (2013).

16. Warhaut, S. *et al.* Ligand-modulated folding of the full-length adenine riboswitch probed by NMR and single-molecule FRET spectroscopy. *Nucleic Acids Res* **45**, 5512–5522 (2017).

17. Qureshi, N. S., Bains, J. K., Sreeramulu, S., Schwalbe, H. & Fürtig, B. Conformational switch in the ribosomal protein S1 guides unfolding of structured RNAs for translation initiation. *Nucleic Acids Res* **46**, 10917–10929 (2018).

18. Ouellet, J. RNA Fluorescence with Light-Up Aptamers. *Front Chem* **4**, 29 (2016).

19. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and Functions of Long Noncoding RNAs. *Cell* **136**, 629–641 (2009).

20. Crick, F. Central Dogma of Molecular Biology. *Nature 1970 227:5258* **227**, 561–563 (1970).

21. Hargarten, J. C. & Williamson, P. R. Epigenetic regulation of autophagy: A path to the control of autoimmunity. *Front Immunol* **9**, 1864 (2018).

22. Yin, H. *et al.* The Therapeutic and Pathogenic Role of Autophagy in Autoimmune Diseases. *Front Immunol* **9**, 1512 (2018).

23. Haemmig, S., Simion, V., Yang, D., Deng, Y. & Feinberg, M. W. Long noncoding RNAs in cardiovascular disease, diagnosis, and therapy. *Current Opinion in Cardiology* vol. 32 776–783 Preprint at https://doi.org/10.1097/HCO.0000000000000454 (2017).

24. Zhou, M., Guo, X., Wang, M. & Qin, R. The patterns of antisense long non-coding RNAs regulating corresponding sense genes in human cancers. *J Cancer* **12**, 1499–1506 (2021).

25. Niu, T., Zhang, W. & Xiao, W. MicroRNA regulation of cancer stem cells in the pathogenesis of breast cancer. *Cancer Cell Int* **21**, 1–14 (2021).

26. Uchida, S. & Dimmeler, S. Long noncoding RNAs in cardiovascular diseases. *Circ Res* **116**, 737–50 (2015).

27. Jean-François, L., Derghal, A. & Mounien, L. MicroRNAs in Obesity and Related Metabolic Disorders. *Cells* **8**, (2019).

28. Wijesinghe, S. N., Nicholson, T., Tsintzas, K. & Jones, S. W. Involvements of long noncoding RNAs in obesity-associated inflammatory diseases. *Obesity Reviews* **22**, e13156 (2021).

29. Oo, J. A., Brandes, R. P. & Leisegang, M. S. Long non-coding RNAs: novel regulators of cellular physiology and function. *Pflugers Arch* **474**, 191–204 (2022).

30. Vieira, A. S., Dogini, D. B. & Lopes-Cendes, I. Role of non-coding RNAs in non-aging-related neurological disorders. *Brazilian Journal of Medical and Biological Research* **51**, (2018).

31. Higgs, P. G. RNA secondary structure: physical and computational aspects. *Q Rev Biophys* **33**, 199–253 (2000).

32. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–8 (1953).

33. Westhof, E. & Fritsch, V. RNA folding: beyond Watson–Crick pairs. *Structure* **8**, R55–R65 (2000).

34. Leontis, N. B. & Westhof, E. Geometric nomenclature and classification of RNA base pairs. *RNA* **7**, 499–512 (2001).

35. Leontis, N. B., Stombaugh, J. & Westhof, E. The non-Watson–Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res* **30**, 3497–3531 (2002).

36. Varani, G. & McClain, W. H. The G·U wobble base pair. *EMBO Rep* **1**, 18–23 (2000).

37. Ladner, J. E. *et al.* Structure of yeast phenylalanine transfer RNA at 2.5 A resolution. *Proceedings of the National Academy of Sciences* **72**, 4414–4418 (1975).

38. Scott, W. G., Finch, J. T. & Klug, A. The crystal structure of an AII-RNAhammerhead ribozyme: A proposed mechanism for RNA catalytic cleavage. *Cell* **81**, 991–1002 (1995).

39. Griffin, B. D. & Bass, H. W. Review: Plant G-quadruplex (G4) motifs in DNA and RNA; abundant, intriguing sequences of unknown function. *Plant Science* **269**, 143–147 (2018).

40. Millevoi, S., Moine, H. & Vagner, S. G-quadruplexes in RNA biology. *Wiley Interdiscip Rev RNA* **3**, 495–507 (2012).

41. Yonkunas, M. J. & Baird, N. J. A highly ordered, nonprotective MALAT1 ENE structure is adopted prior to triplex formation. *RNA* **25**, 975–984 (2019).

42. Brown, J. A. *et al.* Structural insights into the stabilization of MALAT1 noncoding RNA by a bipartite triple helix. *Nature Structural & Molecular Biology 2014 21:7* **21**, 633–640 (2014).

43. Halder, S. & Bhattacharyya, D. RNA structure and dynamics: A base pairing perspective. *Prog Biophys Mol Biol* **113**, 264–283 (2013).

44. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–45 (2004).

45. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics 2015 47:3* **47**, 199–208 (2015).

46. Matera, A. G., Terns, R. M. & Terns, M. P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol* **8**, 209–20 (2007).

47. Liang, X. H., Liu, L. & Michaeli, S. Identification of the First Trypanosome H/ACA RNA That Guides Pseudouridine Formation on rRNA. *Journal of Biological Chemistry* **276**, 40313–40318 (2001).

48. Bachellerie, J. P., Cavaillé, J. & Hüttenhofer, A. The expanding snoRNA world. *Biochimie* **84**, 775–790 (2002).

49. Ojha, S., Malla, S. & Lyons, S. M. snoRNPs: Functions in Ribosome Biogenesis. *Biomolecules* **10**, (2020).

50. Hombach, S. & Kretz, M. Non-coding RNAs: Classification, Biology and Functioning. *Adv Exp Med Biol* **937**, 3–17 (2016).

51. Argetsinger Steitz, J. & Jakes, K. How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli. *Proceedings of the National Academy of Sciences* **72**, 4734–4738 (1975).

52. Bartel, D. P. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* **136**, 215–233 (2009).

53. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science (1979)* **294**, 853–858 (2001).

54. Carthew, R. W. & Sontheimer, E. J. Origins and Mechanisms of miRNAs and siRNAs. *Cell* vol. 136 642–655 Preprint at https://doi.org/10.1016/j.cell.2009.01.035 (2009).

55. Cech, T. R. & Steitz, J. A. The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. *Cell* **157**, 77–94 (2014).

56. Sabin, L. R., Delás, M. J. & Hannon, G. J. Dogma derailed: the many influences of RNA on the genome. *Mol Cell* **49**, 783–94 (2013).

57. Ross, R. J., Weiner, M. M. & Lin, H. PIWI proteins and PIWI-interacting RNAs in the soma. *Nature* **505**, 353–359 (2014).

58. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature 2003 420:6915* **420**, 563–573 (2002).

59. Maeda, N. *et al.* Transcript Annotation in FANTOM3: Mouse Gene Catalog Based on Physical cDNAs. *PLoS Genet* **2**, e62 (2006).

60. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766–D773 (2019).

61. Kung, J. T. Y., Colognori, D. & Lee, J. T. Long noncoding RNAs: past, present, and future. *Genetics* **193**, 651–69 (2013).

62. Wang, K. C. & Chang, H. Y. Molecular mechanisms of long noncoding RNAs. *Mol Cell* **43**, 904–14 (2011).

63. Martianov, I., Ramadass, A., Serra Barros, A., Chow, N. & Akoulitchev, A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445**, 666–70 (2007).

64. Tay, Y., Rinn, J. & Pandolfi, P. P. The multilayered complexity of ceRNA crosstalk and competition. *Nature 2014 505:7483* **505**, 344–352 (2014).

65. Cheng, J.-T. *et al.* Insights into Biological Role of LncRNAs in Epithelial-Mesenchymal Transition. *Cells* **8**, 1178 (2019).

66. Buske, F. A., Mattick, J. S. & Bailey, T. L. Potential in vivo roles of nucleic acid triple-helices. *RNA Biol* **8**, 427–39 (2011).

67. Felsenfeld, G., Davies, D. R. & Rich, A. Formation of a Three-Stranded Polynucleotide Molecule. *J Am Chem Soc* **79**, 2023–2024 (1957).

68. Moser, H. E. & Dervan, P. B. Sequence-specific cleavage of double helical DNA by triple helix formation. *Science* **238**, 645–50 (1987).

69. Olivas, W. M. & Maher, L. J. Competitive triplex/quadruplex equilibria involving guanine-rich oligonucleotides. *Biochemistry* **34**, 278–84 (1995).

70. de los Santos, C., Rosen, M. & Patel, D. NMR Studies of DNA (R+)n·(Y-)n·(Y+)n Triple Helices in Solution: Imino and Amino Proton Markers of T·A·T and C·G·C+ Base-Triple Formation. *Biochemistry* **28**, 7282–7289 (1989).

71. Jain, A., Wang, G. & Vasquez, K. M. DNA triple helices: biological consequences and therapeutic potential. *Biochimie* **90**, 1117–30 (2008).

72. Sandström, K., Wärmländer, S., Gräslund, A. & Leijon, M. A-tract DNA disfavours triplex formation. *J Mol Biol* **315**, 737–48 (2002).

73. Roberts, R. W. & Crothers, D. M. Stability and properties of double and triple helices: dramatic effects of RNA or DNA backbone composition. *Science* **258**, 1463–6 (1992).

74. Escudé, C. *et al.* Stability of triple helices containing RNA and DNA strands: experimental and molecular modeling studies. *Nucleic Acids Res* **21**, 5547–53 (1993).

75. Kunkler, C. N. *et al.* Stability of an RNA•DNA–DNA triple helix depends on base triplet composition and length of the RNA third strand. *Nucleic Acids Res* **47**, 7213–7222 (2019).

76. Wang, Y. & Patel, D. J. Bulge defects in intramolecular pyrimidine.purine.pyrimidine DNA triplexes in solution. *Biochemistry* **34**, 5696–704 (1995).

77. Fox, K. R., Flashman, E. & Gowers, D. Secondary binding sites for triplex-forming oligonucleotides containing bulges, loops, and mismatches in the third strand. *Biochemistry* **39**, 6714–25 (2000).

78. Gowers, D. M. & Fox, K. R. DNA triple helix formation at oligopurine sites containing multiple contiguous pyrimidines. *Nucleic Acids Res* **25**, 3787–94 (1997).

79. Mergny, J. L. *et al.* Sequence specificity in triple-helix formation: experimental and theoretical studies of the effect of mismatches on triplex stability. *Biochemistry* **30**, 9791–8 (1991).

80. Vasquez, K. M., Wensel, T. G., Hogan, M. E. & Wilson, J. H. High-affinity triple helix formation by synthetic oligonucleotides at a site within a selectable mammalian gene. *Biochemistry* **34**, 7243–51 (1995).

81. James, P. L., Brown, T. & Fox, K. R. Thermodynamic and kinetic stability of intermolecular triple helices containing different proportions of C+*GC and T*AT triplets. *Nucleic Acids Res* **31**, 5598–606 (2003).

82. Gilbert, S. D., Rambo, R. P., van Tyne, D. & Batey, R. T. Structure of the SAM-II riboswitch bound to S-adenosylmethionine. *Nat Struct Mol Biol* **15**, 177–82 (2008).

83. Huang, L. & Lilley, D. M. J. Structure and ligand binding of the SAM-V riboswitch. *Nucleic Acids Res* **46**, 6869–6879 (2018).

84. Yan, C. *et al.* Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science (1979)* **349**, 1182–1191 (2015).

85. Theimer, C. A., Blois, C. A. & Feigon, J. Structure of the Human Telomerase RNA Pseudoknot Reveals Conserved Tertiary Interactions Essential for Function. *Mol Cell* **17**, 671–682 (2005).

86. Toscano-Garibay, J. D. & Aquino-Jarquin, G. Transcriptional regulation mechanism mediated by miRNA–DNA•DNA triplex structure stabilized by Argonaute. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1839**, 1079–1083 (2014).

87. Bacolla, A. *et al.* Long homopurine*homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. *Nucleic Acids Res* **34**, 2663–75 (2006).

88. Zain, R. & Sun, J.-S. Do natural DNA triple-helical structures occur and function in vivo? *Cell Mol Life Sci* **60**, 862–70 (2003).

89. Lyamichev, V. I., Mirkin, S. M. & Frank-Kamenetskii, M. D. A pH-dependent structural transition in the homopurine-homopyrimidine tract in superhelical DNA. *J Biomol Struct Dyn* **3**, 327–38 (1985).

90. Panyutin, I. G., Kovalsky, O. I. & Budowsky, E. I. Magnesium-dependent supercoiling-induced transition in (dG)n.(dC)n stretches and formation of a new G-structure by (dG)n strand. *Nucleic Acids Res* **17**, 8257–71 (1989).

91. Karlovsky, P., Pecinka, P., Vojtiskova, M., Makaturova, E. & Palecek, E. Protonated triplex DNA in E. coli cells as detected by chemical probing. *FEBS Lett* **274**, 39–42 (1990).

92. Ussery, D. W. & Sinden, R. R. Environmental influences on the in vivo level of intramolecular triplex DNA in Escherichia coli. *Biochemistry* **32**, 6206–13 (1993).

93. Kohwi, Y., Malkhosyan, S. R. & Kohwi-Shigematsu, T. Intramolecular dG.dG.dC triplex detected in Escherichia coli cells. *J Mol Biol* **223**, 817–22 (1992).

94. Kato, M. & Shimizu, N. Effect of the potential triplex DNA region on the in vitro expression of bacterial beta-lactamase gene in superhelical recombinant plasmids. *J Biochem* **112**, 492–4 (1992).

95. Brahmachari, S. K., Sarkar, P. S., Raghavan, S., Narayan, M. & Maiti, A. K. Polypurine/polypyrimidine sequences as cis-acting transcriptional regulators. *Gene* **190**, 17–26 (1997).

96. Rao, B. S. Regulation of DNA replication by homopurine/homopyrimidine sequences. *Mol Cell Biochem* **156**, 163–8 (1996).

97. Mitton-Fry, R. M., DeGregorio, S. J., Wang, J., Steitz, T. A. & Steitz, J. A. Poly(A) tail recognition by a viral RNA element through assembly of a triple helix. *Science* **330**, 1244–7 (2010).

98. Holland, J. A. & Hoffman, D. W. Structural features and stability of an RNA triple helix in solution. *Nucleic Acids Res* **24**, 2841–8 (1996).

99. Qiao, F. & Cech, T. R. Triple-helix structure in telomerase RNA contributes to catalysis. *Nat Struct Mol Biol* **15**, 634–40 (2008).

100. Theimer, C. A. & Feigon, J. Structure and function of telomerase RNA. *Curr Opin Struct Biol* **16**, 307–18 (2006).

101. Chen, G., Chang, K.-Y., Chou, M.-Y., Bustamante, C. & Tinoco, I. Triplex structures in an RNA pseudoknot enhance mechanical stability and increase efficiency of -1 ribosomal frameshifting. *Proc Natl Acad Sci U S A* **106**, 12706–11 (2009).

102. Duca, M., Vekhoff, P., Oussedik, K., Halby, L. & Arimondo, P. B. The triple helix: 50 years later, the outcome. *Nucleic Acids Res* **36**, 5123–38 (2008).

103. Buske, F. A., Bauer, D. C., Mattick, J. S. & Bailey, T. L. Triplexator: Detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome Res* **22**, 1372–1381 (2012).

104. Kuo, C.-C. *et al.* Detection of RNA-DNA binding sites in long noncoding RNAs. *Nucleic Acids Res* **47**, e32 (2019).

105. Lee, J. S., Burkholder, G. D., Latimer, L. J., Haug, B. L. & Braun, R. P. A monoclonal antibody to triplex DNA binds to eucaryotic chromosomes. *Nucleic Acids Res* **15**, 1047–61 (1987).

106. AGAZIE, Y. M., BURKHOLDER, G. D. & LEE, J. S. Triplex DNA in the nucleus: direct binding of triplex-specific antibodies and their effect on transcription, replication and cell growth. *Biochemical Journal* **316**, 461–466 (1996).

107. Lubitz, I., Zikich, D. & Kotlyar, A. Specific high-affinity binding of thiazole orange to triplex and G-quadruplex DNA. *Biochemistry* **49**, 3567–74 (2010).

108. Dixon, B. P., Lu, L., Chu, A. & Bissler, J. J. RecQ and RecG helicases have distinct roles in maintaining the stability of polypurine·polypyrimidine sequences. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* **643**, 20–28 (2008).

109. Schmitz, K.-M., Mayer, C., Postepska, A. & Grummt, I. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev* **24**, 2264–9 (2010).

110. Maldonado, R., Schwartz, U., Silberhorn, E. & Längst, G. Nucleosomes Stabilize ssRNA-dsDNA Triple Helices in Human Cells. *Mol Cell* **73**, 1243-1254.e6 (2019).

111. Aguilera, A. & García-Muse, T. R loops: from transcription byproducts to threats to genome stability. *Mol Cell* **46**, 115–24 (2012).

112. Scaria, P. v., Will, S., Levenson, C. & Shafer, R. H. Physicochemical Studies of the d(G3T4G3)∗d(G3A4G3)•d(C3T4C3) Triple Helix. *Journal of Biological Chemistry* **270**, 7295–7303 (1995).

113. Jalali, S., Singh, A., Maiti, S. & Scaria, V. Genome-wide computational analysis of potential long noncoding RNA mediated DNA:DNA:RNA triplexes in the human genome. *J Transl Med* **15**, 186 (2017).

114. Manzini, G. *et al.* Triple helix formation by oligopurine-oligopyrimidine DNA fragments. Electrophoretic and thermodynamic behavior. *J Mol Biol* **213**, 833–43 (1990).

115. Tateishi-Karimata, H., Nakano, M. & Sugimoto, N. Comparable Stability of Hoogsteen and Watson–Crick Base Pairs in Ionic Liquid Choline Dihydrogen Phosphate. *Scientific Reports 2014 4:1* **4**, 1–7 (2014).

116. Radhakrishnan, I., Gao, X., de los Santos, C., Live, D. & Patel, D. J. NMR structural studies of intramolecular (Y+)n.(R+)n(Y-)nDNA triplexes in solution: imino and amino proton and nitrogen markers of G.TA base triple formation. *Biochemistry* **30**, 9022–30 (1991).

117. Radhakrishnan, I. & Patel, D. J. Solution structure and hydration patterns of a Pyrimidine·Purine·Pyrimidine DNA triplex containing a novel T·CG base-triple. *J Mol Biol* **241**, 600–619 (1994).

118. Radhakrishnan, I., de los Santos, C. & Patel, D. J. Nuclear magnetic resonance structural studies of intramolecular purine · purine · pyrimidine DNA triplexes in solution. *J Mol Biol* **221**, 1403–1418 (1991).

119. Sørensen, J. J., Nielsen, J. T. & Petersen, M. Solution structure of a dsDNA:LNA triplex. *Nucleic Acids Res* **32**, 6078–6085 (2004).

120. Mukherjee, A. & Vasquez, K. M. Triplex technology in studies of DNA damage, DNA repair, and mutagenesis. *Biochimie* **93**, 1197 (2011).

121. Vasquez, K. M. & Glazer, P. M. Triplex-forming oligonucleotides: principles and applications. *Q Rev Biophys* **35**, 89–107 (2002).

122. Xodo, L. E., Manzini, G., Quadrifoglio, F., van der Marel, G. A. & van Boom, J. H. Effect of 5-methylcytosine on the stability of triple-stranded DNA--a thermodynamic study. *Nucleic Acids Res* **19**, 5625–31 (1991).

123. Krawczyk, S. H. *et al.* Oligonucleotide-mediated triple helix formation using an N3-protonated deoxycytidine analog exhibiting pH-independent binding within the physiological range. *Proc Natl Acad Sci U S A* **89**, 3761–4 (1992).

124. Rao, T. S. *et al.* Incorporation of 2’-deoxy-6-thioguanosine into G-rich oligodeoxyribonucleotides inhibits G-tetrad formation and facilitates triplex formation. *Biochemistry* **34**, 765–72 (1995).

125. Milligan, J. F., Krawazyk, S. H., Wadwani, S. & Matteucci, M. D. An anti-parallel triple helix motif with oligodeoxynucleotidescontaining 2’-deoxyguanosine and 7-deaza-2’-deoxy xanthosine. *Nucleic Acids Res* **21**, 327–333 (1993).

126. Inoue, H. *et al.* Synthesis and hybridization studies on two complementary nona(2’-O-methyl)ribonucleotides. *Nucleic Acids Res* **15**, 6131–48 (1987).

127. Escudé, C., Sun, J. S., Rougée, M., Garestier, T. & Hélène, C. Stable triple helices are formed upon binding of RNA oligonucleotides and their 2’-O-methyl derivatives to double-helical DNA. *C R Acad Sci III* **315**, 521–5 (1992).

128. Sproat, B. S., Lamond, A. I., Beijer, B., Neuner, P. & Ryder, U. Highly efficient chemical synthesis of 2’-O-methyloligoribonucleotides and tetrabiotinylated derivatives; novel probes that are resistant to degradation by RNA or DNA specific nucleases. *Nucleic Acids Res* **17**, 3373–3386 (1989).

129. Wang, S. & Kool, E. T. Relative stabilities of triple helices composed of combinations of DNA, RNA and 2’-O-methyl-RNA backbones: chimeric circular oligonucleotides as probes. *Nucleic Acids Res* **23**, 1157–64 (1995).

130. Blommers, M. J., Natt, F., Jahnke, W. & Cuenoud, B. Dual recognition of double-stranded DNA by 2’-aminoethoxy-modified oligonucleotides: the solution structure of an intramolecular triplex obtained by NMR spectroscopy. *Biochemistry* **37**, 17714–25 (1998).

131. Cuenoud, B. *et al.* Dual Recognition of Double-Stranded DNA by 2’-Aminoethoxy-Modified Oligonucleotides. *Angew Chem Int Ed Engl* **37**, 1288–1291 (1998).

132. Seidman, M. M. *et al.* The development of bioactive triple helix-forming oligonucleotides. *Ann N Y Acad Sci* **1058**, 119–27 (2005).

133. Sørensen, M. D., Meldgaard, M., Raunkjaer, M., Rajwanshi, V. K. & Wengel, J. Branched oligonucleotides containing bicyclic nucleotides as branching points and DNA or LNA as triplex forming branch. *Bioorg Med Chem Lett* **10**, 1853–6 (2000).

134. Sun, B.-W. *et al.* Sequence and pH effects of LNA-containing triple helix-forming oligonucleotides: physical chemistry, biochemistry, and modeling studies. *Biochemistry* **43**, 4160–9 (2004).

135. Brunet, E. *et al.* Exploring cellular activity of locked nucleic acid-modified triplex-forming oligonucleotides and defining its molecular basis. *J Biol Chem* **280**, 20076–85 (2005).

136. Frieden, M., Hansen, H. F. & Koch, T. Nuclease stability of LNA oligonucleotides and LNA-DNA chimeras. *Nucleosides Nucleotides Nucleic Acids* **22**, 1041–3 (2003).

137. Torigoe, H., Hari, Y., Sekiguchi, M., Obika, S. & Imanishi, T. 2’-O,4’-C-methylene bridged nucleic acid modification promotes pyrimidine motif triplex DNA formation at physiological pH: thermodynamic and kinetic studies. *J Biol Chem* **276**, 2354–60 (2001).

138. Koizumi, M. *et al.* Triplex formation with 2’-O,4’-C-ethylene-bridged nucleic acids (ENA) having C3’-endo conformation at physiological pH. *Nucleic Acids Res* **31**, 3267–73 (2003).

139. Brunet, E. *et al.* Exploring cellular activity of locked nucleic acid-modified triplex-forming oligonucleotides and defining its molecular basis. *J Biol Chem* **280**, 20076–85 (2005).

140. Lacoste, J., François, J. C. & Hélène, C. Triple helix formation with purine-rich phosphorothioate-containing oligonucleotides covalently linked to an acridine derivative. *Nucleic Acids Res* **25**, 1991–8 (1997).

141. Escudé, C. *et al.* Stable triple helices formed by oligonucleotide N3’-->P5’ phosphoramidates inhibit transcription elongation. *Proc Natl Acad Sci U S A* **93**, 4365–9 (1996).

142. Basye, J., Trent, J. O., Gao, D. & Ebbinghaus, S. W. Triplex formation by morpholino oligodeoxyribonucleotides in the HER-2/neu promoter requires the pyrimidine motif. *Nucleic Acids Res* **29**, 4873–80 (2001).

143. Nielsen, P. E. & Egholm, M. Strand displacement recognition of mixed adenine-cytosine sequences in double stranded DNA by thymine-guanine PNA (peptide nucleic acid). *Bioorg Med Chem* **9**, 2429–34 (2001).

144. Hanvey, J. C. *et al.* Antisense and antigene properties of peptide nucleic acids. *Science* **258**, 1481–5 (1992).

145. Hu, J. & Corey, D. R. Inhibiting gene expression with peptide nucleic acid (PNA)--peptide conjugates that target chromosomal DNA. *Biochemistry* **46**, 7581–9 (2007).

146. Rich, A. A hybrid helix containing both deoxyribose and ribose polynucleotides and its relation to the transfer of information between the nucleic acids. *Proc Natl Acad Sci U S A* **46**, 1044–53 (1960).

147. Hall, B. D. & Spiegelman, S. Sequence complementarity of T2-DNA and T2-specific RNA. *Proc Natl Acad Sci U S A* **47**, 137–63 (1961).

148. Westover, K. D., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: separation of RNA from DNA by RNA polymerase II. *Science* **303**, 1014–6 (2004).

149. Lesnik, E. A. & Freier, S. M. Relative thermodynamic stability of DNA, RNA, and DNA:RNA hybrid duplexes: relationship with base composition and structure. *Biochemistry* **34**, 10807–15 (1995).

150. Shaw, N. N. & Arya, D. P. Recognition of the unique structure of DNA:RNA hybrids. *Biochimie* **90**, 1026–39 (2008).

151. Postepska-Igielska, A. *et al.* LncRNA Khps1 Regulates Expression of the Proto-oncogene SPHK1 via Triplex-Mediated Changes in Chromatin Structure. *Mol Cell* **60**, 626–36 (2015).

152. Xia, P. *et al.* An oncogenic role of sphingosine kinase. *Current Biology* **10**, 1527–1530 (2000).

153. Blank-Giwojna, A., Postepska-Igielska, A. & Grummt, I. lncRNA KHPS1 Activates a Poised Enhancer by Triplex-Dependent Recruitment of Epigenomic Regulators. *Cell Rep* **26**, 2904-2915.e4 (2019).

154. Bastiaan Holwerda, S. J. & de Laat, W. CTCF: the protein, the binding partners, the binding sites and their chromatin loops. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**, (2013).

155. O'Leary, V. B. *et al.* PARTICLE, a Triplex-Forming Long ncRNA, Regulates Locus-Specific Methylation in Response to Low-Dose Irradiation. *Cell Rep* **11**, 474–85 (2015).

156. Mondal, T. *et al.* MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures. *Nat Commun* **6**, 7743 (2015).

157. Kalwa, M. *et al.* The lncRNA HOTAIR impacts on mesenchymal stem cells via triple helix formation. *Nucleic Acids Res* **44**, 10631–10643 (2016).

158. Leisegang, M. S. *et al.* HIF1α-AS1 is a DNA:DNA:RNA triplex-forming lncRNA interacting with the HUSH complex. *Nat Commun* **13**, 6563 (2022).

159. Perez, J. G., Stark, J. C. & Jewett, M. C. Cell-Free Synthetic Biology: Engineering Beyond the Cell. *Cold Spring Harb Perspect Biol* **8**, a023853 (2016).

160. Matthaei, H. & Nirenberg, M. W. The dependence of cell-free protein synthesis in E. coli upon RNA prepared from ribosomes. *Biochem Biophys Res Commun* **4**, 404–8 (1961).

161. Nirenberg, M. W. & Matthaei, J. H. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U S A* **47**, 1588–602 (1961).

162. Chambers, D. A. & Zubay, G. The stimulatory effect of cyclic adenosine 3’5’-monophosphate on DNA-directed synthesis of beta-galactosidase in a cell-free system. *Proc Natl Acad Sci U S A* **63**, 118–122 (1969).

163. Zalkin, H., Yanofsky, C. & Squires, C. L. Regulated in vitro synthesis of Escherichia coli tryptophan operon messenger ribonucleic acid and enzymes. *Journal of Biological Chemistry* **249**, 465–475 (1974).

164. Whittaker, J. W. Cell-free protein synthesis: The state of the art. *Biotechnol Lett* **35**, 143–152 (2013).

165. Nilsson, B. L., Soellner, M. B. & Raines, R. T. Chemical synthesis of proteins. *Annu Rev Biophys Biomol Struct* **34**, 91–118 (2005).

166. Karikó, K. *et al.* Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. *Molecular Therapy* **16**, 1833–1840 (2008).

167. Kigawa, T. *et al.* Cell-free production and stable-isotope labeling of milligram quantities of proteins. *FEBS Lett* **442**, 15–19 (1999).

168. Liu, D. v., Zawada, J. F. & Swartz, J. R. Streamlining Escherichia Coli S30 Extract Preparation for Economical Cell-Free Protein Synthesis. *Biotechnol Prog* **21**, 460–465 (2008).

169. Schwarz, D. *et al.* Preparative scale expression of membrane proteins in Escherichia coli-based continuous exchange cell-free systems. *Nat Protoc* **2**, 2945–57 (2007).

170. Spirin, A. S. High-throughput cell-free systems for synthesis of functionally active proteins. *Trends Biotechnol* **22**, 538–45 (2004).

171. Rosenblum, G. & Cooperman, B. S. Engine out of the chassis: Cell-free protein synthesis and its uses. *FEBS Lett* **588**, 261–268 (2014).

172. Ohashi, H., Kanamori, T., Shimizu, Y. & Ueda, T. A Highly Controllable Reconstituted Cell-Free System -a Breakthrough in Protein Synthesis Research. *Curr Pharm Biotechnol* **11**, 267–271 (2010).

173. Jackson, A. M., Boutell, J., Cooley, N. & He, M. Cell-free protein synthesis for proteomics. *Brief Funct Genomic Proteomic* **2**, 308–19 (2004).

174. Volyanik, E. v. *et al.* Synthesis of Preparative Amounts of Biologically Active Interleukin-6 Using a Continuous-Flow Cell-Free Translation System. *Anal Biochem* **214**, 289–294 (1993).

175. Zawada, J. F. *et al.* Microscale to manufacturing scale-up of cell-free cytokine production—a new approach for shortening protein production development timelines. *Biotechnol Bioeng* **108**, 1570–1578 (2011).

176. Dondapati, S. K., Stech, M., Zemella, A. & Kubick, S. Cell-Free Protein Synthesis: A Promising Option for Future Drug Development. *BioDrugs* **34**, 327–348 (2020).

177. Shimizu, Y., Kanamori, T. & Ueda, T. Protein synthesis by pure translation systems. *Methods* **36**, 299–304 (2005).

178. Ohashi, H., Kanamori, T., Shimizu, Y. & Ueda, T. A Highly Controllable Reconstituted Cell-Free System -a Breakthrough in Protein Synthesis Research. *Curr Pharm Biotechnol* **11**, 267–271 (2010).

179. Katzen, F., Chang, G. & Kudlicki, W. The past, present and future of cell-free protein synthesis. *Trends Biotechnol* **23**, 150–6 (2005).

180. Sachse, R., Dondapati, S. K., Fenz, S. F., Schmidt, T. & Kubick, S. Membrane protein synthesis in cell-free systems: From bio-mimetic systems to bio-membranes. *FEBS Lett* **588**, 2774–2781 (2014).

181. He, M. Cell-free protein synthesis: applications in proteomics and biotechnology. *N Biotechnol* **25**, 126–132 (2008).

182. King, R. W., Lustig, K. D., Stukenberg, P. T., McGarry, T. J. & Kirschner, M. W. FUNCTIONAL GENOMICS: Expression Cloning in the Test Tube. *Science (1979)* **277**, 973–974 (1997).

183. Bundy, B. C., Franciszkowicz, M. J. & Swartz, J. R. Escherichia coli-based cell-free synthesis of virus-like particles. *Biotechnol Bioeng* **100**, 28–37 (2008).

184. Bundy, B. C. & Swartz, J. R. Efficient disulfide bond formation in virus-like particles. *J Biotechnol* **154**, 230–9 (2011).

185. Tsuboi, T. *et al.* Wheat germ cell-free system-based production of malaria proteins for discovery of novel vaccine candidates. *Infect Immun* **76**, 1702–1708 (2008).

186. Gite, S., Mamaev, S., Olejnik, J. & Rothschild, K. Ultrasensitive Fluorescence-Based Detection of Nascent Proteins in Gels. *Anal Biochem* **279**, 218–225 (2000).

187. Tabuchi, I. Next-generation protein-handling method: puromycin analogue technology. *Biochem Biophys Res Commun* **305**, 1–5 (2003).

188. Wals, K. & Ovaa, H. Unnatural amino acid incorporation in E. coli: current and future applications in the design of therapeutic proteins. *Front Chem* **2**, 15 (2014).

189. Khambhati, K. *et al.* Exploring the Potential of Cell-Free Protein Synthesis for Extending the Abilities of Biological Systems. *Front Bioeng Biotechnol* **7**, (2019).

190. Normanly, J., Kleina, L. G., Masson, J. M., Abelson, J. & Miller, J. H. Construction of Escherichia coli amber suppressor tRNA genes: III. Determination of tRNA specificity. *J Mol Biol* **213**, 719–726 (1990).

191. Ozawa, K., Dixon, N. & Otting, G. Cell-free synthesis of 15 N-labeled proteins for NMR studies. *IUBMB Life (International Union of Biochemistry and Molecular Biology: Life)* **57**, 615–622 (2005).

192. Takeda, M. & Kainosho, M. Cell-free protein production for NMR studies. *Methods in Molecular Biology* **831**, 71–84 (2012).

193. Ozawa, K., Wu, P. S. C., Dixon, N. E. & Otting, G. 15N-Labelled proteins by cell-free protein synthesis. *FEBS J* **273**, 4154–4159 (2006).

194. Guignard, L., Ozawa, K., Pursglove, S. E., Otting, G. & Dixon, N. E. NMR analysis of in vitro-synthesized proteins without purification: a high-throughput approach. *FEBS Lett* **524**, 159–162 (2002).

195. Stark, J. C. *et al.* BioBits^{TM} Bright: A fluorescent synthetic biology education kit. *Sci Adv* **4**, 33 (2018).

196. Huang, A. *et al.* Biobits^{TM} explorer: A modular synthetic biology education kit. *Sci Adv* **4**, (2018).

197. Nielsen, A. A. K., Segall-Shapiro, T. H. & Voigt, C. A. Advances in genetic circuit design: novel biochemistries, deep part mining, and precision gene expression. *Curr Opin Chem Biol* **17**, 878–92 (2013).

198. Jia, H., Heymann, M., Bernhard, F., Schwille, P. & Kai, L. Cell-free protein synthesis in micro compartments: building a minimal cell from biobricks. *N Biotechnol* **39**, 199–205 (2017).

199. Shin, J. & Noireaux, V. An E. coli cell-free expression toolbox: Application to synthetic gene circuits and artificial cells. *ACS Synth Biol* **1**, 29–41 (2012).

200. Chizzolini, F., Forlin, M., Cecchi, D. & Mansy, S. S. Gene position more strongly influences cell-free protein expression from operons than T7 transcriptional promoter strength. *ACS Synth Biol* **3**, 363–71 (2014).

201. Findeiß, S., Etzel, M., Will, S., Mörl, M. & Stadler, P. Design of Artificial Riboswitches as Biosensors. *Sensors* **17**, 1990 (2017).

202. Voyvodic, P. L. *et al.* Plug-and-play metabolic transducers expand the chemical detection space of cell-free biosensors. *Nat Commun* **10**, 1697 (2019).

203. Thavarajah, W. *et al.* Point-of-Use Detection of Environmental Fluoride via a Cell-Free Riboswitch-Based Biosensor. *ACS Synth Biol* **9**, 10–18 (2020).

204. Hunt, J. P. *et al.* Towards detection of SARS-CoV-2 RNA in human saliva: A paper-based cell-free toehold switch biosensor with a visual bioluminescent output. *N Biotechnol* **66**, 53–60 (2022).

205. Pardee, K. *et al.* Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components. *Cell* **165**, 1255–1266 (2016).

206. Pardee, K. *et al.* Paper-Based Synthetic Gene Networks. *Cell* **159**, 940–954 (2014).

207. Bédard, A.-S. V., Hien, E. D. M. & Lafontaine, D. A. Riboswitch regulation mechanisms: RNA, metabolites and regulatory proteins. *Biochim Biophys Acta Gene Regul Mech* **1863**, 194501 (2020).

208. Loh, E. *et al.* A trans-Acting Riboswitch Controls Expression of the Virulence Regulator PrfA in Listeria monocytogenes. *Cell* **139**, 770–779 (2009).

209. Breaker, R. R. Prospects for Riboswitch Discovery and Analysis. *Mol Cell* **43**, 867–879 (2011).

210. Soukup, J. K. & Soukup, G. A. Riboswitches exert genetic control through metabolite-induced conformational change. *Curr Opin Struct Biol* **14**, 344–349 (2004).

211. Sudarsan, N., Cohen-Chalamish, S., Nakamura, S., Emilsson, G. M. & Breaker, R. R. Thiamine pyrophosphate riboswitches are targets for the antimicrobial compound pyrithiamine. *Chem Biol* **12**, 1325–35 (2005).

212. Mayer, G., Raddatz, M.-S. L., Grunwald, J. D. & Famulok, M. RNA Ligands That Distinguish Metabolite-Induced Conformations in the TPP Riboswitch. *Angewandte Chemie* **119**, 563–566 (2007).

213. Serganov, A. & Nudler, E. A decade of riboswitches. *Cell* **152**, 17–24 (2013).

214. Mandal, M. *et al.* A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science (1979)* **306**, 275–279 (2004).

215. Barrick, J. E. & Breaker, R. R. The Power of Riboswitches. *Sci Am* **296**, 50–57 (2007).

216. Gusarov, I. & Nudler, E. The Mechanism of Intrinsic Transcription Termination. *Mol Cell* **3**, 495–504 (1999).

217. Yarnell, W. S. & Roberts, J. W. Mechanism of intrinsic transcription termination and antitermination. *Science* **284**, 611–5 (1999).

218. Omotajo, D., Tate, T., Cho, H. & Choudhary, M. Distribution and diversity of ribosome binding sites in prokaryotic genomes. *BMC Genomics* **16**, 1–8 (2015).

219. Laursen, B. S., Sørensen, H. P., Mortensen, K. K. & Sperling-Petersen, H. U. Initiation of Protein Synthesis in Bacteria. *Microbiology and Molecular Biology Reviews* **69**, 101–123 (2005).

220. Breaker, R. R. Riboswitches and Translation Control. *Cold Spring Harb Perspect Biol* **10**, (2018).

221. Hollands, K. *et al.* Riboswitch control of Rho-dependent transcription termination. *Proc Natl Acad Sci U S A* **109**, 5376–5381 (2012).

222. Caron, M.-P. *et al.* Dual-acting riboswitch control of translation initiation and mRNA decay. *Proceedings of the National Academy of Sciences* **109**, E3444–E3453 (2012).

223. Bastet, L. *et al.* Translational control and Rho-dependent transcription termination are intimately linked in riboswitch regulation. *Nucleic Acids Res* **45**, 7474–7486 (2017).

224. Ariza-Mateos, A., Nuthanakanti, A. & Serganov, A. Riboswitch Mechanisms: New Tricks for an Old Dog. *Biochemistry (Mosc)* **86**, 962–975 (2021).

225. Winkler, W., Nahvi, A. & Breaker, R. R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952–6 (2002).

226. Miranda-Ríos, J. The THI-box riboswitch, or how RNA binds thiamin pyrophosphate. *Structure* **15**, 259–65 (2007).

227. Miranda-Ríos, J., Navarro, M. & Soberón, M. A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc Natl Acad Sci U S A* **98**, 9736–9741 (2001).

228. Sudarsan, N., Barrick, J. E. & Breaker, R. R. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA* **9**, 644–647 (2003).

229. Kubodera, T. *et al.* Thiamine-regulated gene expression of Aspergillus oryzae thiA requires splicing of the intron containing a riboswitch-like domain in the 5′-UTR. *FEBS Lett* **555**, 516–520 (2003).

230. Ontiveros-Palacios, N. *et al.* Molecular basis of gene regulation by the THI-box riboswitch. *Mol Microbiol* **67**, 793–803 (2008).

231. Guedich, S. *et al.* Quantitative and predictive model of kinetic regulation by E. coli TPP riboswitches. *RNA Biol* **13**, 373–90 (2016).

232. Kulshina, N., Edwards, T. E. & Ferré-D'Amaré, A. R. Thermodynamic analysis of ligand binding and ligand binding-induced tertiary structure formation by the thiamine pyrophosphate riboswitch. *RNA* **16**, 186–196 (2010).

233. Edwards, T. E. & Ferré-D'Amaré, A. R. Crystal structures of the thi-box riboswitch bound to thiamine pyrophosphate analogs reveal adaptive RNA-small molecule recognition. *Structure* **14**, 1459–68 (2006).

234. Serganov, A., Polonskaia, A., Phan, A. T., Breaker, R. R. & Patel, D. J. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* **441**, 1167–71 (2006).

235. Edwards, T. E. & Ferré-D'Amaré, A. R. Crystal Structures of the Thi-Box Riboswitch Bound to Thiamine Pyrophosphate Analogs Reveal Adaptive RNA-Small Molecule Recognition. *Structure* **14**, 1459–1468 (2006).

236. Warner, K. D. *et al.* Validating fragment-based drug discovery for biological RNAs: lead fragments bind and remodel the TPP riboswitch specifically. *Chem Biol* **21**, 591–5 (2014).

237. Baird, N. J. & Ferré-D'Amaré, A. R. Idiosyncratically tuned switching behavior of riboswitch aptamer domains revealed by comparative small-angle X-ray scattering analysis. *RNA* **16**, 598–609 (2010).

238. Bastet, L. *et al.* Translational control and Rho-dependent transcription termination are intimately linked in riboswitch regulation. *Nucleic Acids Res* **45**, 7474–7486 (2017).

239. Thore, S., Leibundgut, M. & Ban, N. Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. *Science* **312**, 1208–11 (2006).

240. Rodionov, D. A., Vitreschak, A. G., Mironov, A. A. & Gelfand, M. S. Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. *J Biol Chem* **277**, 48949–59 (2002).

241. Chauvier, A. *et al.* Transcriptional pausing at the translation start site operates as a critical checkpoint for riboswitch regulation. *Nat Commun* **8**, 13892 (2017).

242. Thore, S., Frick, C. & Ban, N. Structural basis of thiamine pyrophosphate analogues binding to the eukaryotic riboswitch. *J Am Chem Soc* **130**, 8116–8117 (2008).

243. Kim, J. N. & Breaker, R. R. Purine sensing by riboswitches. *Biol Cell* **100**, 1–11 (2008).

244. Mandal, M. & Breaker, R. R. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nat Struct Mol Biol* **11**, 29–35 (2004).

245. Serganov, A. *et al.* Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem Biol* **11**, 1729–41 (2004).

246. Noeske, J. *et al.* An intermolecular base triple as the basis of ligand specificity and affinity in the guanine- and adenine-sensing riboswitch RNAs. *Proc Natl Acad Sci U S A* **102**, 1372–1377 (2005).

247. Steinert, H. *et al.* Pausing guides RNA folding to populate transiently stable RNA structures for riboswitch-based transcription regulation. *Elife* **6**, 1–18 (2017).

248. Lemay, J.-F. *et al.* Comparative Study between Transcriptionally- and Translationally-Acting Adenine Riboswitches Reveals Key Differences in Riboswitch Regulatory Mechanisms. *PLoS Genet* **7**, e1001278 (2011).

249. Rieder, R., Lang, K., Graber, D. & Micura, R. Ligand-induced folding of the adenosine deaminase A-riboswitch and implications on riboswitch translational control. *ChemBioChem* **8**, 896–902 (2007).

250. Delfosse, V. *et al.* Riboswitch structure: an internal residue mimicking the purine ligand. *Nucleic Acids Res* **38**, 2057–2068 (2010).

251. Jones, M. K. & Oliver, J. D. Vibrio vulnificus: Disease and Pathogenesis. *Infect Immun* **77**, 1723–1733 (2009).

252. Dixon, N. *et al.* Reengineering orthogonally selective riboswitches. *Proceedings of the National Academy of Sciences* **107**, 2830–2835 (2010).

253. de Jesus, V. *et al.* Switching at the ribosome: riboswitches need rProteins as modulators to regulate translation. *Nat Commun* **12**, 4723 (2021).

254. Paige, J. S., Wu, K. Y. & Jaffrey, S. R. RNA mimics of green fluorescent protein. *Science* **333**, 642–6 (2011).

255. al Mazid, M. F., Shkel, O., Kharkivska, Y. & Lee, J. S. Application of fluorescent turn-on aptamers in RNA studies. *Molecular Omics* vol. 17 483–491 Preprint at https://doi.org/10.1039/d1mo00085c (2021).

256. Stoltenburg, R., Reinemann, C. & Strehlitz, B. SELEX—A (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol Eng* **24**, 381–403 (2007).

257. Trachman, R. J. & Ferré-D'Amaré, A. R. Tracking RNA with light: selection, structure, and design of fluorescence turn-on RNA aptamers. *Q Rev Biophys* **52**, e8 (2019).

258. Grate, D. & Wilson, C. Laser-mediated, site-specific inactivation of RNA transcripts. *Proc Natl Acad Sci U S A* **96**, 6131–6136 (1999).

259. Kraus, G. A. *et al.* Fluorinated analogs of malachite green: synthesis and toxicity. *Molecules* **13**, 986–94 (2008).

260. Flinders, J. *et al.* Recognition of planar and nonplanar ligands in the malachite green-RNA aptamer complex. *Chembiochem* **5**, 62–72 (2004).

261. Chudakov, D. M., Matz, M. v, Lukyanov, S. & Lukyanov, K. A. Fluorescent proteins and their applications in imaging living cells and tissues. *Physiol Rev* **90**, 1103–63 (2010).

262. Warner, K. D. *et al.* Structural basis for activity of highly efficient RNA mimics of green fluorescent protein. *Nat Struct Mol Biol* **21**, 658–63 (2014).

263. Strack, R. L., Disney, M. D. & Jaffrey, S. R. A superfolding Spinach2 reveals the dynamic nature of trinucleotide repeat-containing RNA. *Nat Methods* **10**, 1219–1224 (2013).

264. Autour, A., Westhof, E. & Ryckelynck, M. iSpinach: a fluorogenic RNA aptamer optimized for in vitro applications. *Nucleic Acids Res* **44**, 2491–2500 (2016).

265. Okuda, M., Fourmy, D. & Yoshizawa, S. Use of Baby Spinach and Broccoli for imaging of structured cellular RNAs. *Nucleic Acids Res* **45**, 1404–1415 (2017).

266. Filonov, G. S., Moon, J. D., Svensen, N. & Jaffrey, S. R. Broccoli: rapid selection of an RNA mimic of green fluorescent protein by fluorescence-based selection and directed evolution. *J Am Chem Soc* **136**, 16299–308 (2014).

267. Filonov, G. S., Kam, C. W., Song, W. & Jaffrey, S. R. In-gel imaging of RNA processing using broccoli reveals optimal aptamer expression strategies. *Chem Biol* **22**, 649–660 (2015).

268. Filonov, G. S., Moon, J. D., Svensen, N. & Jaffrey, S. R. Broccoli: rapid selection of an RNA mimic of green fluorescent protein by fluorescence-based selection and directed evolution. *J Am Chem Soc* **136**, 16299–16308 (2014).

269. Dolgosheina, E. v *et al.* RNA mango aptamer-fluorophore: a bright, high-affinity complex for RNA labeling and tracking. *ACS Chem Biol* **9**, 2412–20 (2014).

270. Autour, A. *et al.* Fluorogenic RNA Mango aptamers for imaging small non-coding RNAs in mammalian cells. *Nat Commun* **9**, 656 (2018).

271. Cawte, A. D., Unrau, P. J. & Rueda, D. S. Live cell imaging of single RNA molecules with fluorogenic Mango II arrays. *Nat Commun* **11**, 1283 (2020).

272. Trachman, R. J. *et al.* Structure and functional reselection of the Mango-III fluorogenic RNA aptamer. *Nat Chem Biol* **15**, 472–479 (2019).

273. Trachman, R. J. *et al.* Structure-Guided Engineering of the Homodimeric Mango-IV Fluorescence Turn-on Aptamer Yields an RNA FRET Pair. *Structure* **28**, 776-785.e3 (2020).

274. Trachman, R. J. *et al.* Structural basis for high-affinity fluorophore binding and activation by RNA Mango. *Nat Chem Biol* **13**, 807–813 (2017).

275. Trachman, R. J. *et al.* Crystal Structures of the Mango-II RNA Aptamer Reveal Heterogeneous Fluorophore Binding and Guide Engineering of Variants with Improved Selectivity and Brightness. *Biochemistry* **57**, 3544–3548 (2018).

276. Yaseen, I. M., Ang, Q. R. & Unrau, P. J. Fluorescent Visualization of Mango-tagged RNA in Polyacrylamide Gels via a Poststaining Method. *JoVE (Journal of Visualized Experiments)* **2019**, e59112 (2019).

277. Svensen, N. & Jaffrey, S. R. Fluorescent RNA Aptamers as a Tool to Study RNA-Modifying Enzymes. *Cell Chem Biol* **23**, 415–25 (2016).

278. Rogers, T. A., Andrews, G. E., Jaeger, L. & Grabow, W. W. Fluorescent monitoring of RNA assembly and processing using the split-spinach aptamer. *ACS Synth Biol* **4**, 162–166 (2015).

279. Ausländer, S., Fuchs, D., Hürlemann, S., Ausländer, D. & Fussenegger, M. Engineering a ribozyme cleavage-induced split fluorescent aptamer complementation assay. *Nucleic Acids Res* **44**, (2016).

280. Paige, J. S., Nguyen-Duc, T., Song, W. & Jaffrey, S. R. Fluorescence imaging of cellular metabolites with RNA. *Science* **335**, 1194 (2012).

281. Kellenberger, C. A., Wilson, S. C., Sales-Lee, J. & Hammond, M. C. RNA-based fluorescent biosensors for live cell imaging of second messengers cyclic di-GMP and cyclic AMP-GMP. *J Am Chem Soc* **135**, 4906–9 (2013).

282. Song, W., Strack, R. L. & Jaffrey, S. R. Imaging bacterial protein expression using genetically encoded RNA sensors. *Nat Methods* **10**, 873–5 (2013).

CHAPTER V

Publication

5.1 **Research article:** HIF1α-AS1 is a DNA:DNA:RNA triplex-forming lncRNA interacting with the HUSH complex.

Matthias S. Leisegang*, Jasleen Kaur Bains*, Sandra Seredinski, James A. Oo, Nina M. Krause, Chao-Chung Kuo, Stefan Günther, Nevcin Sentürk Cetin, Timothy Warwick, Can Cao, Frederike Boos, Judit Izquierdo Ponce, Shaza Haydar, Rebecca Bednarz, Chanil Valasarajan, Dominik C. Fuhrmann, Jens Preussner, Mario Looso, Soni S. Pullamsetti, Marcel H. Schulz, Hendrik R. A. Jonker, Christian Richter, Flávia Rezende, Ralf Gilsbach, Beatrice Pflüger-Müller, Ilka Wittig, Ingrid Grummt, Teodora Ribarska, Ivan G. Costa, Harald Schwalbe and Ralf P. Brandes; *Nature Communications* (**13**), 6563 (2022).

* These authors contributed equally to this work.

In this work, *HIF1α-AS1* was identified as a functionally important triplex-forming lncRNA in human endothelial cells using a combination of bioinformatics techniques, RNA/DNA pulldown experiments, and biophysical experiments such as CD and NMR spectroscopy. Through RNA:DNA:DNA triplex formation, *HIF1α-AS1* downregulates the expression the gene expression of the identified target genes, such as EPH receptor A2 (EPHA2) and adrenomedullin ADM in a *trans*-acting manner, by recruiting the repressive human silencing hub (HUSH) complex.

The groups of Leisegang and Brandes performed the bioinformatics studies (T. Warwick), triplex sequencing using RNA and DNA interacting complexes ligated and sequenced (RADICL-seq; S. Seredinski), and further cell culture experiments. The author of this thesis performed the EMSA, CD, and NMR experiments and analyzed the data to characterize and distinguish the ssRNA, ssDNA, RNA:DNA heteroduplex, and RNA:DNA:DNA triplex motifs. The 2D NMR experiments were performed and analyzed together with N.M. Krause and C. Richter. H.R.A. Jonker generated models of the RNA:DNA:DNA triplex. Moreover, the author of this thesis prepared the figures and wrote the manuscript together with M.S. Leisegang, R.P. Brandes and H. Schwalbe.

https://doi.org/10.1038/s41467-022-34252-2

nature communications

Article

https://doi.org/10.1038/s41467-022-34252-2

HIF1α-AS1 is a DNA:DNA:RNA triplex-forming lncRNA interacting with the HUSH complex

Received: 29 September 2021

Accepted: 19 October 2022

Published online: 02 November 2022

Check for updates

Matthias S. Leisegang[1,2,12], Jasleen Kaur Bains[3,12], Sandra Seredinski[1,2], James A. Oo[1,2], Nina M. Krause[3], Chao-Chung Kuo[4], Stefan Günther[5], Nevcin Sentürk Cetin[6], Timothy Warwick[1,2], Can Cao[1], Frederike Boos[1,2], Judit Izquierdo Ponce[1], Shaza Haydar[1,2], Rebecca Bednarz[1], Chanil Valasarajan[5,7], Dominik C. Fuhrmann[8], Jens Preussner[5], Mario Looso[5], Soni S. Pullamsetti[5,7], Marcel H. Schulz[2,9], Hendrik R. A. Jonker[3], Christian Richter[3], Flávia Rezende[1,2], Ralf Gilsbach[1,2], Beatrice Pflüger-Müller[1,2], Ilka Wittig[1,2,10], Ingrid Grummt[6], Teodora Ribarska[6,11], Ivan G. Costa[4], Harald Schwalbe[3,13] ✉ & Ralf P. Brandes[1,2,13] ✉

DNA:DNA:RNA triplexes that are formed through Hoogsteen base-pairing of the RNA in the major groove of the DNA duplex have been observed in vitro, but the extent to which these interactions occur in cells and how they impact cellular functions remains elusive. Using a combination of bioinformatic techniques, RNA/DNA pulldown and biophysical studies, we set out to identify functionally important DNA:DNA:RNA triplex-forming long non-coding RNAs (lncRNA) in human endothelial cells. The lncRNA *HIF1α-AS1* was retrieved as a top hit. Endogenous *HIF1α-AS1* reduces the expression of numerous genes, including EPH Receptor A2 and Adrenomedullin through DNA:DNA:RNA triplex formation by acting as an adapter for the repressive human silencing hub complex (HUSH). Moreover, the oxygen-sensitive *HIF1α-AS1* is down-regulated in pulmonary hypertension and loss-of-function approaches not only result in gene de-repression but also enhance angiogenic capacity. As exemplified here with *HIF1α-AS1*, DNA:DNA:RNA triplex formation is a functionally important mechanism of trans-acting gene expression control.

Long non-coding RNAs (lncRNAs) represent the most diverse, plastic and poorly understood class of ncRNA[1]. Their gene regulatory mechanisms involve formation of RNA-protein, RNA-RNA or RNA-DNA complexes[1]. RNA-DNA interactions occur either in heteroduplex (DNA:RNA) or triplex strands (DNA:DNA:RNA). In triplexes, double-stranded DNA (dsDNA) accommodates the single-stranded RNA in its major groove[2]. The binding occurs via Hoogsteen or reverse Hoogsteen hydrogen bonds with a purine-rich sequence of DNA to which the

[1]Institute for Cardiovascular Physiology, Goethe University, Frankfurt, Germany. [2]German Center of Cardiovascular Research (DZHK), Partner site RheinMain, Frankfurt, Germany. [3]Institute for Organic Chemistry and Chemical Biology, Center for Biomolecular Magnetic Resonance (BMRZ), Goethe University, Frankfurt, Germany. [4]Institute for Computational Genomics, Joint Research Center for Computational Biomedicine, RWTH Aachen Medical Faculty, Aachen, Germany. [5]Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany. [6]Division of Molecular Biology of the Cell II, German Cancer Research Center DKFZ-ZMBH, Heidelberg, Germany. [7]Department of Internal Medicine, Member of the DZL, Member of Cardio-Pulmonary Institute (CPI), Justus Liebig University, Gießen, Germany. [8]Faculty of Medicine, Institute of Biochemistry I, Goethe University, Frankfurt, Germany. [9]Institute of Cardiovascular Regeneration, Goethe University, Frankfurt, Germany. [10]Functional Proteomics, SFB 815 Core Unit, Faculty of Medicine, Goethe University, Frankfurt, Germany. [11]Oslo University Hospital, Oslo, Norway. [12]These authors contributed equally: Matthias S. Leisegang, Jasleen Kaur Bains. [13]These authors jointly supervised this work: Harald Schwalbe, Ralf P. Brandes. ✉e-mail: schwalbe@nmr.uni-frankfurt.de; brandes@vrc.uni-frankfurt.de

RNA strand binds in a parallel or antiparallel manner. Hoogsteen bonds are weaker than Watson-Crick bonds, resulting in Hoogsteen pairing rules being more flexible[3].

Ex vivo characterization of triplex formation relies on a variety of different biophysical methods including circular dichroism- (CD) and nuclear magnetic resonance-spectroscopy (NMR)[4–6]. Even with these techniques it can be challenging to discriminate DNA-RNA heteroduplexes from triplexes and analyses are usually restricted to oligonucleotides of a limited length. Nevertheless, a few lncRNAs have been suggested to form triplexes with dsDNA, however, triplex studies using living cells are still in early development[4,6–13]. In silico analyses of RNA-DNA triplex formation predicted several genomic loci and lncRNAs to form triplexes[14]. In line with this, a global approach in HeLa S3 and U2OS cells to isolate triplex-forming RNAs on a genome-wide scale yielded several RNA:DNA triplex-forming lncRNAs[15].

In addition to the sparse initial findings of triplex formation within cells, several other open questions remain: What is the physiological relevance of triplex-forming lncRNAs and are these cell- and tissue-type specific? What is the mechanism of action of triplex-forming lncRNAs? Do they disturb transcription in a similar way to R-loops[16] or recruit certain protein complexes to DNA in a site-specific manner? Regarding the latter aspect, Polycomb Repressive Complex 2 (PRC2) has been identified as a target of the lncRNAs HOX Transcript Antisense RNA (*HOTAIR*), FOXF1 Adjacent Non-Coding Developmental Regulatory RNA (*FENDRR*) and Maternally Expressed 3 (*MEG3*)[4,12,13], but, given the highly promiscuous nature of PRC2, this function remains controversial. Other examples of protein interactors involve e.g. E2F1 and p300, which are recruited by the triplex-forming antisense lncRNA *KHPS1* to activate gene expression of the proto-oncogene sphingosine kinase 1 (SPHK1) *in cis*[7,10].

Much of today's in vitro RNA research heavily relies on immortalized cell lines. Although such model systems are well suited for transfection or genomic manipulation, they are highly dedifferentiated and exhibit reaction patterns such as unlimited growth and immortalization - characteristics not observed in primary cells[17]. Considering that lncRNAs are expressed in a species-, tissue- and differentiation-specific manner[1], biological evidence for lncRNA functions in primary cells is limited. Among such cells, endothelial cells stand out due to their well documented importance in regeneration, angiogenesis and tissue vascularization. Indeed, endothelial cell dysfunction is one of the main drivers of systemic diseases like diabetes and inflammation[18].

Here, we combined molecular biology and biophysics, bioinformatics and physiology to systematically uncover the role of triplex-forming lncRNAs in endothelial cells. This approach identified *HIF1α-AS1* as a *trans*-acting triplex-forming lncRNA that controls vascular gene expression in endothelial cells with implications for vascular disease.

Results

HIF1α-AS1 is a triplex-associated lncRNA

To identify triplex-associated lncRNAs, we used Triplex-Seq data from U2OS and HeLa S3 cells[15]. Triplex-Seq relies on the isolation of RNase H-resistant RNA-DNA complexes from cells followed by DNA- and RNA-Seq[15]. RNase H cleaves the RNA in DNA-RNA heteroduplexes as present in R-loops[19] and has previously been used to distinguish between heteroduplexes and triplexes[20]. The Triplex-Seq data comprised all RNA entities and was filtered for the number of individual lncRNA genes, resulting in 989 (for HeLa S3, Supplementary Data 1) and 1363 (for U2OS, Supplementary Data 2) different lncRNAs associated with triplexes, with an overlap of 280 lncRNA genes between the two cell lines (Fig. 1a). To further narrow down this set of enriched triplex-associated lncRNAs, parameters for specificity (fold enrichment >10, -log10(P value peak enrichment)) were increased so that 11 lncRNA candidates with high confidence remained. Subsequently, these were correlated to ENCODE and FANTOM5 Cap Analysis of Gene Expression (CAGE)[21–23] data. Of the 11 candidates, only 5 (*RMRP*, *HIF1α-AS1*, *RP5-857K21.4*, *SCARNA2* and *SNHG8*) were expressed in endothelial cells. All 5 candidates were predicted as non-coding by the online tools Coding Potential Assessment Tool (CPAT 3.0.0) and coding potential calculator 2 (CPC2) and at least partially nuclear localized by ENCODE CAGE (Fig. 1a). To further analyze these candidates, the Triplex-Seq enriched regions were manually inspected in the IGV browser. This led to the exclusion of *SNHG8* as the triplex-associated regions within this lncRNA were exclusively within the overlapping small nucleolar RNA 24 (*SNORA24*) gene. In the case of the other candidates, triplex-association was within the individual lncRNA gene body. The cumulative fold enrichment of the remaining lncRNAs in the Triplex-Seq dataset illustrated strong triplex-association (Supplementary Fig. 1a). To verify the candidates experimentally, RNA immunoprecipitation (RIP) with antibodies against dsDNA with or without RNase H treatment in human endothelial cells was performed. Cleavage of the RNA in DNA-RNA heteroduplexes by RNase H[19] revealed that *HIF1α-AS1* was the strongest triplex-associated lncRNA (Fig. 1b).

Genomically, *HIF1α-AS1* is located on the antisense strand of the Hypoxia-inducible factor 1-alpha gene (*HIF1A*) (Fig. 1c). The lncRNA was specifically enriched in nuclear DNA, whereas *HIF1α* mRNA and *18 S* rRNA were not (Fig. 1d). Moreover, RIP with anti-histone 3 (Fig. 1e) indicated that *HIF1α-AS1* is bound to dsDNA in the chromatin environment.

HIF1α-AS1 is disease-relevant

Only a few studies have so far documented the biological relevance of *HIF1α-AS1*. Increased *HIF1α-AS1* expression has been reported in thoracoabdominal aortic aneurysms[24]. *HIF1α-AS1* was also suggested as a biomarker in colorectal carcinoma[25]. Functionally, *HIF1α-AS1* is pro-apoptotic and anti-proliferative in vascular smooth muscle, Kupffer and umbilical vein endothelial cells[26–28].

As HIF1α is a central regulator of oxygen-dependent gene expression[18], we decided to measure the expression of *HIF1α-AS1* in endothelial cells under altered oxygen and disease conditions. Hypoxia led to a decrease in *HIF1α-AS1* expression in endothelial and pulmonary artery smooth muscle cells (paSMC) (Fig. 1f, Supplementary Fig. 1b), which was restored in endothelial cells after 4 h and even surpassed basal levels after 24 h of normoxic conditions (Fig. 1g). Importantly, *HIF1α-AS1* was downregulated in endothelial cells isolated from human glioblastoma (Supplementary Fig. 1c) and in lungs from patients with end stage idiopathic pulmonary arterial hypertension (IPAH) or chronic thromboembolic pulmonary hypertension (CTEPH) (Fig. 1h). In paSMCs isolated from pulmonary arteries of patients with IPAH, *HIF1α-AS1* was strongly decreased (Supplementary Fig. 1d). Together, these data demonstrate that *HIF1α-AS1* is an oxygen-dependent and disease-relevant lncRNA.

HIF1α-AS1-triplex binding suppresses target gene expression

Triplex-Seq can provide evidence for existing triplex forming regions of the RNA (TFR) and triplex target sites (TTS) within the DNA but the details of exactly which TFR and TTS interact cannot be derived from Triplex-Seq. To identify the TFRs within *HIF1α-AS1* as well as *HIF1α-AS1*-dependent TTS, a combination of bioinformatics and wet lab approaches were used: An Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-Seq) was performed after *HIF1α-AS1* knockdown to identify DNA target sites in human endothelial cells. LNA-GapmeRs targeting *HIF1α-AS1* led to a strong knockdown of the lncRNA (Supplementary Fig. 1e). Triplex Domain Finder (TDF), a computational tool for the prediction of RNA and DNA triplex-forming potential[14], predicted the TFRs within *HIF1α-AS1* to target DNA regions around genes that displayed altered ATAC-Seq peaks after *HIF1α-AS1* silencing (Fig. 2a). The software identified three statistically significant TFRs (TFR1-3) within the pre-processed *HIF1α-AS1* RNA

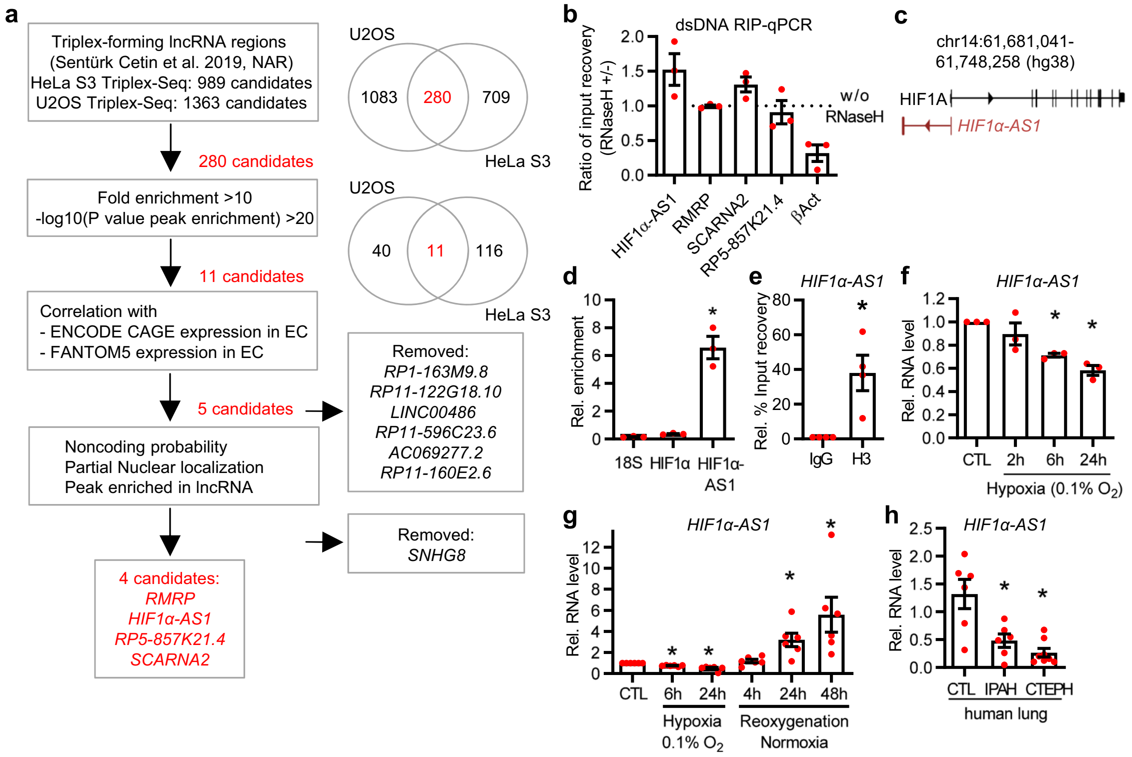


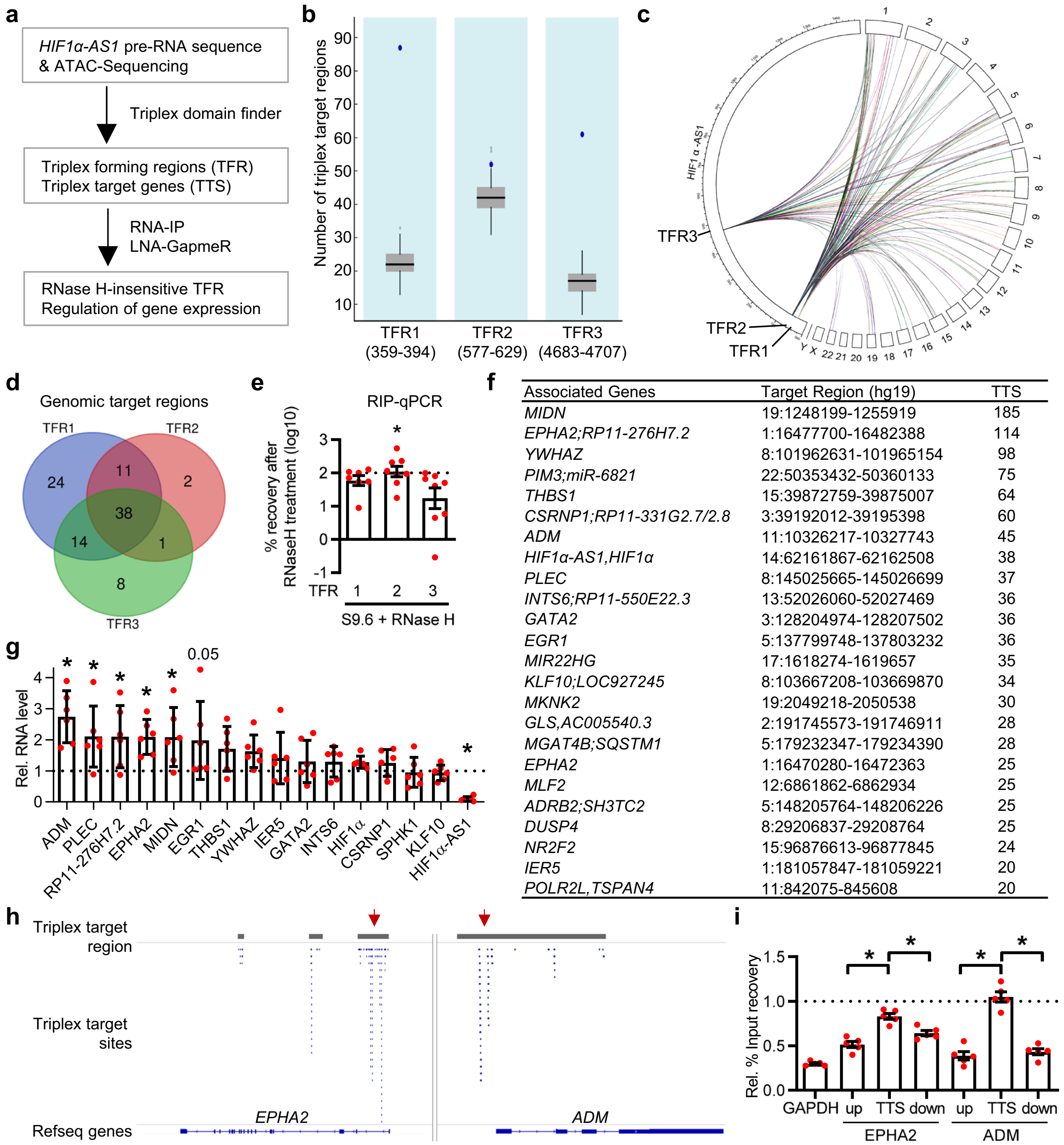
Fig. 1 | *HIF1α-AS1* is a triplex- and DNA-associated RNase H-insensitive lncRNA in endothelial cells. a Overview of the identification of endothelial-expressed triplex-forming lncRNAs. LncRNAs from a previous Triplex-Seq study in HeLa S3 and U2OS were overlapped, filtered with high stringency and analyzed for nuclear expression in endothelial cells with ENCODE and FANTOM5 CAGE data followed by analyses for noncoding probability and enriched peaks in the Triplex-Seq data. **b** RNA-immunoprecipitation with anti-dsDNA antibody followed by qPCR (RIP-qPCR) targeting the lncRNA candidates in HUVEC. Samples were treated with or without RNase H. βAct served as control for RNase H-mediated degradation. $n = 3$. **c** Scheme of the human genomic locus of *HIF1α-AS1*. **d** RT-qPCR after anti-dsDNA-RIP in HUVEC. *HIF1α* and *18 S* rRNA served as negative control. One-way ANOVA with Tukey's test, $n = 3$. *($p = 0.0002$). **e** RIP-qPCR with anti-histone3 (H3) in HUVEC. Data was normalized against GAPDH. Paired *t* test, $n = 4$. *($p = 0.0364$). **f** RT-qPCR of *HIF1α-AS1* in HUVEC treated with hypoxia (0.1% O_2) for the indicated time points. Normoxia served as negative control (CTL). $n = 3$, One-Way ANOVA with Bonferroni test. *6 h ($p = 0.0216$), * 24 h ($p = 0.0035$). **g** RT-qPCR of *HIF1α-AS1* in HUVECs treated with hypoxia (0.1% O_2) followed by reoxygenation with normoxia (after 24 h of hypoxia) for the indicated time points. $n = 6$, paired *t* test. *Hypoxia-6h ($p = 0.0001$), *Hypoxia-24h ($p = 0.002$), *Reoxygenation-24h ($p = 0.0189$), *Reoxygenation-48h ($p = 0.04$). **h** RT-qPCR of *HIF1α-AS1* in lungs from control donors (CTL, $n = 6$) or patients with idiopathic pulmonary arterial hypertension (IPAH, $n = 6$) or chronic thromboembolic pulmonary hypertension (CTEPH, $n = 8$). One-Way ANOVA with Tukey's test. *IPAH ($p = 0.0063$), *CTEPH ($p = 0.0005$). For **b**, **d**–**g**, *n* is defined as number of independent experiments. Data are presented as mean values ± SEM.

(Fig. 2b). There was also a high incidence of triplex-prone motifs predicted in regions whose chromatin state was altered in the ATAC-Seq data after *HIF1α-AS1* knockdown (Fig. 2c, Supplementary Data 3–5). Of these TTS, 38 overlapped within all three TFRs (Fig. 2d). To identify which TFR is most strongly associated with triplexes, RIP with S9.6 antibodies recognizing RNA-DNA association was performed. RNA-DNA associations remaining after RNase H treatment excluded the possibility that these were RNA-DNA heteroduplexes. Of the three *HIF1α-AS1* TFRs, TFR2 was identified as the TFR most resistant to RNase H (Fig. 2e). TFR2 is located intronically 478 nucleotides (nt) downstream of Exon1 and was detected by RT-PCR within nuclear isolated RNA with primers covering the first 714 nt (E1-I) of the pre-processed *HIF1α-AS1* (Supplementary Fig. 1f). Triplex-prone motifs in the TFR1-3-overlapping target regions yielded more than 20 different associated genes, some of which displayed a high number of DNA binding sites (Fig. 2f). If this binding of the lncRNA is relevant for the individual target gene, then a change in target gene expression would be expected. Importantly, in response to the downregulation of *HIF1α-AS1* with LNA-GapmeRs the expression of the following triplex target genes increased: *ADM*, *PLEC*, *RP11-276H7.2*, *EPHA2*, *MIDN* and *EGR1* (Fig. 2g). Interestingly, as exemplified by the target genes *HIF1A*, *EPHA2* and *ADM*, the triplex target sites are often located close to the 5′ end of the gene. In this region, histone modifications, transcription factor binding and chromatin conformation often have the greatest effect on promoter function and gene expression (Fig. 2h, Supplementary Fig. 1g). In order to prove that the triplexes also exist in vivo, Chromatin Immunoprecipitation (ChIP) was performed with S9.6 antibodies. After RNase H treatment, the TTS of *EPHA2* and *ADM* were both more resistent to RNase H treatment compared to DNA regions upstream or downstream of both TTS (Fig. 2i).

These data indicate that *HIF1α-AS1* contains triplex forming regions and target sites important for the regulation of gene expression.

HIF1α-AS1 TFR2 RNA forms triplexes with EPHA2 and ADM

Our analysis identified *HIF1α-AS1* TFR2 as the best suited candidate for verification of triplex formation of the lncRNA using biophysical and biochemical techniques. To monitor triplex formation of *HIF1α-AS1*, *EPHA2* was chosen as the target gene due to its abundance of triplex target sites (Figs. 2f, h), its regulatory potential (Fig. 2g) and its importance for vascularization[29]. Triplex domain finder predicted not the complete TFR2 to bind *EPHA2* TTS, but rather a core TFR2 sequence that binds the TTS. The formation of DNA:DNA:RNA triplexes between lncRNA *HIF1α-AS1* TFR2 and its proposed DNA target site within intron 1 of *EPHA2* was characterized by electrophoretic mobility shift assay (EMSA), CD- and solution NMR-spectroscopy. From electrophoretic mobility shift assay (EMSA) the *HIF1α-AS1* TFR2 RNA was found to form a low-mobility DNA-RNA complex with the



Associated Genes	Target Region (hg19)	TTS
MIDN	19:1248199-1255919	185
EPHA2;RP11-276H7.2	1:16477700-16482388	114
YWHAZ	8:101962631-101965154	98
PIM3;miR-6821	22:50353432-50360133	75
THBS1	15:39872759-39875007	64
CSRNP1;RP11-331G2.7/2.8	3:39192012-39195398	60
ADM	11:10326217-10327743	45
HIF1α-AS1,HIF1α	14:62161867-62162508	38
PLEC	8:145025665-145026699	37
INTS6;RP11-550E22.3	13:52026060-52027469	36
GATA2	3:128204974-128207502	36
EGR1	5:137799748-137803232	36
MIR22HG	17:1618274-1619657	35
KLF10;LOC927245	8:103667208-103669870	34
MKNK2	19:2049218-2050538	30
GLS,AC005540.3	2:191745573-191746911	28
MGAT4B;SQSTM1	5:179232347-179234390	28
EPHA2	1:16470280-16472363	25
MLF2	12:6861862-6862934	25
ADRB2;SH3TC2	5:148205764-148206226	25
DUSP4	8:29206837-29208764	25
NR2F2	15:96876613-96877845	24
IER5	1:181057847-181059221	20
POLR2L,TSPAN4	11:842075-845608	20

EPHA2 DNA target sequence (Fig. 3a). We also used CD-spectroscopy to confirm triplex formation of *HIF1α-AS1* TFR2 on *EPHA2*. The CD spectrum indicated typical features for triplex formation, such as a positive small peak at ~220 nm, two negative peaks at ~210 nm and ~240 nm and a blue-shift of the peak at ~270 nm[30,31], which was distinct from the *EPHA2* DNA duplex or the heteroduplex spectra (Fig. 3b). This confirmed the existence of *EPHA2:HIF1α-AS1* TFR2 triplexes. Additionally, we performed thermal melting assays and obtained melting temperatures T_m (RNA-DNA heteroduplex) = 53.48 ± 0.32 °C, T_m (DNA-DNA duplex) = 70.74 ± 0.22 °C and T_m (DNA-DNA-RNA triplex) = 49.52 ± 0.22 °C with a very broad second melting point around 70 °C. The biphasic melting transition is a distinct feature of triplex formation, which is characterized by a first melting temperature that corresponds to melting of Hoogsteen hydrogen bonds that stabilize the triplex and the second for the melting of the Watson-Crick base pairing at higher temperatures (Fig. 3c). ^{1}H-1D NMR spectra were recorded for *EPHA2* DNA duplex (25 nt), *HIF1α-AS1* TFR2 RNA (TFO2-23, 23 nt), *EPHA2:HIF1α-AS1*_TFR2 heteroduplex and *EPHA2:HIF1α-AS1*_TFR2 triplex at different temperatures (Fig. 3d). Using 10 eq *HIF1α-AS1* TFR2 RNA, triplex ^{1}H NMR imino signals were observed in a spectral region between 9 and 12 ppm providing further evidence that *HIF1α-AS1* was associated with *EPHA2* through Hoogsteen base pairing (Fig. 3e, Supplementary Fig. 2a). Further, we conducted NMR-spectroscopic analysis of the triplex: we first measured a ^{1}H, ^{1}H-NOESY spectrum for *EPHA2* DNA duplex and assigned cross peaks in this spectrum of the DNA duplex. We identified 11 G and 12 T imino proton signals (Fig. 3f). Then, we added the *HIF1α-AS1*_TFR2 RNA triplex-forming strand in 10-fold excess [DNA-DNA]:[RNA] = 1:10 and

Fig. 2 | *HIF1α-AS1* potentially forms DNA:DNA:RNA triplexes. a Overview of the identification of *HIF1α-AS1* triplex forming regions (TFR) and their DNA triplex target sites (TTS) with triplex domain finder (TDF). *HIF1α-AS1* pre-RNA and ATAC-Seq of HUVECs treated with or without LNA GapmeRs against *HIF1α-AS1* were used as input for TDF. RIP and LNA GapmeRs were used to validate the findings of TDF. **b** All three TFRs of *HIF1α-AS1* have a significantly higher number of DNA triplex target regions (blue dots) than the random background (boxplots in gray). Boxplot visualizes the median, first and third quartiles. The whiskers present the 1.5 inter-quartile range. External gray dots represent outliers. Numbers in brackets are positions of the individual TFR within *HIF1α-AS1* pre-RNA. Analyzed with TDF, n=200 randomizations. **c** Circos plot showing the localization of the individual TFR within *HIF1α-AS1* pre-RNA and its interaction with the chromosomal TTS. **d** Overlap of TTS of the individual TFRs of *HIF1α-AS1*. **e** Identification of RNase H-resistant TFRs. RIP with anti-S9.6 with/without RNase H treatment in HUVEC followed by qPCR for the TFRs. Ratio of %-input recovery with/without RNase H treatment is shown. *n* = 8 independent experiments, paired *t* test. Dotted line represents normalized values without RNase H treatment. The asterisk indicates that the %-recovery after RNase H is significantly different for TFR2 compared to TFR1 (*p* = 0.0452) and TFR3 (*p* = 0.0244). **f** *HIF1α-AS1* TFR1-3 overlapping top target genes, their genomic location and number of TTS identified by TDF. **g** RT-qPCR of triplex target genes of TFR2 after knockdown of *HIF1α-AS1* in HUVEC. *n* = 6 independent experiments. One-Way ANOVA with Holm's Sidak test. *ADM (*p* = <0.0001), *HIF1α-AS1 (*p* = <0.0001), *PLEC (*p* = 0.0238), *RP11-276H7.2 (*p* = 0.0238), *EPHA2 (*p* = 0.0238), *MIDN (*p* = 0.023). **h** Different triplex target regions of *HIF1α-AS1* are shown. Triplex target regions are highlighted in gray, triplex target sites are shown in blue. Arrows indicate TTS at the 5′end. **i** ChIP-qPCR with S9.6 antibody with/without RNase H treatment. Dotted line represents normalized values without RNase H. QPCR was performed against *EPHA2* or *ADM* TTS or regions up- and downstream of the individual TTS. One-Way ANOVA with Bonferonni test. *n* = 5 independent experiments. *EPHA2:up/TTS (*p* = <0.0001), *ADM:up/TTS (*p* = <0.0001), *ADM:TTS/down (*p* = <0.0001), *EPHA2:TTS/down (*p* = 0.0321). Data are presented as mean values ± SEM.

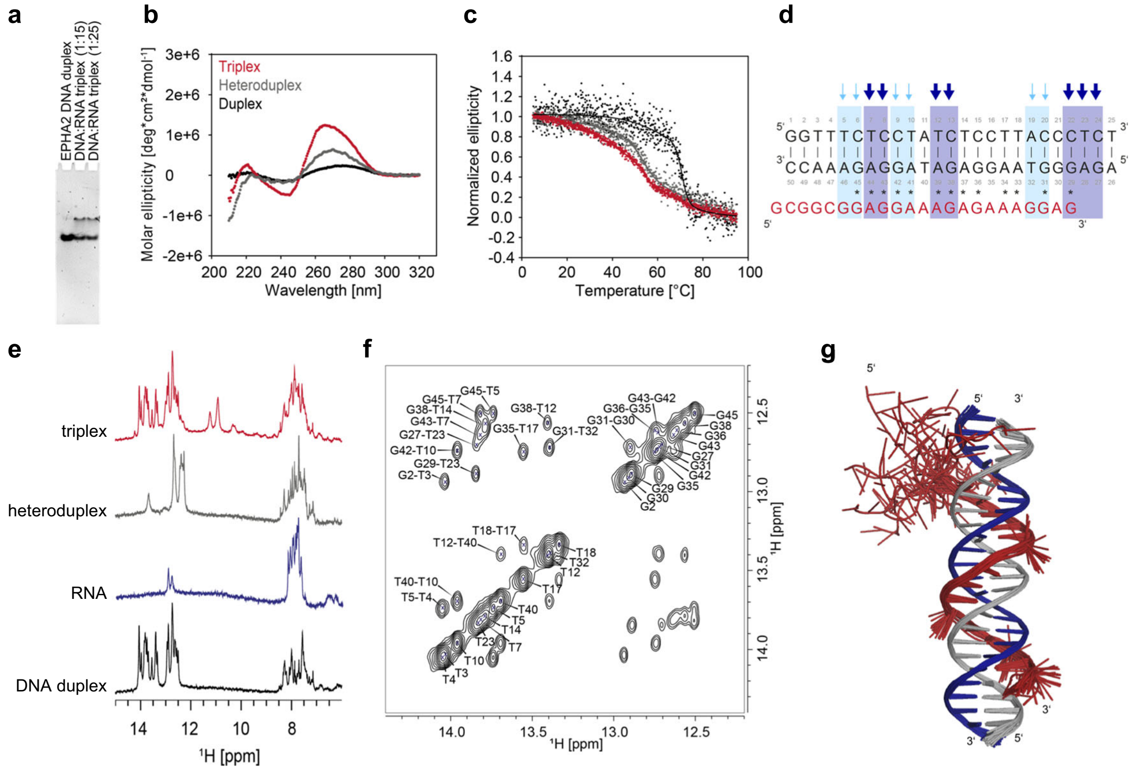


Fig. 3 | *HIF1α-AS1* TFR2 RNA forms in vitro DNA:DNA:RNA triplexes with the predicted DNA target region in *EPHA2*. a Electromobility shift assay of *EPHA2* DNA duplex and *EPHA2:HIF1α-AS1*_TFR2 triplex with 15 and 25 equivalents of RNA (TFO2-23). **b** Circular dichroism spectra of the *EPHA2* DNA duplex (black), the heteroduplex (dark gray) and *EPHA2:HIF1α-AS1*-TFR2 (red) measured at 298 K. **c** Thermal melting assay of the *EPHA2* DNA duplex (black), the heteroduplex (dark gray) and *EPHA2:HIF1α-AS1*-TFR2 (red). **d** Sequence of *EPHA2* DNA (black) and *HIF1α-AS1*-TFR2 RNA (red). Watson-Crick base pairing is indicated with | and the Hoogsteen base pairing is indicated with *. Changes in the DNA duplex were quantitatively analyzed using NOESY spectra of duplex and triplex. Imino protons with strong attenuation (dark blue arrows) or medium attenuation (light blue arrows) of cross peak intensities in the imino-imino region were observed. **e** ^{1}H-1D NMR spectra of the *EPHA2* DNA duplex (black), *HIF1α-AS1* TFR2 RNA (blue), heteroduplex (dark gray) and *EPHA2:HIF1α-AS1*-TFR2 triplex (red) at 288 K. **f** Assignment of the imino region of the ^{1}H,^{1}H-NOESY spectrum of *EPHA2* DNA duplex measured at 800 MHz and 288 K in NMR buffer with 5% D_2O. **g** Cartoon representation of DNA:DNA:RNA triplex docking studies with the following color code: DNA strand (blue and gray) and RNA strand (red). This shows an ensemble of the 20 top-ranked modeled structures for a DNA:DNA:RNA triplex. The figure was generated by using PyMol 2.5 (Schrödinger, LLC). Source data (for a) are provided as a Source Data file.

semi-quantitatively analyzed the change in the DNA duplex spectrum. For 7 G- and 6 T-imino protons either a strong or medium attenuation of cross peak intensities in the imino-imino region of the NOESY spectrum was observed (Supplementary Fig. 2b). We rationalize this attenuation as to arise from weakening of the Watson-Crick base pairing induced by the Hoogsteen interaction with the RNA strand. From this analysis, we compared predicted Hoogsteen interactions in the triplex with the detected changes in the NOESY spectrum for different positions i of RNA relative to DNA duplex strand. Interestingly, the previously predicted position (i = 0) is supported by the observed attenuations in the NOESY, where 3 G- and 2 T-imino sites disappear completely and 3 G- and 1 T-imino sites are significantly attenuated. In total, 12 sites remain unaffected in the DNA duplex (Supplementary Fig. 2c). Further, based on our interpretation that Hoogsteen interactions can be mapped from the analysis of cross peak attenuation in the NOESY, we generated structural models for the

*EPHA2:HIF1α-AS1*_TFR2 triplex. The ensemble of the 20 top-ranked structures for the triplex are displayed as cartoon (Fig. 3g).

To confirm the formation of triplexes with lower equivalents, stabilized triplex formation was investigated: the intermolecular dsDNA formed by two complementary antiparallel DNA strands was changed into a hairpin construct, where both DNA strands were linked with a 5 nt thymidine-linker and duplex formation thus became intramolecular. With this approach, triplex formation was obtained with 3 eq RNA, indicating that triplex formation is favored under those conditions as expected. ^{1}H-1D NMR spectra of hairpin *EPHA2*_CTGA and ^{15}N-labeled *HIF1α-AS1* TFR2:*EPHA2*_CTGA triplex indicated changes in the Hoogsteen region (9-12 ppm) and the spectral region of imino (12-14 ppm) and amino signals (7–8.5 ppm) (Supplementary Fig. 3a, b). In addition to *EPHA2*, we also tested *ADM*, a preprohormone involved in endothelial cell function[32]. For *ADM*_CTGA:*HIF1α-AS1* TFR2 triplex, the new imino protons in the Hoogsteen region arose at lower temperatures (Supplementary Fig. 3c and d). For both *ADM*_CTGA and *EPHA2*_CTGA triplex constructs the CD spectra showed an increased negative ellipticity at ~240 nm and positive ellipticity at ~270 nm (Supplementary Fig. 3e, f). Further, the thermal melting data verified the triplex stabilization with higher melting temperatures and defined melting transitions upon DNA hairpin formation. For the *EPHA2*_CTGA:*HIF1α-AS1* TFR2 (TFO2-23) triplex we obtained a first melting point at T_m (1st triplex) = 50.08 ± 0.51 °C, a second melting point T_m (2nd triplex) = 79.90 ± 0.10 °C and T_m (DNA hairpin) = 80.41 ± 0.10 °C (Supplementary Fig. 3g). The melting temperature of *ADM* DNA duplex T_m (DNA-DNA duplex) = 63.80 ± 0.20 °C increased for the *ADM*_CTGA hairpin T_m (DNA hairpin) = 95.76 ± 16.69 °C. For the *ADM*_CTGA:*HIF1α-AS1* TFR2 (TFO2-23), we obtained a first melting point T_m (1st triplex) = 40.24 ± 2.62 °C and a second T_m (2nd triplex) = 81.78 ± 0.59 °C (Supplementary Fig. 3h). The data demonstrate that *HIF1α-AS1* TFR2 forms triplexes with *EPHA2* and *ADM* dsDNA under regular and triplex-stabilized conditions upon DNA hairpin formation.

TFR2 represses EPHA2 and ADM gene expression

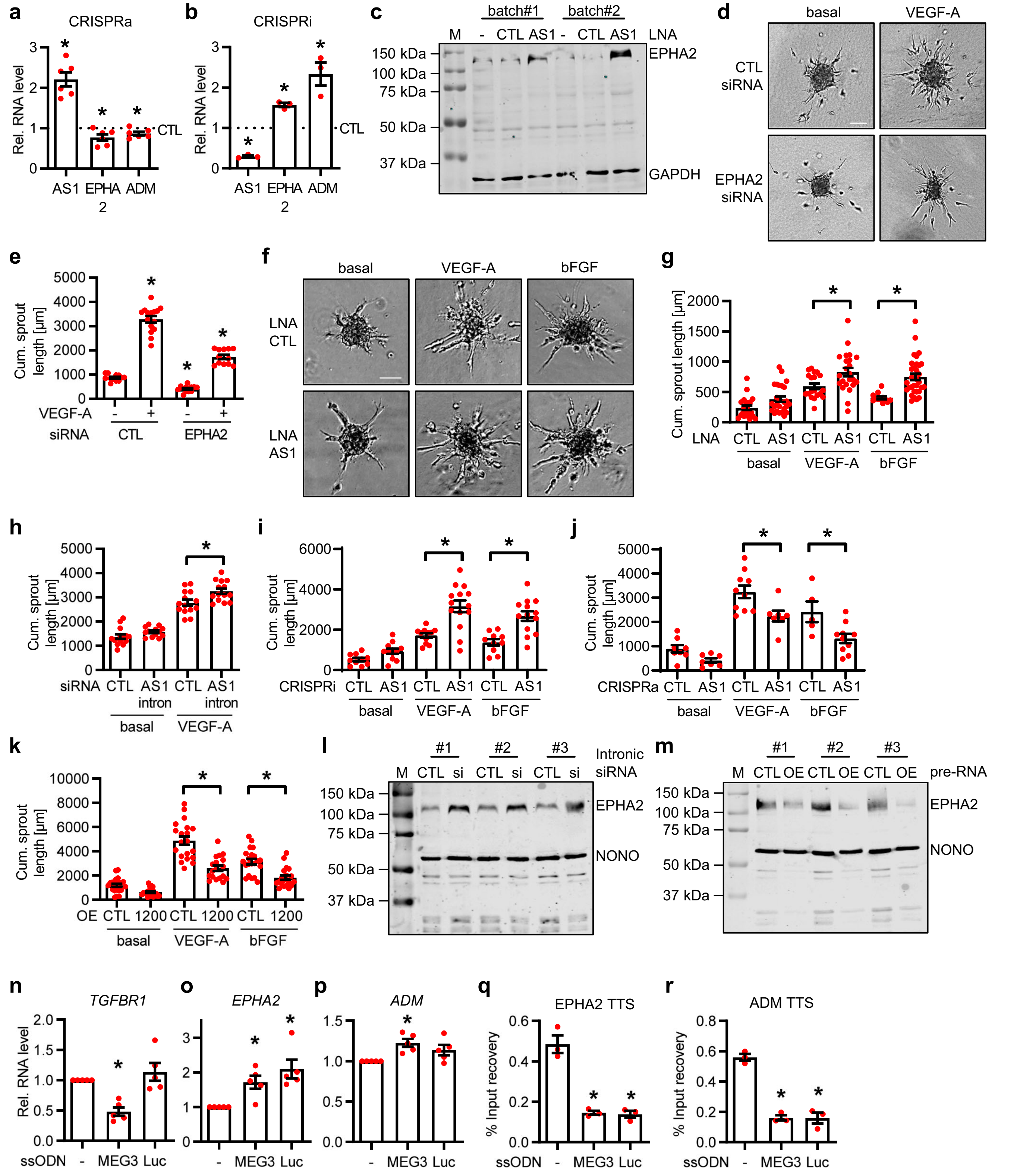
The current data indicate that *HIF1α-AS1* forms triplexes with *EPHA2* and *ADM*, however, the mechanistic and functional consequences of this phenomenon are unclear. To investigate these aspects, gain and loss of function approaches were performed. Increasing the expression of *HIF1α-AS1* using a dCas9-VP64 CRISPR activation system (CRISPRa) reduced the expression of *EPHA2* and *ADM* (Fig. 4a). Conversely, downregulation of *HIF1α-AS1* with a dCas9-KRAB repression system (CRISPRi) increased the expression of *EPHA2* and *ADM* (Fig. 4b). Consistent with *HIF1α-AS1* repressing *EPHA2* and *ADM* gene expression, EPHA2 levels increased after knockdown of *HIF1α-AS1* (Figs. 2g, 4c). EPHA2 has a multi-faceted role in angiogenesis[29,33,34]. In HUVEC, knockdown of *EPHA2* with siRNAs strongly reduced its RNA and protein expression and inhibited angiogenic sprouting (Fig. 4d and e, Supplementary Fig. 4a–c). Conversely, a knockdown of *HIF1α-AS1* with LNA-GapmeRs increased VEGF-A- and bFGF-mediated angiogenic sprouting (Fig. 4f, g, Supplementary Fig. 4d), confirming the repressive effect of *HIF1α-AS1* on *EPHA2*. Additionally, CRISPRi targeting *HIF1α-AS1* or an siRNA-mediated knockdown of the *HIF1α-AS1* pre-RNA targeting the intron region next to the TFR2 were performed. Targeting the intron of *HIF1α-AS1* not only decreased the expression of TFR2, but also increased *EPHA2* and *ADM*, whereas *HIF1α* was not significantly altered (Supplementary Fig. 4e). As expected, both CRISPRi and siRNA against *HIF1α-AS1* intron induced VEGF-A-mediated sprouting (Fig. 4h and i, Supplementary Fig. 4f, g), whereas CRISPRa and an overexpression of the first 1200 nt of the *HIF1α-AS1* gene (containing Exon1, the beginning of the intron including TFR2) had the opposite effect (Fig. 4j and k, Supplementary Fig. 4h, i). The repressive effect of *HIF1α-AS1* on *EPHA2* was further confirmed by Western analysis, where siRNA-mediated knockdown of the *HIF1α-AS1* pre-RNA increased EPHA2 and overexpression of the first 1200 nt of the *HIF1α-AS1* gene decreased EPHA2 protein levels (Fig. 4l and m). The beneficial effect on sprouting is at least partially based on an anti-apoptotic effect as knockdown of *HIF1α-AS1* increased caspase 3&7 activity as measured by a cell-permeant fluorescent probe (SR-DEVD-FMK) that bound to active caspase 3 & 7 (Supplementary Fig. 4j). To demonstrate directly that TFR2 is responsible for the regulation of *EPHA2*, we replaced TFR2 by genome editing using a recombinant Cas9-eGFP, a gRNA targeting TFR2 and different single-stranded oligodeoxynucleotides (ssODN) harboring either the published *MEG3* TFR[4] or a luciferase control sequence (Supplementary Fig. 4k). Replacement of the TFR2 with the *MEG3* TFR, which served as a positive control for a functional TFR repressing *TGFBR1* expression[4], yielded a reduction in *TGFBR1* levels compared to the luciferase control (Fig. 4n). More importantly, the loss of TFR2 consequently led to a loss of *HIF1α-AS1* TFR2, an upregulation of *EPHA2* and partially of *ADM* (Fig. 4o, p, Supplementary Fig. 4l), and also ChIP with anti-S9.6 led to a reduced detection of the TTS of *EPHA2* and *ADM* (Fig. 4q, r). These data demonstrate that TFR2 is functional as a TFR and represses *EPHA2* and *ADM* gene expression.

HIF1α-AS1 binds to and recruits HUSH to triplex targets

To elucidate the mechanism by which *HIF1α-AS1* represses gene expression, *HIF1α-AS1*-associated proteins were studied using RNA pulldown experiments. 3′-biotinylated spliced *HIF1α-AS1* lncRNA or 3′-biotinylated pcDNA3.1+ negative control were incubated in nuclear extracts from HUVECs and RNA-associated proteins were identified by electrospray ionization mass spectrometry, which retrieved M-phase phosphoprotein 8 (MPP8)-a component of the human silencing hub (HUSH) complex[35]- as top hit (Fig. 5a–b, Supplementary Data 6). The HUSH-complex is a nuclear machinery consisting of the chromodomain-containing protein MPP8, TASOR (FAM208A) and PPHLN1 (Periphilin), and was originally thought to mediate gene silencing during viral infection by recruiting the SET Domain Bifurcated Histone Lysine Methyltransferase 1 (SETDB1) which methylates H3K9[35]. The HUSH complex has not yet been studied in vascular cells and an interaction of its core protein MPP8 with lncRNAs has not been reported. To support our finding, RIP revealed that *HIF1α-AS1* and its TFR2, but not *HIF1A* mRNA, interact with MPP8 (Fig. 5c, Supplementary Fig. 5a–b). Furthermore, *HIF1α-AS1* was highly enriched with H3K9me3 (Fig. 5d).

To map the RNA binding region of MPP8 on *HIF1α-AS1*, we used *cat*RAPID fragments[36], an algorithm involving division of polypeptide and nucleotide sequences into fragments to estimate the interaction propensity of protein-RNA pairs. This highlighted potential binding regions within Exon1 (Supplementary Fig. 5c). To substantiate these data experimentally, ex vivo bindings assays were performed between fragments of *HIF1α-AS1* and recombinant MPP8 (Fig. 5e) as well as with *HIF1α-AS1* and in vitro translated MPP8 mutants, among them the mutation in the chromodomain W80A, a chromodomain deletion, an N- or C-terminal half deletion and a deletion of the Ankyrin repeats (ANK) (Fig. 5f). MPP8 interacted directly with *HIF1α-AS1* full length and a *HIF1α-AS1* mutant lacking Exon2 (Fig. 5g). In contrast and in accordance with the *cat*RAPID prediction, deletion of Exon1 (nucleotides 26-78 in particular) prevented the interaction (Fig. 5g), indicating that this region of *HIF1α-AS1* is critical for the interaction of *HIF1α-AS1* with MPP8. On the protein side, RIP of the MPP8 mutants with anti-His antibodies followed by RT-qPCR for *HIF1α-AS1* revealed that the interaction of *HIF1α-AS1* with MPP8 was strongly reduced by a deletion of the C-terminal half of MPP8, but not by deletion or mutation of the chromodomain (Fig. 5h). Further, the Ankyrin repeats in the C-terminus seem to effect the interaction only to a minor extent (Fig. 5h). Uniprot[37] listed three disordered regions, two of them in the N-terminal half and one in the C-terminal half (Fig. 5f), which could potentially be involved in the interaction.

To demonstrate that *HIF1α-AS1* acts through HUSH complex recruitment, we first tested whether parts of this complex exist in



endothelial cells. Proximity ligation assays with antibodies against MPP8, dsDNA, H3K9me3 and SETDB1 confirmed the association of MPP8 with dsDNA (Supplementary Fig. 5d), H3K9me3 (Fig. 5i) and SETDB1 (Fig. 5j) in the nuclei of endothelial cells, indicating that parts of the complex are present at endothelial chromatin.

ChIP with and without RNase A revealed that targeting of MPP8 and SETDB1, but not NP220, which is another protein associated with the HUSH complex[38] and interacting with *HIF1α-AS1* (Fig. 5b), to the *HIF1α-AS1* TTS of *EPHA2* and *ADM* were attenuated after RNA depletion (Fig. 6a, Supplementary Fig. 6a, b). A region 5.7 kb downstream of *EPHA2*, which harbors different triplex target sites to the one studied here (Fig. 2h), also appeared to be reduced after RNase treatment the binding of MPP8 and SETDB1, but not NP220, indicating that MPP8 and SETDB1 might also act there (Supplementary Fig. 6c–e). To demonstrate the dependence of the interactions with the TTS on *HIF1α-AS1*, ChIP experiments with antibodies targeting SETDB1, MPP8 and NP220

Fig. 4 | *HIF1α-AS1* limits EPHA2 and ADM expression through TFR2. a, **b** CRISPRa (**a**, $n = 6$ independent experiments) or CRISPRi (**b**, $n = 3$ independent experiments) targeting *HIF1α-AS1* in HUVECs followed by RT-qPCR for *HIF1α-AS1*, *EPHA2* and *ADM*. Paired *t* test. A non-targeting gRNA served as negative control (CTL). a: *AS1 ($p = 0.0009$), *EPHA2 ($p = 0.0335$), *ADM ($p = 0.0359$); b: *AS1 ($p = 0.0012$), *EPHA2 ($p = 0.008$), *ADM ($p = 0.0428$). **c** Western blot with (AS1)/without (-, CTL) LNA knockdown of *HIF1α-AS1* in two independent experiments using two different batches of HUVEC. GAPDH served as control. **d**–**g** Representative images (**d**, **f**) of a spheroid assay and quantification (**e**, **g**) of the cumulative sprout length of HUVECs treated with/without siRNAs against *EPHA2* (**d** and **e**) or LNA GapmeRs targeting *HIF1α-AS1* (**f** and **g**). Scale bar, 200 µm. One-Way ANOVA with Bonferroni test. e: CTL-VEGF-A ($n = 12$), CTL+VEGF-A ($n = 15$), EPHA2-VEGF-A ($n = 13$), EPHA2+VEGF-A ($n = 12$); *CTL-/+VEGF-A ($p = <0.0001$), *EPHA2-/+VEGF-A ($p = <0.0001$), *CTL/EPHA2+VEGF-A ($p = <0.0001$), *CTL/EPHA2-VEGF-A ($p = 0.0079$); g: CTL-basal ($n = 21$), AS1-basal ($n = 26$), CTL+VEGF-A ($n = 19$), AS1+VEGF-A ($n = 23$), CTL+bFGF ($n = 12$), AS1+bFGF ($n = 32$); *VEGF-A ($p = 0.0495$), *bFGF ($p = 0.0012$). **h**–**k** Quantification of the cumulative sprout length from the spheroid assays with siRNA targeting the *HIF1α-AS1* intron (**h**, $n = 14$ replicates), with CRISPRi (**i**, CTL-basal($n = 10$), AS1-basal ($n = 11$), CTL+VEGF-A ($n = 11$), AS1+VEGF-A ($n = 14$), CTL+bFGF ($n = 10$), AS1+bFGF ($n = 13$)), with CRISPRa (**j**, CTL-basal ($n = 8$), AS1-basal ($n = 7$), CTL+VEGF-A ($n = 10$), AS1+VEGF-A ($n = 7$), CTL+bFGF ($n = 5$), AS1+bFGF ($n = 10$)) or after overexpression (**k**, CTL-basal ($n = 21$), 1200-basal ($n = 22$), CTL+VEGF-A ($n = 20$), 1200+VEGF-A ($n = 18$), CTL+bFGF ($n = 21$), 1200+bFGF ($n = 19$)) of the first 1200 nt of the *HIF1α-AS1* gene (included TFR2, named as 1200). One-Way ANOVA with Bonferroni test. h: *VEGF-A ($p = 0.0109$); i: *VEGF-A ($p = <0.0001$), *bFGF ($p = 0.0006$); j: *VEGF-A($p = 0.0429$), *bFGF($p = 0.0489$); k: *VEGF-A($p = <0.0001$), *bFGF($p = 0.0004$). **l**, **m** Western blot with (si) or without (CTL) siRNA-mediated knockdown of *HIF1α-AS1* targeting the intron (**l**) or with (OE) or without (CTL) overexpression of the first 1200 nt of *HIF1α-AS1* (**m**) in three independent experiments using three different batches of HUVEC. NONO served as control. **n**, **o**, **p** RT-qPCR of *TGFBR1* (**n**), *EPHA2* (**o**) or *ADM* (**p**) after replacement of *HIF1α-AS1*-TFR2 with single-stranded oligodeoxynucleotides (ssODN) containing *MEG3*-TFR or a DNA fragment of a luciferase negative control (Luc). -, no ssODN. $n = 5$ independent experiments, Paired *t* test. n: *($p = 0.0018$); o: *MEG3($p = 0.0187$), *Luc ($p = 0.015$); p: *($p = 0.0106$). **q**, **r** ChIP with anti-S9.6 after replacement of *HIF1α-AS1*-TFR2 and qPCR for *EPHA2* (**q**) or *ADM* (**r**) TTS. -, no ssODN. One-Way ANOVA with Bonferroni test. $n = 3$ independent experiments. q: *MEG3 ($p = 0.0025$), *Luc ($p = 0 .0024$); r: *Meg3 ($p = 0.0006$), *Luc ($p = 0.0006$). Data are presented as mean values ± SEM. AS1, HIF1α-AS1. M, marker. Source data (for c, l, m) are provided as a Source Data file.

after LNA-GapmeR-mediated knockdown, CRISPRi and CRISPRa of *HIF1α-AS1* were performed. The binding of SETDB1 and MPP8, but not of NP220, to the triplex target sites of *HIF1α-AS1* -and not to the regions up- or downstream of the TTS- required the presence of the lncRNA (Fig. 6b-e, Supplementary Figs. 6f–i and 7a–j) suggesting that these interactions facilitate epigenetic processes and ultimately regulate gene expression. ATAC-Seq confirmed that these factors act in the region of the TTS: After knockdown of *HIF1α-AS1*, SETDB1 or MPP8, the chromatin accessibility of both the *EPHA2* and *ADM* transcriptional start sites were reduced, which was also seen by CRISPRi of *HIF1α-AS1* and by LentiCRISPR-mediated deletion of *HIF1α-AS1* TFR2 (Fig. 6f, Supplementary Figs. 7k and 8). LentiCRISPR-mediated deletion of *EPHA2* TTS or *ADM* TTS showed similar effects for the TSS of their gene locus and CRISPRa of *HIF1α-AS1* was confirmative by showing the opposite effect with increased open chromatin at the TSS of *EPHA2* and *ADM* (Fig. 6f, Supplementary Fig. 8). An increase in accessibility to the region downstream of the EPHA2 TTS was detected after knockdown of *HIF1α-AS1*, SETDB1 and MPP8; however, this could not be validated with CRISPRi/a for *HIF1α-AS1* or LentiCRISPR-dependent *HIF1α-AS1* TFR2, EPHA2 TTS or ADM TTS experiments, suggesting that the regions within the TSS, but not downstream of the EPHA2 TTS, may contain the most essential repressor regions. These data indicate that triplex formation by *HIF1α-AS1* is important for fine-tuning chromatin accessibility locally and thereby gene expression of *EPHA2* and *ADM* through SETDB1 and MPP8.

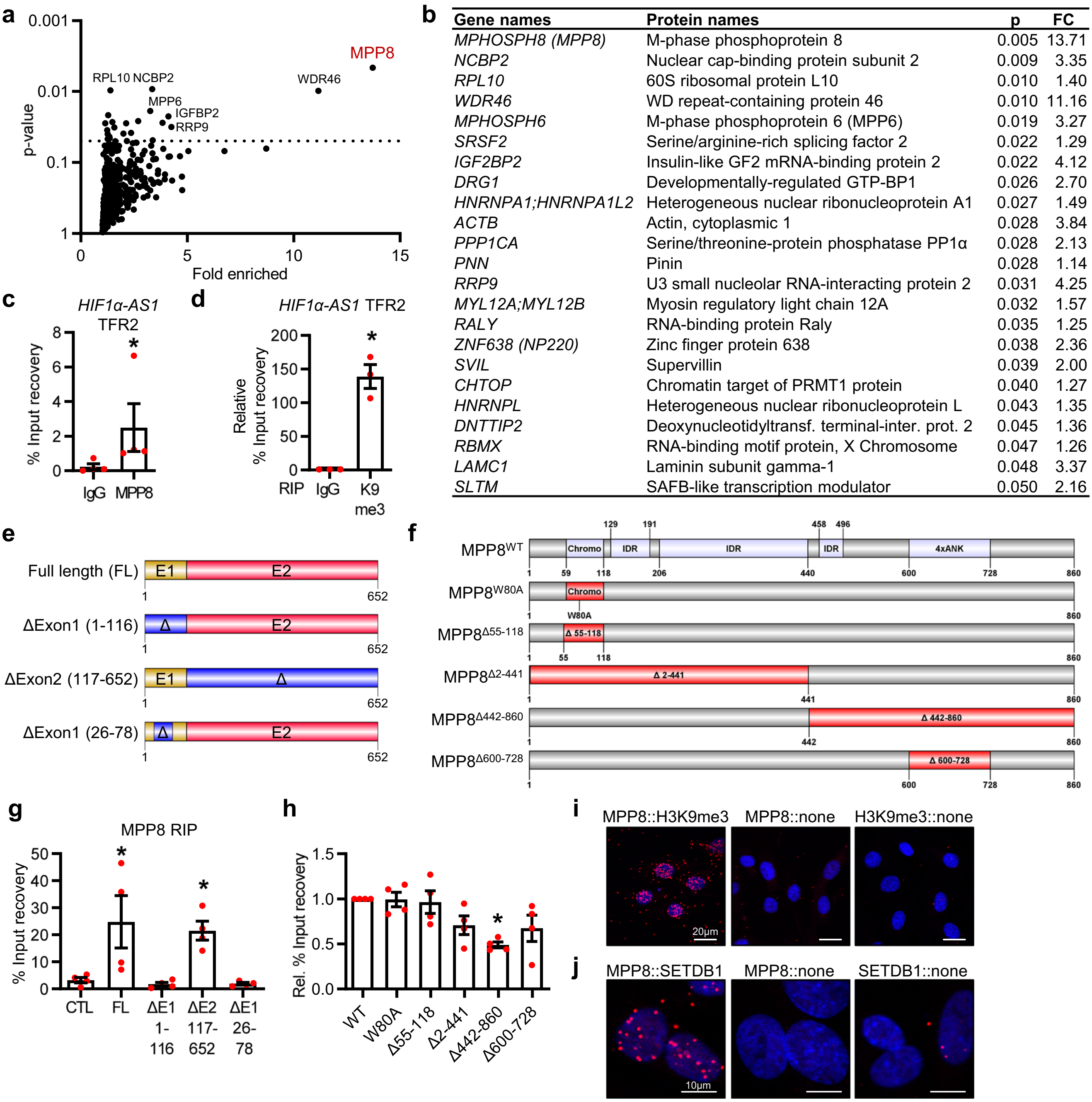
Discussion

The present study combined molecular biology, bioinformatics, physiology and structural analysis to identify and establish the lncRNA *HIF1α-AS1* as a triplex-forming lncRNA in human endothelial cells. Through *trans*-acting triplex formation by a specific region within *HIF1α-AS1*, *EPHA2* and *ADM* DNA target sites are primed for their interaction with the HUSH complex members MPP8 and SETDB1 to mediate gene repression through control of chromatin accessibility. Physiologically, the anti-angiogenic lncRNA *HIF1α-AS1* is dysregulated in hypoxia and severe angiogenic and pulmonary diseases like CTEPH, IPAH and GBM. Thus, the present work establishes a putative link of a disease-relevant lncRNA and the HUSH complex by triplex formation resulting in the inhibition of endothelial gene expression.

The interaction of chromatin modifying complexes with lncRNAs suggests that lncRNAs have targeting or scaffolding functions within these complexes to modulate chromatin structure and thereby gene expression. Most studied lncRNAs have been identified to interact with complexes such as PRC2, SWI/SNF, E2F1 and p300, e.g. *MEG3*[4], *FENDRR*[12], *MANTIS*[39], and *KHPS1*[7,10]. In the present work, we identified another silencing complex that can be targeted by lncRNAs: We demonstrated that *HIF1α-AS1* interacts with proteins of the HUSH complex, which mediates gene silencing. HUSH is also involved in silencing extrachromosomal retroviral DNA[38]. Recently it has been shown that the HUSH complex, particularly MPP8, which is downregulated in many cancer types and whose depletion caused overexpression of long interspersed element-1 (LINE-1s) and Long Terminal Repeats, controls type I Interferon signaling involving a mechanism with dsRNA sensing by MDA5 and RIG-I[40]. Here we report a direct interaction of the HUSH complex member MPP8 with *HIF1α-AS1*. Moreover, we identified Exon1 of *HIF1α-AS1* as being critical for this function. The HUSH complex has not yet been studied in vascular cells; it is not known whether its published composition with MPP8, TASOR and PPHLN1[35] is valid for endothelial cells. Our data propose that, in endothelial cells, the HUSH complex member MPP8 interacts with H3K9me3 and DNA and that SETDB1 and MPP8, but not NP220, repress gene expression of *HIF1α-AS1*-specific target genes.

The finding that *HIF1α-AS1* interacts with the C-terminal domain of MPP8, and not with the chromodomain, was unexpected. The N-terminus of MPP8 was reported to interact with H3K9me3[41]. Substitution of Trp80 to alanine (W80A) within the chromodomain showed that it is important for H3K9me3 binding[41], but also for other interactions, such as with DNMT3A[42]. Douse et al. removed the first 499 aa of MPP8 without impairing HUSH function; the function of HUSH was, however, only affected by the deletion of the amino acids 500-860, which also contain the predicted ankyrin repeats[43]. A similar finding was made for the maintenance of the self-renewal of ground-state murine embryonic stem cells, where not the chromodomain, but a C-terminal region of MPP8 was required for function[44]. The function of the IDRs within MPP8 were not investigated so far, but disordered regions are discussed to be potential linkers or binding partners of RNA[45].

We propose that *HIF1α-AS1* mediates the anti-angiogenic effects through triplex-formation with the receptor tyrosine kinase *EPHA2* and the preprohormone *ADM* genes. EPHA2 is a major regulator of angiogenic processes since EphA2-deficient mice displayed impaired angiogenesis in response to ephrin-A1 stimulation in vivo[46]. EphA2-deficient endothelial cells failed to undergo cell migration and vascular assembly in response to ephrin-A1 and only adenovirus-mediated transduction of *Epha2* restored the defect[46]. Additionally, ADM promotes arterio- and angiogenesis[32]. Both genes were upregulated after *HIF1α-AS1* knockdown, explaining why *HIF1α-AS1* knockdown increased sprouting. However, other *HIF1α-AS1* targets are likely to contribute to the phenotype, such as the proangiogenic genes *HIF1A*[47], *THBS1*[48], *EGR1*[49] or *NR2F2*[50].



Gene names	Protein names	p	FC
MPHOSPH8 (MPP8)	M-phase phosphoprotein 8	0.005	13.71
NCBP2	Nuclear cap-binding protein subunit 2	0.009	3.35
RPL10	60S ribosomal protein L10	0.010	1.40
WDR46	WD repeat-containing protein 46	0.010	11.16
MPHOSPH6	M-phase phosphoprotein 6 (MPP6)	0.019	3.27
SRSF2	Serine/arginine-rich splicing factor 2	0.022	1.29
IGF2BP2	Insulin-like GF2 mRNA-binding protein 2	0.022	4.12
DRG1	Developmentally-regulated GTP-BP1	0.026	2.70
HNRNPA1;HNRNPA1L2	Heterogeneous nuclear ribonucleoprotein A1	0.027	1.49
ACTB	Actin, cytoplasmic 1	0.028	3.84
PPP1CA	Serine/threonine-protein phosphatase PP1α	0.028	2.13
PNN	Pinin	0.028	1.14
RRP9	U3 small nucleolar RNA-interacting protein 2	0.031	4.25
MYL12A;MYL12B	Myosin regulatory light chain 12A	0.032	1.57
RALY	RNA-binding protein Raly	0.035	1.25
ZNF638 (NP220)	Zinc finger protein 638	0.038	2.36
SVIL	Supervillin	0.039	2.00
CHTOP	Chromatin target of PRMT1 protein	0.040	1.27
HNRNPL	Heterogeneous nuclear ribonucleoprotein L	0.043	1.35
DNTTIP2	Deoxynucleotidyltransf. terminal-inter. prot. 2	0.045	1.36
RBMX	RNA-binding motif protein, X Chromosome	0.047	1.26
LAMC1	Laminin subunit gamma-1	0.048	3.37
SLTM	SAFB-like transcription modulator	0.050	2.16

Fig. 5 | *HIF1α-AS1* interacts directly with the HUSH complex member MPP8. **a** Volcano plot of *HIF1α-AS1* protein interaction partners after RNA pulldown assay and ESI-MS/MS measurements with fold enrichment and *p* value. *n* = 5. Significant proteins are shown above the line ($p < 0.05$). **b** List of proteins enriched after RNA pulldown assay, their *p* value (p, unpaired *t* test, two-tailed) and absolute fold change (FC). **c** RIP with MPP8 antibodies and qPCR for *HIF1α-AS1* TFR2. IgG served as negative control. *n* = 4 independent experiments, Mann−Whitney test. *($p = 0.0286$). **d** RIP with histone3-lysine9-trimethylation antibodies and qPCR for *HIF1α-AS1* TFR2. IgG served as negative control. *n* = 3 independent experiments, Paired *t* test. *($p = 0.0162$). **e** Scheme of the different *HIF1α-AS1* RNAs used for in vitro RNA immunoprecipitation. E1, Exon1; E2, Exon2. **f** Scheme of the different MPP8 mutants used for in vitro RNA-immunoprecipitation. IDR intrinsically disordered region, ANK Ankyrin repeat; Chromo, Chromodomain. **g** RT-qPCR after in vitro binding assay of purified MPP8 with in vitro transcribed *HIF1α-AS1* RNAs. MPP8 antibodies were used for RNA immunoprecipitation (RIP). An T7-MCS in vitro transcribed RNA served as negative control (CTL). FL, full length; E1, Exon1; E2, Exon2. Δ indicates the deleted nt from *HIF1α-AS1* full length. *n* = 4 independent experiments, One-Way ANOVA with Dunnett's test. *FL ($p = 0.018$), *ΔE2 117-652 ($p = 0.0453$). **h** RT-qPCR after in vitro binding assay of different in vitro translated His-tagged MPP8 mutants with in vitro transcribed *HIF1α-AS1* full length. Anti-His was used for RNA immunoprecipitation. *n* = 4 independent experiments, One-Way ANOVA with Dunnett's test. *($p = 0.0029$). **i, j** Proximity ligation assay of HUVECs with antibodies against MPP8 and H3K9me3 (**i**) or MPP8 and SETDB1 (**j**). The individual antibody alone served as negative control. Red dots indicate polymerase amplified interaction signals. Scale bar indicates 20 µm (**i**) or 10 µm (**j**). Images were representative of three independent experiments. Data are presented as mean values ± SEM.

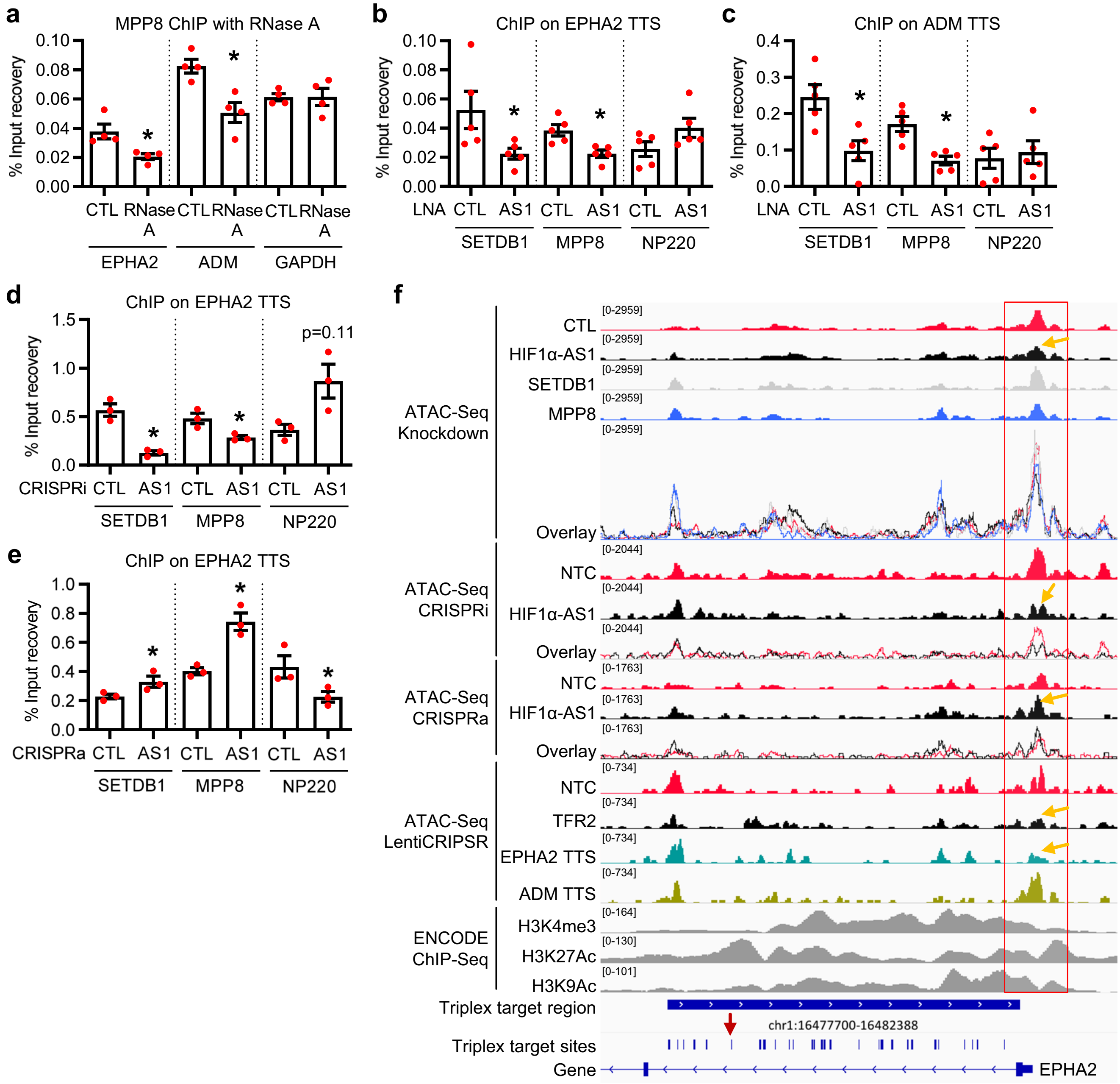


Fig. 6 | *HIF1α-AS1* directs the HUSH complex member MPP8 and SETDB1 to triplex target sites. a Chromatin immunoprecipitation (ChIP) with MPP8 antibodies with or without RNase A treatment and qPCR for the triplex target sites of *EPHA2* and *ADM*. Primers against a promoter sequence of GAPDH served as negative control. $n = 4$ independent experiments, paired t test. *EPHA2 ($p = 0.0223$), *ADM ($p = 0.0221$). **b, c** ChIP with antibodies against SETDB1, MPP8 or NP220 in HUVECs treated with (AS1) or without (CTL) LNA GapmeRs against *HIF1α-AS1*. QPCR was performed for *EPHA2* TTS (b) or *ADM* TTS (c). $n = 5$ independent experiments, paired t test. b: *SETDB1 ($p = 0.0487$), *MPP8 ($p = 0.0287$); c: *SETDB1 ($p = 0.0358$), *MPP8 ($p = 0.0109$). **d, e** ChIP with SETDB1, MPP8 or NP220 antibodies after CRISPRi (d) or CRISPRa (e) for *HIF1α-AS1* and qPCR for *EPHA2* TTS. $n = 3$ independent experiments, paired t test. d: *SETDB1 ($p = 0.0371$), *MPP8 ($p = 0.0117$); e: *SETDB1 ($p = 0.0159$), *MPP8 ($p = 0.0465$), *NP220 ($p = 0.0202$). **f** Genome tracks for *EPHA2* of ATAC-Seq in HUVECs separately and as an overlay after knockdown of *HIF1α-AS1* (black), SETDB1 (gray), MPP8 (blue) or the negative control (red), after CRISPRi and CRISPRa of *HIF1α-AS1* or after LentiCRISPR-mediated deletions of *HIF1α-AS1* TFR2, *EPHA2* TTS or *ADM* TTS. ChIP-Seq data (H3K4me3, H3K27Ac, H3K9Ac) in HUVECs was derived from ENCODE. Numbers in square brackets indicate data range values. Red arrow indicates the TTS analyzed in this study, orange arrows indicate strong changes. The red box indicates the relevant location. NTC non-targeting control, TFR triplex forming region, TTS triplex target site. Data are presented as mean values ± SEM.

Our data indicate that the TSS region of *EPHA2*, where changes in chromatin accessibility were found consistently with knockdown, CRISPRi/a and LentiCRISPR, may contain repressive elements required to control the transcription of *EPHA2*. Other regions, such as the *EPHA2* region downstream of the triplex target sites, showed sensitivity to *HIF1α-AS1*, MPP8 and SETDB1 knockdown, but could not be validated by CRISPRi/a or LentiCRISPR. This finding could indicate that this region is probably sensitive to different transfection methods (electroporation versus transfection reagents) or the use of small RNA molecules, which was not the case in CRISPR experiments. The relevance of that region for the control of *EPHA2* transcription needs to be further clarified with experiments involving region-specific mutations.

In our unbiased approach, a large number of DNA binding sites were identified for *HIF1α-AS1* with triplex domain finder analysis. The large number is not unusual as many of these binding sites overlap and are not identical. Also for other lncRNAs, such as *GATA6-AS*, *FENDRR*, *HOTAIR* and *PARTICLE*, many DNA binding sites have been predicted within their target genes[9,14]. *EPHA2* and *ADM*, as well as *PLEC*, *RP11-276H7.2*, *MIDN* and *EGR1* contained a large number of DNA binding sites for *HIF1α-AS1* and were upregulated after *HIF1α-AS1* knockdown. It is therefore tempting to speculate that similar regulatory mechanisms may play a role in the regulation of these genes. Despite containing triplex target sites, several genes were unaffected by changing the expression level of the triplex-forming RNA. Given the large number of target sites, this could be a consequence of redundancy with respect to target sites or lncRNAs, steric hindering or additional local factors so far unknown. In fact, beyond Hoogsteen base pairing, the local factors required for triplex formation have not yet been identified. For example, it is possible that large protein complexes, like those involved in splicing interfere with binding. Also, the binding of transcription factors could compete with the lncRNA binding. Obviously, also the local epigenetic landscape and chromatin state impacts on triplex binding.

On several occasions, our study takes advantage on the fact that RNase H cleaves the RNA in DNA-RNA heteroduplexes[19], and therefore enriches triplex-forming RNAs within the pool of DNA-interacting RNAs. Although this approach has been widely used in the field[4,6,7,13,15,20,51], it is indirect and therefore not perfect. Proteins or specific local factors may shield RNAs resulting in false positive results and dynamic triplexes, with weak RNA interactions might also be digested. This is why additional methods, in particular bioinformatics prediction of Hoogsteen base pairing and ex vivo demonstration of triplex forming potential are needed to confirm the data obtained with the aid of differential RNase H-digestion.

The evidence for triplex formation by *HIF1α-AS1* is substantiated by a number of findings: Firstly, target recognition by *HIF1α-AS1* occurs via triplex formation involving GA-rich sequences of the DNA targets and GA-rich sequences within *HIF1α-AS1* lncRNA. This has also been observed for other lncRNAs such as *HOTAIR*[52] and *MEG3*[4]. Secondly, the ^{1}H-1D NMR and CD spectroscopy data for *HIF1α-AS1* provided similar but more detailed characteristics for triplex formation, compared with other studies[4,5]. Thorough NMR analysis of attenuations of the individual DNA Watson-Crick base-paired nucleotides allows delineation of those base pairs that are markedly affected by triplex formation. From a total of 25 base pairs in the *EPHA2*_DNA duplex target, only 13 base-pairs are affected. This observation in turn implies that not the entire *HIF1α-AS1* (TFO2-23) RNA is engaged in interaction with the DNA target duplex within the major groove of the DNA duplex, but substantial parts of the RNA strand retain dynamic flexibility which was further assured by our structural modeling. Through the use of heteroduplex samples, measurements at different temperatures, a reduction of equivalents of RNA and triplex analysis with stabilized DNA hairpin sequences, our study allowed an improved and extended analysis of triplex formation. Thirdly, in agreement with previous work[5], most of the triplex target sites were located in the promoter region or introns of the DNA target genes. Fourthly, the triplex formation of *HIF1α-AS1* resulted in gene repression, a finding also observed for other triplex forming RNAs[3]. We could extend this finding by replacing the TFR2 of *HIF1α-AS1* with other sequences, which abolished the repressive effects.

HIF1α-AS1 was downregulated in the lungs of patients with specific forms of pulmonary arterial hypertension (PAH). PAH is characterized by several structural changes, remodeling and lesion development in the pulmonary arteries. A study by Masri *et al.* demonstrated the impairment of pulmonary artery endothelial cells from IPAH patients to form tube-like structures[53]. CTEPH, a complex disorder with major vessel remodeling and small vessel arteriopathy, is characterized by medial hypertrophy, microthrombi formation and plexiform lesions[54]. It has been further shown that TGF-ß-induced angiogenesis was increased by circulating CTEPH microparticles co-cultured with pulmonary endothelial cells, indicating a pro-angiogenic feedback of endothelial injury[55]. Since *HIF1α-AS1* knockdown led to an increase in sprouting, we assume that the loss of *HIF1α-AS1* is a compensatory mechanism, which could be putatively included in the above mentioned pro-angiogenic feedback loop. *HIF1α-AS1* was also reduced in endothelial cells isolated from glioblastoma. Typically this pathology represents a highly angiogenic situation with defective endothelium and abnormal morphology[56]. Additionally, *HIF1α-AS1* is pro-apoptotic[27] and so the reduction of *HIF1α-AS1* could explain the observed sprouting phenotype by the inhibition of apoptosis. Therefore, it is tempting to speculate that *HIF1α-AS1* harbors atheroprotective roles, which could be exploited to alter angiogenesis in patients. Strategies to design such therapeutics require data in other species and in different tissues. *HIF1α-AS1* is not endothelial-specific according to CAGE analysis. A comprehensive analysis on *HIF1α-AS1* conservation, especially of TFR2, is lacking. Initial attempts with BLAT showed that the first 1000 nt of the pre-processed *HIF1α-AS1* including TFR2 were conserved in primates and pigs, but not in rodents (data not shown). A potential application could be the promotion of vascular regeneration after an ischemia damage to promote early blood supply. Indeed, the post-ischemic healing response is not solely dependent on cardiomyocyte loss and adaptation but also on the damage response of the stroma-vascular compartment[57]. *HIF1α-AS1* was downregulated in hypoxia, but upregulated in the damage-relevant re-oxygenation phase. This suggests that specifically in that phase where endothelial proliferation is most needed, *HIF1α-AS1* limits the angiogenic response and therefore advocates itself as a target. Therefore, we propose an anti-HIF1α-AS1 approach to promote the early angiogenic response to promote post-ischemia regeneration.

Additionally, the data indicate that triplex formation could have therapeutic potential. The single nucleotide polymorphism (SNP) rs5002 (chr11:10326521 (hg19)) was found within the triplex target site of *ADM* with phenoscanner, which lists an association with hemoglobin concentration, red blood cell count and hematocrit[58]. Another link between a triplex forming lncRNA and PAH was reported by a massive upregulation of *MEG3* in paSMCs from IPAH patients. This prevented hyperproliferation after *MEG3* knockdown and a reduced apoptosis phenotype of IPAH-paSMCs involving a mechanism with miR-328-3p and IGF1R[59]. Although triplex formation was not studied, another study provided evidence that a ribonucleotide sequence can be used to form a potential triple helix to inhibit gene expression of the *IGF1R* gene in rat glioblastoma cells[60]. *MEG3* is known to impair cell proliferation and to promote apoptosis in glioma cells[61]. This argues that the binding of a lncRNA to DNA is potentially involved in PAH and GBM.

Taken together, the findings presented here highlight an important pathway of a scaffolding lncRNA within an epigenetic-silencer complex that has a crucial role in the regulation of endothelial genes.

Methods

Materials

The following chemicals and concentrations were used: Human recombinant VEGF-A 165 (R&D, 293-VE), Recombinant Human FGF-basic (154 a.a.) (bFGF, Peprotech, 100-18B), RNase A (NEB, EN0531), RNase H (NEB, M0297L), Cycloheximide (Sigma, C1988) and human recombinant TNF-α (Peprotech, 300-01 A). The following antibodies were used: Anti-H3-pan (Diagenode, C15200011), Anti-dsDNA [3519 DNA] (Abcam, ab27156), Anti-DNA-RNA Hybrid [S9.6] (Kerafast, ENH001), Anti-EPHA2 (Bethyl, A302-025-M), Anti-GAPDH (Sigma, G8795), Anti-HSC70/HSP70 (Enzo Life Sciences, ADI-SPA-820), Anti-NONO (Bethyl, A300-587A), Anti-MPP8 (Bethyl, A303-051A-M), Recombinant Anti-6X His tag® antibody [EPR20547] (Abcam, ab213204, ChIP grade), Anti-H3K9me3 (Diagenode, SN-146-100), Anti-SETDB1 (Bethyl, A300-121A, for chromatin immunoprecipitation; Santa

Cruz Biotechnology, ESET (G-4): sc-271488, for Proximity ligation assay) and Anti-ZNF638/NP220 (Bethyl, A301-548A-M).

Cell culture

Pooled human umbilical vein endothelial cells (HUVECs) were purchased from PromoCell (C-12203, Lot No. 405Z013, 408Z014, 416Z042, Heidelberg, Germany) and originate from umbilical cord/ umbilical vein of caucasians (405Z013: 2 males, 1 female; 408Z014: 2 males, 1 female; 416Z042: 2 males, 2 females). HUVECs were cultured in a humidified atmosphere of 5% CO_2 at 37 °C. Fibronectin-coated (356009, Corning Incorporated, USA) dishes were used to culture the cells. Endothelial growth medium (EGM), consisting of endothelial basal medium (EBM) supplemented with human recombinant epidermal growth factor (EGF), EndoCGS-Heparin (PeloBiotech, Germany), 8% fetal calf serum (FCS) (S0113, Biochrom, Germany), penicillin (50 U/mL) and streptomycin (50 μg/mL) (15140-122, Gibco/ Lifetechnologies, USA) was used. For each experiment except ATAC-Seq, at least three different batches of HUVEC from passage 3 were used. In case of hypoxic treatments, cells were incubated in a SciTive Workstation (Baker Ruskinn, Leeds, UK) at 0.1% O_2 and 5% CO_2 for the times indicated.

Human embryonic kidney 293 cells (HEK293) (ATCC, Manassas, USA) and Lenti-X 293 T cells (Takara, 632180, Japan) were cultured in Dulbecco's Modified Eagle Medium High Glucose (Gibco) supplemented with 8% FCS, penicillin (50 U/mL) and streptomycin (50 μg/ mL) (15140-122, Gibco/ Lifetechnologies, USA), in a humidified atmosphere of 5% CO_2 at 37 °C.

Analyses of Triplex-Seq data to identify candidate lncRNAs

Triplex-Seq data of U2OS and HeLa S3 was used from[15], aligned using STAR[62] and peak-calling was performed with MACS2[63]. Peaks were intersected with Ensembl hg38 gene coordinates to produce a list of gene-associated peaks, which was filtered for lncRNAs. The overlap of U2OS and HeLa S3 lncRNAs was filtered for high confidence candidates by applying two cut-off filters for fold enrichment (>10) and -log10(P value peak enrichment) (>20). The *P* value for the enrichment of the peak (i-log10(*P* value peak enrichment)) was calculated using a Poisson distribution to estimate the expected number of reads which should lie within the peak region. The enrichments (fold enrichment, P) are then calculated based on the ratio of observed reads at the peak location (i.e. the real peak) vs. the expected peak. Next, the candidates were filtered for the presence of a nuclear value (>0) in ENCODE and for the presence of a signal (>0) in aorta, artery, lymphatic, microvascular, thoracic, umbilical vein and vein in FANTOM5 CAGE data[21–23]. Subsequently, the remaining candidates (*RMRP*, *HIF1α-AS1*, *RP5-857K21.4*, *SCARNA2* and *SNHG8*) were tested for their non-coding probability with the online tools CPAT[64] and CPC2[65]. Lastly, regions enriched in the Triplex-Seq were manually inspected in the IGV browser to rule out the possibility that the signals belong to overlapping genes.

Total and nuclear RNA isolation, Reverse transcription and RT-qPCR

Total RNA isolation was performed with the RNA Mini Kit (Bio&Sell). Reverse transcription was performed with SuperScript III Reverse Transcriptase (Thermo Fisher) and oligo(dT)23 together with random hexamer primers (Sigma). CopyDNA amplification was measured with RT-qPCR using ITaq Universal SYBR Green Supermix and ROX as reference dye (Bio-Rad, 1725125) in an AriaMX cycler (Agilent). Relative expression of target genes was normalized to ß-*Actin* or *18 S* ribosomal RNA. Expression levels were analyzed by the delta-delta Ct method with the Agilent Aria 1.7 qPCR software. Oligonucleotides used for amplification are listed in Table 1.

For nuclear RNA isolation, cells were resuspended in buffer A1 (10 mM HEPES pH 7.6, 10 mM KCl, 0.1 mM EDTA pH 8.0, 0.1 mM EGTA pH 8.0, 1 mM DTT, 40 μg/mL PMSF) and incubated on ice for 15 min. Nonidet was added to a final concentration of 0.75% and cells were centrifuged (1 min, 4 °C, 16,000 × *g*). The pellet was washed twice in buffer A1, lysed in buffer C1 (20 mM HEPES pH 7.6, 400 mM NaCl, 1 mM EDTA pH 8.0, 1 mM EGTA pH 8.0, 1 mM DTT, 40 μg/mL PMSF) and centrifuged (5 min, 4 °C, 16,000 × *g*). The supernatant was used for RNA isolation with RNA Isolation the RNA Mini Kit (Bio&Sell).

Knockdown procedures

For small interfering RNA (siRNA) treatments, endothelial cells (80–90% confluent) were transfected with GeneTrans II according to the instructions provided by MoBiTec (Göttingen, Germany). The following siRNAs were used: siEPHA2 (Thermo Fisher Scientific, HSS176396), siSETDB1 (Thermo Fisher Scientific, s19112) and siMPP8 (Thermo Fisher Scientific, HSS123184). The stealth siRNA targeting the intron of *HIF1α-AS1* (approx. 100 nt downstream of TFR2) was designed with the Invitrogen BLOCK-iT RNAi designer (Thermo Fisher) and had the following sequence: 5′-GCC TGG TCC CAA ACA TGC ATC ATA T-3′. As negative control, scrambled Stealth RNAi™ Med GC (Life technologies) was used. All siRNA experiments were performed for 48 h.

For Locked nucleic acid (LNA)-GapmeR (Exiqon) treatment, the transfection was performed with the Lipofectamine RNAiMAX (Invitrogen) transfection reagent according to manufacturer's protocol. All LNA-GapmeR transfections were performed for 48 h. LNA-GapmeRs were designed with the Exiqon LNA probe designer and contained the following sequences: *HIF1α-AS1* (1) 5′-GAAAGAGCAAGGAAC A-3′ and as a negative Control 5′-AACACGTCTATACGC-3′.

Protein isolation and western blot analyses

HUVECs were washed in Hanks solution (Applichem) and afterwards lysed with Triton X-100 buffer (20 mM Tris/HCl pH 7.5, 150 mM NaCl, 10 mM NaPPi, 20 mM NaF, 1% Triton, 2 mM Orthovanadat (OV), 10 nM Okadaic Acid, protein-inhibitor mix (PIM), 40 μg/mL Phenylmethylsulfonylfluorid (PMSF)). The cells were centrifuged (10 min, 16,000 × *g*) and protein concentration of the supernatant was determined with the Bradford assay. The cell extract was boiled in Laemmli buffer and equal amounts of protein were separated with SDS-PAGE. The gels were blotted onto a nitrocellulose membrane and blocked in Rotiblock (Carl Roth, Germany). After incubation with the first antibody, infrared-fluorescent-dye-conjugated secondary antibodies (Licor, Bad Homburg, Germany) were used and signals detected with an infrared-based laser scanning detection system (Odyssey Classic, Licor, Bad Homburg, Germany). Images were acquired with Image Studio 5.2 (Licor). The following first antibodies and dilutions were used: Anti-EPHA2 (Bethyl, A302-025-M, 1:1000), Anti-GAPDH (Sigma, G8795, 1:10000), Anti-HSC70/HSP70 (Enzo Life Sciences, ADI-SPA-820, 1:2000) and Anti-NONO (Bethyl, A300-587A, 1:5000). The following secondary antibodies and dilutions were used: IRDye® 680RD Donkey anti-Rabbit IgG Secondary Antibody (LICOR, 926-68073, 1:15000), IRDye® 800CW Donkey anti-Rabbit IgG Secondary Antibody (LICOR, 926-32213, 1:15000), IRDye® 680RD Donkey anti-Mouse IgG Secondary Antibody (LICOR, 926-68072, 1:15000) and IRDye® 800CW Donkey anti-Mouse IgG Secondary Antibody (LICOR, 926-32212, 1:15000).

Human lung samples

The study protocol for tissue donation from human idiopathic pulmonary hypertension patients was approved by the ethics committee (Ethik Kommission am Fachbereich Humanmedizin der Justus Liebig Universität Giessen) of the University Hospital Giessen (Giessen, Germany) in accordance with national law and with Good Clinical Practice/ International Conference on Harmonisation guidelines. Written informed consent was obtained from each individual patient or the patient's next of kin (AZ 31/93, 10/06, 58/15)[66].

Human explanted lung tissues from subjects with IPAH, CTEPH or control donors were obtained during lung transplantation. Samples of

Table 1 | List of primers for RT-qPCR

Name	Forward primer (5′–3′)	Reverse primer (5′–3′)
b-actin	AAAGACCTGTACGCCAACAC	GTCATACTCCTGCTTGCTGAT
HIF1α-AS1 (TFR2)	CCGAAATCCCTTCTCAGCAG	TCTGTGTTTAGCGGCGGAGG
HIF1α-AS1 (E1)	GCCCTCCATGGTGAATCGGTCCCCGCG	CCTTCTCTTCTCCGCGTGTGGAGGGAG
HIF1α-AS1 (E2)	AGGGCTGTTCCATGTTTAGG	GTCTATGGATGCCCACATGC
HIF1α-AS1 (E1-I)	GCCCTCCATGGTGAATCGGTCCCCGCG	CAACCGAAATCCCTTCTCAGCAGCG
RMRP	TCCGCCAAGAAGCGTATCCC	ACAGCCGCGCTGAGAATGAG
SCARNA2	AGTGTGAGTGGACGCGTGAG	AAGTGTAAGCGGGAGGAGGG
RP5-857K21.4	AGAGTGAGGAGAAGGCTTAC	TTCTGAGTCCCAGAGGTTAC
HIF1α	GCTCATCAGTTGCCACTTCC	ACCAGCATCCAGAAGTTTCC
18 S rRNA	CTTTGGTCGCTCGCTCCTC	CTGACCGGGTTGGTTTTGAT
HIF1α-AS1 (TFR1)	TCAGACGAGGCAGCACTGTGCACTGAGG	TCGCTCGCCATTGGATCTCGAGGAACCC
HIF1α-AS1 (TFR3)	GAGCCCTAATCATAGGACTG	AGGGTCTGAGGTTTGAGTTC
KLF10	AGCCAGCATCCTCAACTATC	GCAGCACTTGCTTTCTCATC
SPHK1	GGAGATGCGCTTCACTCTGG	GGAGGCAGGTGTCTTGGAAC
CSRNP1	TGTGGCTGTCACTGCGATAG	TGTGGTCCATCTGGCACTTG
INTS6	GCCTGGCACCATGTCAGTAG	GCACCAAGGACTCCAGACAC
GATA2	GCAACCCCTACTATGCCAACC	CAGTGGCGTCTTGGAGAAG
IER5	AGACCGGGAACGTGGCTAAC	TCTCAGCACCGGCTTATCGC
YWHAZ	GTGTTCTATTATGAGATTCTGAAC	ATGTCCACAATGTCAAGTTGTCTC
THBS1	TGTACGCCATCAGGGTAAAG	AAGAAGGTGCCACTGAAGTC
EGR1	ACCCAGCAGCCTTCGCTAAC	AGAAGCGGCGATCACAGGAC
MIDN	AAGACACCCGGCTCAGTTCG	TGAGACATGAGGCCCGCTTC
EPHA2	GGCTGAGCGTATCTTCATTG	ACTCGGCATAGTAGAGGTTG
RP11-276H7.2	CCAGACTCCCTTTGCCTACC	GCAGAGAAGACCCACGTACC
PLEC	CCAAGGGCATCTACCAATCC	CACTCCAGCCTCTCAAACTC
ADM	TTCCGTCGCCCTGATGTACC	ATCCGCAGTTCCCTCTTCCC
TGFBR1	GAGCGGTCTTGCCCATCTTC	TTCAGGGGCCATGTACCTTTT

donor lung tissue were taken from the lung that was not transplanted. All lungs were reviewed for pathology and the IPAH lungs were classified as grade III or IV.

PASMC isolation and culture

Pulmonary arterial smooth muscle cells (PASMCs) were handled and treated as described before[67]. Briefly, segments of PASMCs, which were derived from human pulmonary arteries (<2 mm in diameter) of patients with IPAH or from control donors, were cut to expose them to the luminal surface. Gentle scraping with a scalpel blade was used to remove the endothelium. The media was peeled away from the underlying adventitial layer. 1–2 mm^2 sections of medial explants were cultured in Promocell smooth Muscle Cell Growth Medium 2 (Promocell, Heidelberg, Germany). For each experiment, cells from passage 4-6 were used. A primary culture of human PASMCs was obtained from Lonza (CC-2581, Basel, Switzerland), grown in SmGM-2 Bulletkit medium (Lonza) and cultured in a humidified atmosphere of 5% CO_2 at 37 °C. Cells from passages 4–6 were used for experiments. For hypoxia experiments, PASMCs were incubated in hypoxia or normoxia chambers for 24 h in hypoxic medium (basal medium containing 1% FCS for human PASMCs). Hypoxia chambers were equilibrated with a water-saturated gas mixture of 1% O_2, 5% CO_2, and 94% N_2 at 37 °C.

Brain microvessel isolation from glioblastoma (GBM) patients

Studies for human glioblastoma were covered by an ethics statement according to the guidelines of the University of Frankfurt, whose approval number for autopsy material is GS-249/11 and for resection material GS-04/09. Human Brain microvessel (HMBV) isolation from GBM patients was performed exactly as described before[39]. Within 3 h post surgery, fresh brain specimens were obtained from GBM patients. For patients without available normal appearing healthy tissue, healthy material was obtained from epilepsy or dementia patients or autopsy material within a day postmortem. To isolate HMBV, specimens obtained in ice-cold MVB (15 mM HEPES, 147 mM NaCl, 4 mM KCl, 3 mM $CaCl_2$, 1.2 mM $MgCl_2$, 5 mM glucose and 0.5% BSA, pH 7.4) were used. These were cleared using forceps and the tissue was homogenized in 3-fold ice-cold MVB buffer by 15 up and down strokes in a tight-fitting douncer (0.25 mm clearance, 10 mL Wheaton) attached to an electrical overhead stirrer (2000 rpm, VOS 14, VWR). The homogenate was centrifuged (400 × *g*, 10 min, 4 °C) and the pellet was resuspended in fourfold 25% BSA (in PBS). After an additional centrifugation (2000 × *g*, 30 min, 4 °C), myelin fat in the top layer was aspirated. Next, the pellet containing the microvessels was resuspended in 3 mL ice-cold MVB/ gram starting material. To remove large vessels and tissue aggregates, the sample was filtered through 100-micron sterile nylon mesh cell strainer (BD) and the microvessels were trapped onto a 40-micron sterile nylon mesh (BD). Afterwards, the mesh was washed once with ice-cold MVB and the microvessels were lysed directly with ice-cold RLT-Plus RNA lysis buffer (Qiagen), vortexed and stored at −80 °C until use.

CRISPR/dCas9 activation (CRISPRa) and inactivation (CRISPRi)

Guide RNAs (gRNA) were designed with the help of the web-interfaces of CRISPR design (http://crispr.mit.edu/). CRISPR activation (CRISPRa) was performed with a catalytically inactive Cas9 (dCas9), which is fused to the transcription activator VP64 (pHAGE EF1α dCas9-VP64), whereas CRISPRi was performed with a dCas9 fusion to the KRAB repressive domain. Both were used together with a sgRNA(MS2) vector containing the individual guide RNA (gRNA) to induce or repress *HIF1α-AS1* gene expression. pHAGE EF1α dCas9-VP64 and pHAGE EF1α dCas9-KRAB were a gift from Rene Maehr and Scot Wolfe (Addgene plasmid # 50918, # 50919)[68] and sgRNA(MS2) cloning backbone was a gift from Feng Zhang (Addgene plasmid # 61424)[69]. The following oligonucleotides were used for cloning of the guide RNAs into the sgRNA(MS2) vector: For CRISPRa of *HIF1α-AS1* 5′-CACCGGGGCCGGCCTCGGCGTTAAT-3′ and 5′-AAACATTAACGCCGAGGCCGGCCCC-3′, and for CRISPRi of *HIF1α-AS1* 5′-CACCGGTCTGGTGAGGATCGCATGA-3′ and 5′-AAACTCATGCGATCCTCACCAGACC-3′. After cloning, plasmids were purified and sequenced. The transfection of the plasmids in HUVEC was performed using the NEON electroporation system (Invitrogen).

CRISPR-Cas9 genome editing with LentiCRISPR

Guide RNAs (gRNA) were selected using the publicly available CRISPOR algorithm 5.01 (http://crispor.tefor.net/)[70]. A dual gRNA approach consisting of gRNA-A and gRNA-B was used to facilitate the individual deletions. The gRNAs were cloned into lentiCRISPRv2 vector backbone with Esp3I (Thermo Fisher, FD0454) according to the standard protocol[71]. lentiCRISPRv2 was a gift from Feng Zhang (Addgene plasmid #52961; http://n2t.net/addgene:52961; RRID:Addgene_52961)[71].

For annealing, the following oligonucleotides were used: *HIF1α-AS1 TFR2*: gRNA-A, 5′-CACCGGCTCGTCTGTGTTTAGCGG-3′ and 5′-AAACCCGCTAAACACAGACGAGCC-3′, gRNA-B, 5′-CACCGGTGCGGCTCAGCCCGAGTC-3′ and 5′-AAACGACTCGGGCTGAGCCGCACC-3′; *EPHA2 TTS*: gRNA-A, 5′-CACCGTTGCATAGGTTCTATGCCC-3′ and 5′-AAACGGGCATAGAACCTATGCAAC-3′, gRNA-B, 5′- CACCGAAGTGCTACCCTCCCTAGA-3′ and 5′-AAACTCTAGGGAGGGTAGCACTTC-3′; *ADM TTS*: gRNA-A, 5′- CACCGCCGAGAGCAGGAGCGCGCG-3′ and 5′-AAACCGCGCGCTCCTGCTCTCGGC-3′, gRNA-B, 5′- CACCGCGCGTGGCTGAGGAAAGAA-3′ and 5′-AAACTTCTTTCCTCAGCCACGCGC-3′. After cloning, the gRNA-containing LentiCRISPRv2 vectors were sequenced and purified. Lentivirus was produced in Lenti-X 293 T cells (Takara, 632180) using Polyethylenamine (Sigma-Aldrich, 408727), psPAX2 and pVSVG (pMD2.G). pMD2.G was a gift from Didier Trono (Addgene plasmid #12259; http://n2t.net/addgene:12259; RRID:Addgene_12259). psPAX2 was a gift from Didier Trono (Addgene plasmid #12260; http://n2t.net/addgene:12260; RRID:Addgene_12260). LentiCRISPRv2-produced virus was transduced in HUVEC with polybrene transfection reagent (Merck-Millipore, TR-1003-G) and selection was performed with puromycin (1 µg/mL) for 6 d. Afterwards, genomic DNA was isolated, PCR was performed followed by agarose gel electrophoresis and ethidium-bromide staining. The following primers were used: *HIF1α-AS1* TFR2 del, 5′-GCGGAGGAAAGAGAAAGGAG-3′ and 5′-GAACAGAGAGCCCAGCAGAG-3′; EPHA2 TTS del, 5′-TCTCCTTACCCTCTAGGGAG-3′ and 5′-ATTCTAGGCCCAGAGACCAG-3′; ADM TTS del, 5′-GCGTGGCTGAGGAAAGAAAG-3′ and 5′-GAGAGTGATCTGCCAAGTAC-3′; GAPDH, 5′-TGGTGTCAGGTTATGCTGGGCCAG-3′ and 5′- GTGGGATGGGAGGGTGCTGAACAC-3′.

CRISPR-Cas9 ArciTect genome editing

For genome editing, the ArciTect Cas9-eGFP system was used according to the manufacturer's conditions (STEMCELL Technologies, Köln, Germany). Briefly, ArciTect™ CRISPR-Cas9 RNP Complex solution was generated with 60 µM gRNA and tracrRNA and 3.6 µg ArciTect™ Cas9-eGFP Nuclease. Afterwards, 20 µM single-strand oligodeoxynucleotide (ssODN) was added to the RNP solution. The following gRNA was used to target TFR2 of *HIF1α-AS1*: 5′-ACGTGCTCGTCTGTGTTTAG-3′. The following ssODNs (Integrated DNA Technologies, Leuven, Belgium) were used to replace TFR2: MEG3, 5′-GAG GCACAGCTGGGACGGGCTGCGACGCTCACGTGCTCGTCTGTGTTGTAATCGCTCCCTCTCTGCTCTCCGATGGGGGTGCGGCTCAGCCCGAGTCTGGGGACTCTGCGCCTTCTCCGAAGGAAGGCGG-3′, negative control Luc 5′-GCTGAGGCACAGCTGGGACGGGCTGCGACGCTCACGTGCTCGTCTGTGTTGTAATTATCACGCTCGTCGTTCGGTATGATGGGGGTGCGGCTCAGCCCGAGTCTGGGGACTCTGCGCCTTCTCCGAAGGAAG-3′. 400.000 HUVECs were seeded in a 12-well plate and electroporated in E2 buffer with the NEON electroporation system (Invitrogen) (1,400 V, 1 × 30 ms pulse). A full medium exchange was done every 24 h and cells were incubated for 72 h. For FACS, eGFP-positive cells were sorted in PBS supplemented with 5% FCS with a Cell Sorter SH800S (Sony).

HIF1α-AS1 mutants, pCMV6-MPP8-10xHis and MPP8 mutants

To clone pcDNA3.1 + HIF1α-AS1, *HIF1α-AS1* was amplified with PCR from cDNA (forward primer: 5′-ATATTAGGTACCCGCCGCCGGCGCCCTCCATGGTG-3′, reverse primer: 5′-ACGGGAATTCTAATGGAACATTTCTTCTCCCTAG-3′) and insert and vector (pcDNA3.1+) were digested with Acc65I/EcoRI and ligated. pCMV6-MPP8-MYC-DDK was obtained from Origene (#RC202562L3). The plasmid pcDNA3.1 + HIF1α-AS1_1200 (called TFR2) included the first 1200 nt (hg19, chr14:62,161,342-62,162,541) of the genomic DNA of the *HIF1α-AS1* gene and was synthesized from Biomatik (Canada).

To create pcDNA3.1 + HIF1-AS1-Δexon1 (1-116), pcDNA3.1 + HIF1-AS1-Δexon2 (117-652), pcDNA3.1 + HIF1-AS1-Δexon1 (26-78) and pCMV6-MPP8-10xHIS (replacement of c-terminally MYC-DDK by 10xHIS), site-directed mutagenesis was performed with the Q5 Site-Directed Mutagenesis Kit (NEB) according to the instructions of the manufacturer. Oligonucleotides and annealing temperatures for mutagenesis were calculated with the NEBaseChanger online tool from NEB. The pcDNA3.1 + HIF1α-AS1 and pCMV6-MPP8-Myc-DDK plasmids served as templates and were amplified with PCR with the following oligonucleotides to obtain the individual constructs: for pcDNA3.1 + HIF1α-AS1-Δexon1 (1-116), 5′-ACTACAGTTCAACTGTCAATTG-3′ and 5′-GGTACCAAGCTTAAGTTTAAAC-3′, for pcDNA3.1 + HIF1-AS1-Δexon2 (117-652), 5′-GAATTCTGCAGATATCCAG-3′ and 5′-CTTTCCTTCTCTTCTCCG-3′, for pcDNA3.1 + HIF1α-AS1-Δexon1 (26-78), 5′-AGCGCTGGCTCCCTCCAC-3′ and 5′-TTCACCATGGAGGGCGCC-3′, for pCMV6-MPP8-10xHIS, 5′-CACCATCATCACCACCATCACTAAACGGCCGGCCGCGGTCAT-3′ and 5′-GTGATGGTGAGAGCCTCCACCCCCCTGCAGCTGCACTCTGTATGCACCTATTAGC-3′. The plasmids were verified by sequencing.

The individual MPP8 mutants were generated with the Q5 Site-Directed Mutagenesis Kit (#E0554S, NEB) according to the instructions of the manufacturer. To generate primer sequences and calculate annealing temperatures, the NEBaseChanger™ (NEB) was used. The pCMV6-MPP8-10xHis plasmid served as template and was amplified with PCR with the following oligonucleotides to obtain the individual mutants: W80A, 5′-CAAAGTTCGCgcGAAAGGCTATAC-3′ and 5′-TAAAGAACTTTACCCCCC-3′, Δ55-118, 5′-AGGAAGGATATTCAGAGACTATCC-3′ and 5′-GTCCTCCTCACTGTCGCC-3′, Δ2-441, 5′-AAGGAAATCAGAAATGCATTTGATTTATTTAAATTAACTCCAGAAGAAAAAAATGATGTTTCTG-3′ and 5′-CATGGCGATCGCGGCGGC-3′, Δ442-860, 5′-GGGGGTGGAGGCTCTCAC-3′ and 5′-AAGTGTCTTTAATCCTTTTGGCTCTTTTCTG-3′, Δ600-728, 5′-GTAGCAGAAGAGACAATAAAG-3′ and 5′-GGAATCCTCTTGGTCCAG-3′. The final plasmids were verified by sequencing. In vitro protein synthesis of the MPP8 mutants was performed with the PURExpress kit (E6800, NEB) according to the manufacturer's protocol.

Purification of pCMV6-MPP8-10xHis

To generate purified MPP8-10xHIS protein, pCMV6-MPP8-10xHIS was overexpressed in HEK293 by transfection with Lipofectamine 2000 according to the manufacturer's protocol. After 24 h, cells were lysed with three cycles of snap freezing in liquid nitrogen and 2% triton X-100 with protease inhibitors. Recombinant MPP8-10xHis was purified using HisTrap FF crude columns (Cytiva Europe, Freiburg, Germany, #11000458) with a linear gradient of imidazole (from 20 to 500 mM, Merck, Burlington, United States, #104716) in an Äkta Prime Plus FPLC system (GE Healthcare/Cytiva Europe).

In vitro transcription and RNA 3′end biotinylation

Prior to in vitro transcription, pcDNA3.1 + HIF1α-AS1, pcDNA3.1 + HIF1α-AS1-Δexon1 (1-116), pcDNA3.1 + HIF1α-AS1-Δexon2 (117-652), pcDNA3.1 + HIF1α-AS1-Δexon1 (26-78) or control pcDNA3.1+ were linearized with SmaI (Thermo Fisher, FD0663). After precipitation and purification of linearized DNA, RNA was in vitro transcribed according to the manufacturers protocol with T7 Phage RNA Polymerase (NEB), and DNA was digested with RQ DNase I (Promega). The remaining RNA was purified with the RNeasy Mini Kit (Qiagen) and used for binding reactions with MPP8-10xHis in RIP experiments. For RNA pulldown experiments, RNA of *HIF1α-AS1* or of the control pcDNA3.1+ were further biotinylated at the 3′end with the Pierce RNA 3′end biotinylation kit (Thermo Fisher).

RNA pulldown assay and mass spectrometry

The RNA pulldown assay was performed similar to[39]. For proper RNA secondary structure formation, 150 ng of 3′end biotinylated *HIF1α-AS1* or control RNA was heated for 2 min at 90 °C in RNA folding buffer (10 mM Tris pH 7.0, 0.1 M KCl, 10 mM $MgCl_2$), and then put on RT for 20 min. 1×10^7 HUVECs were used per sample. Isolation of nuclei was performed with the truCHIP™ Chromatin Shearing Kit (Covaris, USA) according to the manufacturers protocol without shearing the samples. Folded Bait RNA was incubated in nuclear cell extracts for 3 h at 4 °C. After incubation, samples were UV crosslinked. Afterwards, Streptavidin M-270 Dynabeads (80 μL Slurry, Thermo Fisher) were incubated with cell complexes for 2 h at 4 °C. After 4 washing steps with the lysis buffer of the truCHIP chromatin Shearing Kit (Covaris, USA), beads were put into a new Eppendorf tube. For RNA analysis, RNA was extracted with TRIzol (Thermo Fisher). Afterwards, RNA purification was performed with the RNeasy Mini Kit (Qiagen). If indicated, RT-qPCR was performed. For mass spectrometric measurements in order to reduce complexity, samples were eluted stepwise from the beads. Beads were resuspended in 50 mM ammoniumhydrogencarbonate and 1 μL RNAse A. Supernatant was reduced and alkylated with DTT and chloracetamid, respectively. Remaining Beads were resuspended in 20 μL 6 M Guanidinhydrochlorid (GdmCl), 100 mM Tris/HCl, pH 8.5, 10 mM DTT and incubated at 95 °C for 5 min. Reduced thiols were alkylated with 40 mM chloroacetamid and samples were diluted with 25 mM Tris/HCl, pH 8.5, 10% acetonitrile to obtain a final GdmCl concentration of 0.6 M. Proteins of both fractions were digested with 1 μg Trypsin/LysC (sequencing grade, Promega) overnight at 37 °C under gentle agitation. Digestion was stopped by adding trifluoroacetic acid to a final concentration of 0.1%. Peptides were loaded on multi-stop-and-go tip (StageTip) containing three a stack of three C18-disks. Both fractions were eluted in wells of microtiter plates and peptides were dried and resolved in 1% acetonitrile, 0.1% formic acid. Liquid chromatography/mass spectrometry (LC/MS) was performed on Thermo Scientific™ Q Exactive Plus equipped with an ultra-high performance liquid chromatography unit (Thermo Scientific Dionex Ultimate 3000) and a Nanospray Flex Ion-Source (Thermo Scientific). Peptides were loaded on a C18 reversed-phase precolumn (Thermo Scientific) followed by separation on a with 2.4 μm Reprosil C18 resin (Dr. Maisch GmbH) in-house packed picotip emitter tip (diameter 100 μm, 15 cm long from New Objectives) using an gradient from mobile phase A (4 % acetonitrile, 0.1% formic acid) to 40% mobile phase B (80% acetonitrile, 0.1% formic acid) for 60 min followed by a second gradient to 80% B for 30 min with a flow rate 400 nL/min. Run was finished by washout with 99% B for 5 min and reequilibration in 1% B. MS data were recorded by data dependent acquisition Top 10 method selecting the most abundant precursor ions in positive mode for HCD fragmentation. The Full MS scan range was 300 to 2000 m/z with resolution of 70000, and an automatic gain control (AGC) value of 3E6 total ion counts with a maximal ion injection time of 160 ms. Only higher charged ions (2 +) were selected for MS/MS scans with a resolution of 17500, an isolation window of 2 m/z and an automatic gain control value set to E5 ions with a maximal ion injection time of 150 ms. Selected ions were excluded in a time frame of 20 s following fragmentation event. Fullscan data were acquired in profile and Fragments in centroid mode by Xcalibur software. For data analysis MaxQuant 1.5.3.30 and Perseus 1.5.4.1 were used. The enzyme specificity was set to Trypsin, missed cleavages were limited to 2. Following variable modifications were selected: at N-terminus acetylation (+42.01), oxidation of methionine (+15.99), as fixed modification carbamidomethylation (+57.02) on cysteines. Human reference proteome set from Uniprot (Download 4/2015, 68506 entries) was used to identify peptides and proteins. False discovery rate (FDR) was set to 1 %. Protein group file was uploaded to Perseus and data set was cleaned from reverse identifications and common contaminants. Data were Log2 transformed. Identification were filtered for 4 valid values in at least one group. To enable calculation of ratios between sample and control, missing values were replaced from normal distribution. Positive hits from *p* values ($p < 0.05$) of students *t* test between experimental groups were highlighted. The samples were labeled H1-H5 for HIF1α-AS1 and C1-C5 for the negative control RNA. MaxQuant 1.5.3.30 and Perseus 1.5.4.1 were used to analyze the data.

RNA immunoprecipitation

1×10^7 HUVECs were used per sample. Nuclei isolation was performed with the truCHIP™ Chromatin Shearing Kit (Covaris, USA) according to the manufacturers protocol without shearing the samples. After pre-clearing with 20 μL DiaMag Protein A and Protein G (Diagenode), 10% of the pre-cleared sample served as input and the lysed nuclei were incubated with the indicated antibody or IgG alone for 12 h at 4 °C. The following antibodies and dilutions were used: Anti-H3-pan (Diagenode, C15200011, 1:200), Anti-dsDNA [35I9 DNA] (Abcam, ab27156, 1:200), Anti-DNA-RNA Hybrid [S9.6] (Kerafast, ENH001, 1:250), Anti-MPP8 (Bethyl, A303-051A-M, 1:250) and Anti-H3K9me3 (Diagenode, SN-146-100, 1:200). The complexes were then incubated with 50 μL DiaMag Protein A and Protein G (Diagenode) beads for 3 h at 4 °C, followed by 4 washing steps in Lysis Buffer from the truCHIP™ Chromatin Shearing Kit (Covaris, USA). In case of RNase treatments, the samples were washed once in TE-buffer and then incubated for 30 min at 37 °C in buffer consisting of 50 mM Tris-HCl pH 7.5-8.0, 150 mM NaCl, 1 mM $MgCl_2$ containing 2 μL RNase H per 100 μL buffer. Afterwards the samples were washed in dilution buffer (20 mmol/L Tris/HCl pH 7.4, 100 mmol/L NaCl, 2 mmol/L EDTA, 0.5% Triton X-100, 1 μL Superase In (per 100 μL) and protease inhibitors). Prior to elution, beads were put into a new Eppendorf tube. RNA was extracted with TRIzol (Thermo Fisher) followed by RNA purification with the RNeasy Mini Kit (Qiagen), reverse transcription and qRT-PCR.

For the in vitro RIP assays, the individual RNAs were folded as mentioned above in RNA folding buffer (10 mM Tris pH 7.0, 0.1 M KCl, 10 mM $MgCl_2$), and then put on RT for 20 min. The binding reactions with purified MPP8-10xHIS or in vitro translated His-tagged mutants were performed for 2 h at 4 °C in binding buffer (20 mmol/L Tris/HCl pH 8.0, 150 mmol/L KCl, 2 mmol/L EDTA pH 8.0, 5 mmol/L $MgCl_2$, 2 μL/mL Superase In and protease inhibitors). After pre-clearing with 20 μL DiaMag Protein A and Protein G (Diagenode), 5% of the pre-cleared sample served as input. The mixture was incubated with Anti-MPP8 (Bethyl, A303-051A-M, 1:250) or Recombinant Anti-6X His tag® antibody [EPR20547] (Abcam, ab213204, ChIP grade, 1:500) for 3 h at 4 °C. The complexes were then incubated with 50 μL DiaMag Protein A and Protein G (Diagenode) beads for 1 h at 4 °C, followed by 4 washing steps (5 min, 4 °C, each) in binding buffer. Elution, RNA extraction and RT-qPCR were performed as mentioned above. RT-qPCR was performed with primers targeting the remaining multiple cloning site (MCS) within the in vitro transcribed sequences before (5′-GTG CTGGATATC TGCAGAATTC-3′) and after (5′-GTGCTGGATATCTGCA GAATTC-3′) the *HIF1α-AS1* sequences.

Table 2 | DNA oligos used for ^{1}H-1D NMR, CD and melting curve analysis analysis

Name	Sequence (5′–3′)	Size	Genomic location (hg19)
EPHA2 (GA-rich)	GGTTTCTCCTATCTCCTTACCCTCT	25 nt	chr1:16,478,543-16,478,567
EPHA2 (CT-rich)	AGAGGGTAAGGAGATAGGAGAAACC	25 nt	chr1:16,478,543-16,478,567
EPHA2-hairpin	GGTTTCTCCTATCTCCTTACCCTCTTTTTTAGAGGGTAAGGAGATAGGAGAAACC	55 nt	chr1:16,478,543-16,478,567
ADM (CT-rich)	TCTTTCCTCAGCCAC	15 nt	chr11:10,326,521-10,326,535
ADM (GA-rich)	GTGGCTGAGGAAAGA	15 nt	chr11:10,326,521-10,326,535
ADM-hairpin	TCTTTCCTCAGCCACTTTTTGTGGCTGAGGAAAGA	35 nt	chr11:10,326,521-10,326,535

The constructs used were not smaller than the predicted core TFR2:TTS regions.

Assay for transposase accessibility (ATAC)-sequencing

ATAC-Seq was performed similar to[39]. 100.000 HUVECs were used for ATAC library preparation using Tn5 Transposase from Nextera DNA Sample Preparation Kit (Illumina). Cell pellets were resuspended in 50 μL PBS and mixed with 25 μL TD-Buffer, 2.5 μL Tn5, 0.5 μL 10% NP-40 and 22 μL H_2O. The mixture was incubated at 37 °C for 30 min followed by 30 min at 50 °C together with 500 mM EDTA pH 8.0 for optimal recovery of digested DNA fragments. 100 μL of 50 mM $MgCl_2$ was added for neutralization. The DNA fragments were purified with the MinElute PCR Purification Kit (Qiagen). Amplification of library together with indexing was performed as described elsewhere[72]. Libraries were mixed in equimolar ratios and sequenced on NextSeq500 platform using V2 chemistry and assessed for quality by FastQC. Reaper version 13-100 was employed to trim reads after a quality drop below a mean of Q20 in a window of 5 nt[73]. Only reads above 15 nt were cleared for further analyses. These were mapped versus the hg19 version of the human genome with STAR 2.5.2b using only unique alignments to exclude reads with uncertain arrangement. Reads were further deduplicated using Picard 2.6.0 (Picard: A set of tools (in Java)[74] for working with next generation sequencing data in the BAM format) to avoid PCR artefacts leading to multiple copies of the same original fragment. The Macs2 peak caller (version 2.1.0)[63] as employed in punctate mode to accommodate for the range of peak widths typically expected for ATAC-seq. The minimum qvalue was set to −4 and FDR was changed to 0.0001. Peaks overlapping ENCODE blacklisted regions (known misassemblies, satellite repeats) were excluded. Peaks were annotated with the promoter (TSS + /− 5000 nt) of the gene most closely located to the center of the peak based on reference data from GENCODE v19. To compare peaks in different samples, significant peaks were overlapped and unified to represent identical regions. The counts per unified peak per sample were computed with BigWigAverageOverBed (UCSC Genome Browser Utilities, http://hgdownload.cse.ucsc.edu/downloads.html). Raw counts for unified peaks were submitted to DESeq2 (version 1.14.1) for normalization[75]. Spearman correlations were produced to identify the degree of reproducibility between samples using R. To permit a normalized display of samples in IGV, the raw BAM files were normalized for sequencing depth (number of mapped deduplicated reads per sample) and noise level (number of reads inside peaks versus number of reads not inside peaks). Two factors were computed and applied to the original BAM files using bedtools genomecov resulting in normalized BigWig files.

For samples used after siRNA-mediated silencing of MPP8 and *SETDB1* as well as the corresponding LNA GapmeR knockdown of *HIF1α-AS1* or samples from CRISPRa, CRISPRi and LentiCRISPR experiments, the improved OMNI-ATAC protocol[76] was used and samples were sequenced on a Nextseq2000. The resulting data were trimmed and mapped using Bowtie2[77]. Data were further processed using deepTools[78]. For visualization, the Integrative Genomics Viewer[79] was used.

Electrophoretic mobility shift assay (EMSA)

DNA:DNA:RNA triplex samples were analyzed with EMSA using native RNA-PAGE. The samples were prepared with 50% glycerol with 0.3 to 0.5 μM concentration. Native-PAGE gels were prepared using 15% (v/v) polyacrylamide and TA buffer (50 mM Tris/acetate, 50 mM sodium acetate, pH 8.3). Bands were separated at constant power (<1 W) for 5 h and were stained with GelRed® (Biotium, USA) and visualized with the gel documentation imager Gel Doc XR + (Bio-Rad, USA). The following DNA sequences (Dharmacon) were used for triplex target sites: *EPHA2*_3_GA, 5′-AGAGGGTAAGGAGATAGGAGAAACC-3′ and *EPHA2*_3_CT, 5′-GGTTTCTCCTATCTCCTTACCCTCT-3′. HIF1α-AS1-TFR2 (TFO2-23) had the following sequence: 5′-GCGGCGGAGGAAAGAGAAAGGAG-3′.

RNA and DNA Hybridization

By hybridization of the RNA strand to the DNA duplex or DNA hairpin DNA:DNA:RNA triplexes were formed. First the complementary DNA single strands were incubated at 95 °C for 5 min in hybridization buffer (25 mM HEPES, 50 mM NaCl, 10 mM $MgCl_2$ (pH 7.4)) and afterwards cooled down to RT. Triplex formation was performed by adding RNA to previously hybridized double stranded DNA for 1 h at 60 °C and then cooled down to RT[13]. For the ^{1}H-1D NMR, CD and melting curve experiments, the *HIF1α-AS1*-TFR2 (TFO2-23) sequence 5′-GCGGCGGAGGAAAGAGAAAGGAG-3′ (length 23 nt, GC = 50.9%) was used in combination with the DNA sequences listed in Table 2.

CD spectroscopy and melting curve analysis

Circular dichroism spectra were acquired on a Jasco J-810 spectropolarimeter. The measurements were recorded from 210 to 320 nm at 25 °C using 1 cm path length quartz cuvette. CD spectra were recorded on 8 μM samples of each DNA duplex, DNA:RNA heteroduplex and DNA:DNA:RNA-triplex in 25 mM HEPES, 50 mM NaCl, 10 mM $MgCl_2$ (pH 7.4). Spectra were acquired with 8 scans and the data was smoothed with Savitzky-Golay filters. Observed ellipticities recorded in millidegree (mdeg) were converted to molar ellipticity $[\theta] = \text{deg x cm}^2 \text{ x dmol}^{-1}$. Melting curves were acquired at constant wavelength using a temperature rate of 1 °C/min in a range from 5 °C to 95 °C. All melting temperature data was converted to normalized ellipticity and evaluated by the following Eq. (1) using SigmaPlot 12.5:

$$f = \frac{a}{(1 + \exp(-\frac{(x-x0)}{b}))} + \frac{c}{(1 + \exp(-\frac{(x-x2)}{d}))} \qquad (1)$$

NMR spectroscopy

All NMR samples were prepared in NMR buffer containing 25 mM HEPES-d18, 50 mM NaCl, 10 mM $MgCl_2$ (pH 7.4) with addition of 5 to 10% D_2O. All samples were internally referenced with 2,2-dimethyl-2-silapentane-5-sulfonate (DSS). The final NMR sample concentrations ranged between 50 μM to 300 μM. NMR spectra were recorded in a temperature range from 278 K to 308 K on Bruker 600, 800, 900 and 950 MHz spectrometers. ^{1}H NMR spectra were recorded with jump-return-Echo[80] and gradient-assisted excitation sculpting[81] for water suppression. 2D ^{1}H,^{1}H-NOESY spectra were recorded with jump-return-Echo[80] water suppression on a Bruker 800 MHz spectrometer at 288 K and mixing times of 150 ms. NMR data were collected, processed and analyzed using TopSpin 3.6.2 (Bruker) and Sparky 3.115[82].

Structural modeling

Models of the DNA:DNA:RNA triplex were generated by using the ARIA/CNS software packages[83–85]. To generate and keep the B-form DNA duplex, ample modeling distances and dihedral angle restraints were used. For flexible docking of the RNA on the B-DNA, a starting structure was generated containing a DNA duplex template and an extended RNA molecule. The docking was solely driven by hydrogen-bonds and base-planarity restraints for the triplex. No further restraints were added for the RNA; leaving it fully flexible during the conventional simulated annealing stages with cartesian angle dynamics. In total, 2000 models were generated and the 200 best structures (lowest energy) were used as input for a further refinement in explicit water using the nucleic acid forcefield with OPLS charges and non-bonded parameters[86]. The final ensemble of 20 top-ranked structures was validated and had no violations. Figure production was done by using PyMol 2.5 (Schrödinger, LLC).

Spheroid outgrowth assay

Spheroid outgrowth assays in HUVEC were performed as described in[87]. Briefly, spheroids were generated by making drops containing 400 HUVECs in a methyl cellulose (20%) (Sigma-Aldrich, M-0512)/culture medium (80%) mixture onto a square petri dish (Greiner Bio-One, 688102). The dish was incubated overnight in upside down direction. Afterwards, spheroids were washed gently with PBS and resuspended in a methyl cellulose (88%)/FCS (12%) mixture and embedded in collagen type I (Corning, 354236) with Medium 199 (Sigma-Aldrich, M0650−100ML). Stimulation of Spheroids was performed with VEGF-A 165 (1 ng/mL) or bFGF (3 ng/mL) for 16 h. Images were generated with an Axiovert135 microscope (Zeiss). Sprout numbers and cumulative sprout lengths were quantified by analysis with the AxioVision software 4.8 (Zeiss).

Caspase-3/7 activity assay

The Caspase-3/7 activity assays were carried out using 1×10^6 HUVEC. The assay was performed using SR-FLICA Caspase-3/7 assay Kit (ImmunoChemistry Technologies LLC, 931) following the manufacturer's instructions. Briefly, cells were washed and a 1:5 dilution of FLICA was added in a dilution of 1:30 to the cell suspension. After an incubation of 1 h, cells were washed three times with buffer provided by the kit, counted and diluted to 3000 cells/µL before measuring emission at 595 nm in a TECAN infinite M200 Pro plate reader using the TECAN i-control 3.7.3.0 software (Männedorf, Switzerland).

Proximity ligation assay (PLA)

The PLA was performed as described in the manufacturer's protocol (Duolink II Fluorescence, OLink, Upsalla, Sweden). Briefly, HUVECs were fixed in phosphate buffered formaldehyde solution (4%), permeabilized with Triton X-100 (0.2%), blocked with serum albumin solution (3%) in phosphate-buffered saline, and incubated overnight with Anti-dsDNA [35I9 DNA] (Abcam, ab27156, 1:500), Anti-MPP8 (Bethyl, A303-051A-M, 1:500), Anti-H3K9me3 (Diagenode, SN-146-100, 1:500) or Anti-SETDB1 (Santa Cruz Biotechnology, ESET (G-4): sc-271488, 1:500). Samples were washed and incubated with the respective PLA-probes for 1 h at 37 °C. After washing, samples were ligated for 30 min (37 °C). After an additional washing step, the amplification with polymerase was performed for 100 min (37 °C). The nuclei were stained using DAPI. Images (with Alexa Fluor, 546 nm) were acquired by confocal microscopy (LSM 510, Zeiss) using the ZEN 3.2 software.

Chromatin Immunoprecipitation

Preparation of HUVEC extracts, crosslinking and isolation of nuclei was performed with the truCHIP™ Chromatin Shearing Kit (Covaris, USA) according to the manufacturers protocol. The procedure was similar to[88]. The lysates were sonified with the Bioruptur Plus (10 cycles, 30 s on, 90 s off, 4 °C; Diagenode, Seraing, Belgium). Cell debris was removed by centrifugation and the lysates were diluted 1:3 in dilution buffer (20 mmol/L Tris/HCl pH 7.4, 100 mmol/L NaCl, 2 mmol/L EDTA, 0.5% Triton X-100 and protease inhibitors). Pre-clearing was done with DiaMag protein A and protein G coated magnetic beads (Diagenode, Seraing, Belgium) for 1 h at 4 °C. As indicated, the samples were incubated over night at 4 °C with the following antibodies and dilutions: Anti-DNA-RNA Hybrid [S9.6] (Kerafast, ENH001, 1:250), Anti-MPP8 (Bethyl, A303-051A-M, 1:250), Anti-SETDB1 (Bethyl, A300-121A, 1:250) and Anti-ZNF638/NP220 (Bethyl, A301-548A-M, 1:250). 5% of the samples served as input. The complexes were collected with 50 µL DiaMag protein A and protein G coated magnetic beads (Diagenode, Seraing, Belgium) for 3 h at 4 °C, washed twice for 5 min with each of the wash buffers 1–3 (Wash Buffer 1: 20 mmol/L Tris/HCl pH 7.4, 150 mmol/L NaCl, 0.1% SDS, 2 mmol/L EDTA, 1% Triton X-100; Wash Buffer 2: 20 mmol/L Tris/HCl pH 7.4, 500 mmol/L NaCl, 2 mmol/L EDTA, 1% Triton X-100; Wash Buffer 3: 10 mmol/L Tris/HCl pH 7.4, 250 mmol/L lithium chloride, 1% Nonidet p-40, 1% sodium deoxycholate, 1 mmol/L EDTA) and finally washed with TE-buffer pH 8.0. In case of RNase treatments, the samples were washed once in TE-buffer and then incubated for 30 min at 37 °C in buffer consisting of 50 mM Tris-HCl pH 7.5-8.0, 150 mM NaCl, 1 mM $MgCl_2$ containing 2 µL RNase H or 2 µL RNase A per 100 µL buffer. Elution of the beads was done with elution buffer (0.1 M $NaHCO_3$, 1% SDS) containing 1x Proteinase K (Diagenode, Seraing, Belgium) and shaking at 600 rpm for 1 h at 55 °C, 1 h at 62 °C and 10 min at 95 °C. After removal of the beads, the eluate was purified with the QiaQuick PCR purification kit (Qiagen, Hilden, Germany) and subjected to qPCR analysis. As a negative control during qPCR, primer for the promoter of GAPDH were used. The primers are listed in Table 3.

Triplex domain finder analysis

Triplex formation of *HIF1α-AS1* was predicted using the Triplex Domain Finder 0.13.2 (TDF)[14] with the human pre-spliced *HIF1α-AS1* sequence (NR_047116.1, gene ID 100750246) to target DNA regions around genes

Table 3 | List of primers for ChIP-qPCR

Name	Forward primer (5′–3′)	Reverse primer (5′–3′)
GAPDH promoter	TGGTGTCAGGTTATGCTGGGCCAG	GTGGGATGGGAGGGTGCTGAACAC
EPHA2 TTS	CAGGTAGCTGCCAATAAGTG	AGGGCTTTACCCTCTGAATC
ADM TTS	CGCGTGGCTGAGGAAAGAAAGG	GCTTTATAAGCGCACGGGTGGG
EPHA2 up (12 kb)	TCAGCTGGGAAGCCACTATG	TTGCTGCTGCTCTGTGAGTC
EPHA2 down (5.7 kb)	CCTCGAATGCATACTCTCAG	CATTCTTGTGCGAGGATGTC
ADM up (2.1 kb)	GGAGGTCAAGGACAGCTAGGC	AGCGAGGTACAGTCGCAGAG
ADM down (3.2 kb)	ACGTGCGGTTTAATAAGTTC	TGGCATCTGCAAACTGTTTC

Number in brackets indicate the approximal distance to the *EPHA2* or *ADM* TTS.

with ATAC-Seq peaks upon *HIF1α-AS1* silencing. For annotation of *HIF1α-AS1* triplex forming regions across DNA triplex target sites, genome version hg19 was used. Boxplots of Fig. 2b show the distribution of triplex prediction from 200 randomizations by shuffling the positions of the same DNA target regions in the genome. Enrichment was given at a p-value <0.05.

Statistics and reproducibility

Unless otherwise indicated, data are given as means ± standard error of mean (SEM). Calculations were performed with Prism 8.0 or BiAS.10.12. The latter was also used to test for normal distribution and similarity of variance. For multiple group comparisons ANOVA followed by post hoc testing was performed and multiplicity adjusted *p* values were shown, if indicated. Individual statistics of dependent samples were performed by two-tailed Student's *t* test (paired or unpaired), and if not normally distributed by Mann–Whitney test. *P* values of <0.05 were considered as significant. Unless otherwise indicated, n indicates the number of individual experiments.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The ATAC-Seq data generated in this study have been deposited in the Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) under BioProject ID . The CRISPR ATAC-Seq data generated in this study have been deposited in the NCBI's Gene Expression Omnibus under the GEO Series accession number GSE203252. The mass spectrometry proteomics data about *HIF1α-AS1* interaction partners identified in this study have been deposited to the the ProteomeXchange Consortium via the PRIDE partner repository[89] with identifier PXD023512. Triplex-Seq data was used from[15] and is deposited in NCBI GEO under accession number GSE120850. Ensembl hg38 was used for the identification of candidate lncRNAs from the Triplex-Seq data. FANTOM5 ENCODE CAGE expression data was obtained from FANTOM5 website (Gencode v19)[21–23]. ChIP-Seq datasets were taken from ENCODE[90] and are deposited at NCBI GEO under accession number GSM733673 for HUVEC H3K4me3, for H3K27Ac under accession code GSM733691 and for H3K9Ac under GSM733735. Source data are provided with this paper.

References

1. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).
2. Felsenfeld, G., Davies, D. R. & Rich, A. Formation of a three-stranded polynucleotide molecule. *J. Am. Chem. Soc.* **79**, 2023–2024 (1957).
3. Li, Y., Syed, J. & Sugiyama, H. RNA-DNA triplex formation by long noncoding RNAs. *Cell Chem. Biol.* **23**, 1325–1333 (2016).
4. Mondal, T. et al. MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures. *Nat. Commun.* **6**, 7743 (2015).
5. Jalali, S., Singh, A., Scaria, V. & Maiti, S. Genome-wide computational analysis and validation of potential long noncoding RNA-mediated DNA-DNA-RNA triplexes in the human genome. *Methods Mol. Biol. (Clifton, N. J.)* **2254**, 61–71 (2021).
6. Trembinski, D. J. et al. Aging-regulated anti-apoptotic long non-coding RNA Sarrah augments recovery from acute myocardial infarction. *Nat. Commun.* **11**, 2039 (2020).
7. Postepska-Igielska, A. et al. LncRNA Khps1 regulates expression of the proto-oncogene SPHK1 via triplex-mediated changes in chromatin structure. *Mol. cell* **60**, 626–636 (2015).
8. O'Leary, V. B. et al. PARTICLE, a triplex-forming long ncRNA, regulates locus-specific methylation in response to low-dose irradiation. *Cell Rep.* **11**, 474–485 (2015).
9. O'Leary, V. B. et al. PARTICLE triplexes cluster in the tumor suppressor WWOX and may extend throughout the human genome. *Sci. Rep.* **7**, 7163 (2017).
10. Blank-Giwojna, A., Postepska-Igielska, A. & Grummt, I. lncRNA KHPS1 activates a poised enhancer by triplex-dependent recruitment of epigenomic regulators. *Cell Rep.* **26**, 2904–2915.e4 (2019).
11. Yari, H. et al. LncRNA REG1CP promotes tumorigenesis through an enhancer complex to recruit FANCJ helicase for REG3A transcription. *Nat. Commun.* **10**, 5334 (2019).
12. Grote, P. & Herrmann, B. G. The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. *RNA Biol.* **10**, 1579–1585 (2013).
13. Kalwa, M. et al. The lncRNA HOTAIR impacts on mesenchymal stem cells via triple helix formation. *Nucleic Acids Res.* **44**, 10631–10643 (2016).
14. Kuo, C.-C. et al. Detection of RNA-DNA binding sites in long noncoding RNAs. *Nucleic Acids Res.* **47**, e32 (2019).
15. Sentürk Cetin, N. et al. Isolation and genome-wide characterization of cellular DNA:RNA triplex structures. *Nucleic Acids Res.* **47**, 2306–2321 (2019).
16. Aguilera, A. & García-Muse, T. R loops: from transcription byproducts to threats to genome stability. *Mol. Cell* **46**, 115–124 (2012).
17. Kaur, G. & Dufour, J. M. Cell lines: Valuable tools or useless artifacts. *Spermatogenesis* **2**, 1–5 (2012).
18. Wong, B. W., Marsch, E., Treps, L., Baes, M. & Carmeliet, P. Endothelial cell metabolism in health and disease: impact of hypoxia. *EMBO J.* **36**, 2187–2203 (2017).
19. Cerritelli, S. M. & Crouch, R. J. Ribonuclease H: the enzymes in eukaryotes. *FEBS J.* **276**, 1494–1505 (2009).
20. Zhou, Z., Giles, K. E. & Felsenfeld, G. DNA·RNA triple helix formation can function as a cis-acting regulatory mechanism at the human β-globin locus. *Proc. Natl. Acad. Sci. USA* **116**, 6130–6139 (2019).
21. Forrest, A. R. R. et al. A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
22. Lizio, M. et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16**, 22 (2015).
23. Noguchi, S. et al. FANTOM5 CAGE profiles of human and mouse samples. *Sci. data* **4**, 170112 (2017).
24. Zhao, Y., Feng, G., Wang, Y., Yue, Y. & Zhao, W. Regulation of apoptosis by long non-coding RNA HIF1A-AS1 in VSMCs: implications for TAA pathogenesis. *Int. J. Clin. Exp. Pathol.* **7**, 7643–7652 (2014).
25. Gong, W., Tian, M., Qiu, H. & Yang, Z. Elevated serum level of lncRNA-HIF1A-AS1 as a novel diagnostic predictor for worse prognosis in colorectal carcinoma. *Cancer biomarkers: Sect. A Dis. markers* **20**, 417–424 (2017).
26. Wu, Y. et al. Long noncoding RNA hypoxia-inducible factor 1 alpha-antisense RNA 1 promotes tumor necrosis factor-α-induced apoptosis through caspase 3 in Kupffer cells. *Medicine* **97**, e9483 (2018).
27. Wang, S. et al. BRG1 expression is increased in thoracic aortic aneurysms and regulates proliferation and apoptosis of vascular smooth muscle cells through the long non-coding RNA HIF1A-AS1 in vitro. *Eur. J. Cardio-Thorac. Surg.* **47**, 439–446 (2015).
28. Wang, J. et al. Clopidogrel reduces apoptosis and promotes proliferation of human vascular endothelial cells induced by palmitic acid via suppression of the long non-coding RNA HIF1A-AS1 in vitro. *Mol. Cell. Biochem.* **404**, 203–210 (2015).
29. Boyd, A. W., Bartlett, P. F. & Lackmann, M. Therapeutic targeting of EPH receptors and their ligands. *Nat. Rev. Drug Discov.* **13**, 39–62 (2014).

30. Scaria, P. V., Will, S., Levenson, C. & Shafer, R. H. Physicochemical studies of the d(G3T4G3)*d(G3A4G3).d(C3T4C3) triple helix. *J. Biol. Chem*. **270**, 7295–7303 (1995).
31. Jalali, S., Singh, A., Maiti, S. & Scaria, V. Genome-wide computational analysis of potential long noncoding RNA mediated DNA:DNA:RNA triplexes in the human genome. *J. Transl. Med*. **15**, 186 (2017).
32. Maki, T. et al. Angiogenic and vasoprotective effects of adrenomedullin on prevention of cognitive decline after chronic cerebral hypoperfusion in mice. *Stroke* **42**, 1122–1128 (2011).
33. Barquilla, A. & Pasquale, E. B. Eph receptors and ephrins: therapeutic opportunities. *Annu. Rev. Pharmacol. Toxicol*. **55**, 465–487 (2015).
34. Miao, H. et al. EphA2 mediates ligand-dependent inhibition and ligand-independent promotion of cell migration and invasion via a reciprocal regulatory loop with Akt. *Cancer cell* **16**, 9–20 (2009).
35. Tchasovnikarova, I. A. et al. GENE SILENCING. Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells. *Sci. (N. Y., N. Y.)* **348**, 1481–1485 (2015).
36. Cirillo, D. et al. Neurodegenerative diseases: quantitative predictions of protein-RNA interactions. *RNA* **19**, 129–140 (2013).
37. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. **49**, D480–D489 (2021).
38. Zhu, Y., Wang, G. Z., Cingöz, O. & Goff, S. P. NP220 mediates silencing of unintegrated retroviral DNA. *Nature* **564**, 278–282 (2018).
39. Leisegang, M. S. et al. Long noncoding RNA MANTIS facilitates endothelial angiogenic function. *Circulation* **136**, 65–79 (2017).
40. Tunbak, H. et al. The HUSH complex is a gatekeeper of type I interferon through epigenetic regulation of LINE-1s. *Nat. Commun*. **11**, 5387 (2020).
41. Kokura, K., Sun, L., Bedford, M. T. & Fang, J. Methyl-H3K9-binding protein MPP8 mediates E-cadherin gene silencing and promotes tumour cell motility and invasion. *EMBO J*. **29**, 3673–3687 (2010).
42. Chang, Y. et al. MPP8 mediates the interactions between DNA methyltransferase Dnmt3a and H3K9 methyltransferase GLP/G9a. *Nat. Commun*. **2**, 533 (2011).
43. Douse, C. H. et al. TASOR is a pseudo-PARP that directs HUSH complex assembly and epigenetic transposon control. *Nat. Commun*. **11**, 4940 (2020).
44. Müller, I. et al. MPP8 is essential for sustaining self-renewal of ground-state pluripotent stem cells. *Nat. Commun*. **12**, 3034 (2021).
45. Ottoz, D. S. M. & Berchowitz, L. E. The role of disorder in RNA binding affinity and specificity. *Open Biol*. **10**, 200328 (2020).
46. Brantley-Sieders, D. M. et al. EphA2 receptor tyrosine kinase regulates endothelial cell migration and vascular assembly through phosphoinositide 3-kinase-mediated Rac1 GTPase activation. *J. Cell Sci*. **117**, 2037–2049 (2004).
47. Krock, B. L., Skuli, N. & Simon, M. C. Hypoxia-induced angiogenesis: good and evil. *Genes Cancer* **2**, 1117–1133 (2011).
48. Daubon, T. et al. Deciphering the complex role of thrombospondin-1 in glioblastoma development. *Nat. Commun*. **10**, 1146 (2019).
49. Fahmy, R. G., Dass, C. R., Sun, L.-Q., Chesterman, C. N. & Khachigian, L. M. Transcription factor Egr-1 supports FGF-dependent angiogenesis during neovascularization and tumor growth. *Nat. Med*. **9**, 1026–1032 (2003).
50. Pereira, F. A., Qiu, Y., Zhou, G., Tsai, M.-J. & Tsai, S. Y. The orphan nuclear receptor COUP-TFII is required for angiogenesis and heart development. *Genes Dev*. **13**, 1037–1049 (1999).
51. Grote, P. et al. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Developmental cell* **24**, 206–214 (2013).
52. Chu, C., Qu, K., Zhong, F., Artandi, S. E. & Chang, H. Y. Genomic maps of lincRNA occupancy reveal principles of RNA-chromatin interactions. *Mol. cell* **44**, 667–678 (2011).
53. Masri, F. A. et al. Hyperproliferative apoptosis-resistant endothelial cells in idiopathic pulmonary arterial hypertension. *Am. J. Physiol. Lung Cell. Mol. Physiol*. **293**, L548–L554 (2007).
54. Lang, I. M., Pesavento, R., Bonderman, D. & Yuan, J. X.-J. Risk factors and basic mechanisms of chronic thromboembolic pulmonary hypertension: a current understanding. *Eur. Respiratory J*. **41**, 462–468 (2013).
55. Belik, D. et al. Endothelium-derived microparticles from chronically thromboembolic pulmonary hypertensive patients facilitate endothelial angiogenesis. *J. Biomed. Sci*. **23**, 4 (2016).
56. Jain, R. K. et al. Angiogenesis in brain tumours. *Nat. Rev. Neurosci*. **8**, 610–622 (2007).
57. Bell, R. M. & Yellon, D. M. Conditioning the whole heart–not just the cardiomyocyte. *J. Mol. Cell. Cardiol*. **53**, 24–32 (2012).
58. Kamat, M. A. et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinforma. (Oxf., Engl.)* **35**, 4851–4853 (2019).
59. Xing, Y. et al. Long noncoding RNA-maternally expressed gene 3 contributes to hypoxic pulmonary hypertension. *Mol. Ther*. **27**, 2166–2181 (2019).
60. Rininsland, F. et al. Suppression of insulin-like growth factor type I receptor by a triple-helix strategy inhibits IGF-I transcription and tumorigenic potential of rat C6 glioblastoma cells. *Proc. Natl. Acad. Sci. USA* **94**, 5854–5859 (1997).
61. Wang, P., Ren, Z. & Sun, P. Overexpression of the long non-coding RNA MEG3 impairs in vitro glioma cell proliferation. *J. Cell. Biochem*. **113**, 1868–1874 (2012).
62. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinforma. (Oxf., Engl.)* **29**, 15–21 (2013).
63. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. **9**, R137 (2008).
64. Wang, L. et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res*. **41**, e74 (2013).
65. Kang, Y.-J. et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res*. **45**, W12–W16 (2017).
66. Savai, R. et al. Pro-proliferative and inflammatory signaling converge on FoxO1 transcription factor in pulmonary hypertension. *Nat. Med*. **20**, 1289–1300 (2014).
67. Dabral, S. et al. A RASSF1A-HIF1α loop drives Warburg effect in cancer and pulmonary hypertension. *Nat. Commun*. **10**, 2130 (2019).
68. Kearns, N. A. et al. Cas9 effector-mediated regulation of transcription and differentiation in human pluripotent stem cells. *Dev. (Camb., Engl.)* **141**, 219–223 (2014).
69. Konermann, S. et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583–588 (2015).
70. Haeussler, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol*. **17**, 148 (2016).
71. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. methods* **11**, 783–784 (2014).
72. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. methods* **10**, 1213–1218 (2013).
73. Davis, M. P. A., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. & Enright, A. J. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods (San. Diego, Calif.)* **63**, 41–49 (2013).
74. Broad Institute. Picard Toolkit (2019).
75. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol*. **11**, R106 (2010).

76. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
77. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. methods* **9**, 357–359 (2012).
78. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids Res.* **42**, W187–W191 (2014).
79. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinforma.* **14**, 178–192 (2013).
80. Sklenář, V. & Bax, A. Spin-echo water suppression for the generation of pure-phase two-dimensional NMR spectra. *J. Magn. Reson. (1969)* **74**, 469–479 (1987).
81. Hwang, T. L. & Shaka, A. J. Water suppression that works. excitation sculpting using arbitrary wave-forms and pulsed-field gradients. *J. Magn. Reson., Ser. A* **112**, 275–279 (1995).
82. Lee, W., Tonelli, M. & Markley, J. L. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinforma. (Oxf., Engl.)* **31**, 1325–1327 (2015).
83. Brünger, A. T. et al. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. Sect. D., Biol. Crystallogr.* **54**, 905–921 (1998).
84. Linge, J. P., O'Donoghue, S. I. & Nilges, M. Automated assignment of ambiguous nuclear overhauser effects with ARIA. In *Nuclear Magnetic Resonance of Biological Macromolecules - Part B* (Elsevier, 2001), Vol. 339, pp. 71–90.
85. Linge, J. P., Habeck, M., Rieping, W. & Nilges, M. ARIA: automated NOE assignment and NMR structure calculation. *Bioinforma. (Oxf., Engl.)* **19**, 315–316 (2003).
86. Nozinovic, S., Fürtig, B., Jonker, H. R. A., Richter, C. & Schwalbe, H. High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. *Nucleic Acids Res.* **38**, 683–694 (2010).
87. Korff, T. & Augustin, H. G. Integration of endothelial cells in multicellular spheroids prevents apoptosis and induces differentiation. *J. Cell Biol.* **143**, 1341–1352 (1998).
88. Leisegang, M. S. et al. Pleiotropic effects of laminar flow and statins depend on the Krüppel-like factor-induced lncRNA MANTIS. *Eur. heart J.* **40**, 2523–2533 (2019).
89. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic acids Res.* **47**, D442–D450 (2019).
90. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

Acknowledgements

We thank Cindy F. Höper for excellent technical assistance and Jana Meisterknecht for help with mass spectrometry. We are grateful to Katalin Pálfi and Tanja Lüneburg for the help with cell culture. This work was supported by the Goethe University Frankfurt am Main, the German Center for Cardiovascular Research (DZHK, Postdoc Start-up Grant, Förderkennzeichen: 81×3200107, to M.S.L.), the DFG excellence cluster Cardiopulmonary Institute (CPI) EXS2026 (to R.P.B.) and the DFG Transregio TRR267 (Project-ID 403584255—TRR 267; TP A04 (to M.S.L. and R.P.B.), TP A06 (to R.P.B. and H.S.), TP Z02 (to I.W.) and TP Z03 (to M.H.S.)). R.G. was supported by the Rolf M. Schwiete Stiftung (Mannheim, Germany). Work at BMRZ (to H.S.) is supported by the state of Hesse.

Author contributions

M.Le., J.K.B., S.S., S.S.P., F.R., R.G., I.W., T.R., I.G.C., H.S., and R.Br. designed the experiments. M.Le., J.K.B., S.S., J.A.O., N.M.K., C.C.K., S.G., C.C., F.B., J.I.P., S.H., R.Be., C.V., D.C.F., C.R., F.R., B.P.M., I.W., T.R. performed the experiments. M.Le., J.K.B., S.S., J.A.O., N.M.K., N.S.C., H.R.A.J., C.R., B.P.M., I.W., I.G., T.R., H.S. and R.Br. analyzed the data. S.S., J.A.O., C.C.K., S.G., T.W., J.P., M.Lo., M.H.S., R.G., I.G.C. performed bioinformatics. N.S.C., M.H.S., R.G., and I.G. helped with research design and advice. M.Le., J.K.B., H.S., and R.Br. wrote the manuscript. All authors interpreted the data and approved the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-34252-2.

Correspondence and requests for materials should be addressed to Harald Schwalbe or Ralf P. Brandes.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5.2 **Research article:** A universal model of RNA·DNA:DNA triplex formation accurately predicts genome-wide RNA-DNA interactions.

Timothy Warwick, Sandra Seredinski, Nina M. Krause, Jasleen Kaur Bains, Lara Althaus, James A. Oo, Alessandro Bonetti, Anne Dueck, Stefan Engelhardt, Harald Schwalbe, Matthias S. Leisegang, Marcel H. Schulz and Ralf P. Brandes; *Briefings in Bioinformatics* **23**, bbac445 (2022).

In this article, bioinformatics studies were performed to describe RNA:DNA interactions based on triplex sequencing data. The program *TriplexAligner*, a DNA:RNA triplex prediction tool, was implemented with RNA:DNA binding probabilities including DNA:RNA base pairing, Hoogsteen interactions that were consistent with *in vivo* data. The predicted DNA:RNA triplexes were further validated by biochemical and biophysical experiments, including EMSA and CD spectroscopy.

T. Warwick performed the bioinformatics studies, analyzed the data, and wrote the manuscript. S. Seredinski performed the triplex sequencing using RNA and DNA interacting complexes ligated and sequenced (RADICL-seq). The author of this thesis was involved in the CD measurements, which were performed and analyzed by N.M. Krause under the supervision of the author. The corresponding sections of the manuscript were written by the author together with N.M. Krause and H. Schwalbe.

https://doi.org/10.1093/bib/bbac445

Briefings in Bioinformatics, 2022, **23(6)**, 1–12

https://doi.org/10.1093/bib/bbac445

Advance access publication date 13 October 2022

Problem Solving Protocol

A universal model of RNA·DNA:DNA triplex formation accurately predicts genome-wide RNA–DNA interactions

Timothy Warwick, Sandra Seredinski, Nina M. Krause, Jasleen Kaur Bains, Lara Althaus, James A. Oo, Alessandro Bonetti, Anne Dueck, Stefan Engelhardt, Harald Schwalbe, Matthias S. Leisegang, Marcel H. Schulz and Ralf P. Brandes

Corresponding authors. Ralf P. Brandes, Institute for Cardiovascular Physiology, Goethe University, Theodor-Stern-Kai 7, D-60590, Frankfurt am Main, Germany. E-mail: brandes@vrc.uni-frankfurt.de; Marcel H. Schulz, Institute for Cardiovascular Physiology, Goethe University, Theodor-Stern-Kai 7, D-60590, Frankfurt am Main, Germany. E-mail: marcel.schulz@em.uni-frankfurt.de

Abstract

RNA·DNA:DNA triple helix (triplex) formation is a form of RNA–DNA interaction which regulates gene expression but is difficult to study experimentally *in vivo*. This makes accurate computational prediction of such interactions highly important in the field of RNA research. Current predictive methods use canonical Hoogsteen base pairing rules, which whilst biophysically valid, may not reflect the plastic nature of cell biology. Here, we present the first optimization approach to learn a probabilistic model describing RNA–DNA interactions directly from motifs derived from triplex sequencing data. We find that there are several stable interaction codes, including Hoogsteen base pairing and novel RNA–DNA base pairings, which agree with *in vitro* measurements. We implemented these findings in *TriplexAligner*, a program that uses the determined interaction codes to predict triplex binding. *TriplexAligner* predicts RNA–DNA interactions identified in all-to-all sequencing data more accurately than all previously published tools in human and mouse and also predicts previously studied triplex interactions with known regulatory functions. We further validated a novel triplex interaction using biophysical experiments. Our work is an important step towards better understanding of triplex formation and allows genome-wide analyses of RNA–DNA interactions.

Keywords: RNA, DNA, Triplex, machine learning, RNA–DNA interaction.

Introduction

Numerous regulatory roles have been ascribed to RNAs [1, 2], which include interactions with both DNA and proteins. The RNA–protein interface includes functions such as transcription factor addressing and recruitment [3], scaffolding of transcription factor machinery [4] and mediation of histone modifications [5]. Various epigenomic consequences have been attributed to RNA–DNA interactions, including the functional role of the *XIST* transcript in the silencing of the X chromosome during dosage compensation [6]. Other examples of RNA–DNA interactions include

Timothy Warwick is a PhD student at the Institute for Cardiovascular Physiology at Goethe University, Frankfurt. His research pertains to regulatory networks underlying gene expression patterns.
Sandra Seredinski is a PhD student at the Institute for Cardiovascular Physiology at Goethe University, Frankfurt. Her research focuses on lncRNAs and their interactions with chromatin.
Nina M. Krause is a PhD student at the Institute for Organic Chemistry and Chemical Biology at Goethe University, Frankfurt. Her research focuses on the biophysics of RNA–DNA interaction.
Jasleen Kaur Bains is a PhD student at the Institute for Organic Chemistry and Chemical Biology at Goethe University, Frankfurt. Her research focuses on the biophysics of RNA conformation.
Lara Althaus is a PhD student at the Institute of Pharmacology and Toxicology, Technical University of Munich. Her research focuses on control of cellular behaviour by lncRNAs.
James A. Oo is a post-doctoral researcher at the Institute for Cardiovascular Physiology at Goethe University, Frankfurt. His research focuses on lncRNA-mediated control of transcription factors.
Alessandro Bonetti is a post-doctoral researcher at AstraZeneca. His research focuses on reporting of RNA–DNA interactions via next-generation sequencing methods.
Anne Dueck is a group leader at the Institute of Pharmacology and Toxicology, Technical University of Munich. Her research focuses on how lncRNAs define divergent cellular functions.
Stefan Engelhardt is a professor at the Institute of Pharmacology and Toxicology, Technical University of Munich. His research interests cover cardiovascular signal transduction, amongst many other areas.
Harald Schwalbe is a professor at the Institute for Organic Chemistry and Chemical Biology at Goethe University, Frankfurt. His group works on structure and dynamics of biological molecules.
Matthias S. Leisegang is a group leader at the Institute for Cardiovascular Physiology at Goethe University, Frankfurt. His group works on lncRNAs of relevance to the cardiovascular system.
Marcel H. Schulz is a professor at the Institute of Cardiovascular Regeneration at Goethe University, Frankfurt. His group works on computational epigenomics and systems cardiology.
Ralf P. Brandes is a professor at the Institute for Cardiovascular Physiology at Goethe University, Frankfurt. His group works on the physiology and epigenetics of the cardiovascular system.

Received: June 20, 2022. **Revised:** August 16, 2022. **Accepted:** September 17, 2022

RNA–DNA hybrid G-quadruplex [7] and R-loop [8] formation. R-loops consist of interactions between single-stranded DNA and RNA via Watson-Crick base pairing and have been implicated in chromatin condensation and tumorigenesis [9, 10]. Alongside these, there exists another form of RNA–DNA interaction where DNA structure is maintained and single-stranded RNA binds in the major groove of the double helix, resulting in the formation of an RNA·DNA:DNA triple helix (triplex) [11].

Triplex formation represents an area of epigenetics which, although known of in a biophysical sense for many years [12], remains incompletely understood. There are several reasons for this, chief amongst them being the experimental complexities of studying triplex formation on a genome-wide scale in living cells. Experimental probing of triplex formation in the cellular context has previously relied upon methods capturing the genomic interaction sites of single transcripts [13, 14]. However, regulatory transcripts have been implicated in epigenetic mechanisms across many species, tissues and cell types [15–18]. Herein lies the importance of developing tools which accurately predict points of RNA–DNA interaction as putative sites of triplex formation.

Previously published computational tools have relied upon Hoogsteen base pairing rules [19], which are canonically responsible for triplex formation. Tools implementing Hoogsteen rules include *Triplexator/Triplex Domain Finder* [20, 21] and *LongTarget* [22]. Whilst usage of canonical rules to predict triplex formation provides insight into putative RNA–DNA interactions, benchmarking of *Triplexator* and *LongTarget* using *MEG3* ChOP-seq data [13] revealed substantial room for improvement in this area [23]. These benchmarking data suggest that accurate prediction of triplex formation in a cellular context may require the implementation of as yet unknown base pairing rules which go beyond the currently used Hoogsteen base pairing rules. Related to this, there has been work on prediction of triplex-forming RNA and DNA sequences using previously published triplex interactions [24], although prediction of complete triplex interactions is not possible with this method.

Accompanying computational methods for genome-wide prediction of RNA–DNA interaction have been experimental methods with similar aims. Foremost amongst these, with regard to triplex formation, is genome-wide isolation of triplexes followed by sequencing (triplex-seq) [25]. This method permits the identification of triplex-forming sequences across the genome (triplexDNA-seq) and transcriptome (triplexRNA-seq) but lacks information on the pairing of the sequences with one another. Outside specifically triplex-mediated RNA–DNA interactions, there have been a number of published methods designed to identify all-to-all interactions between transcripts and chromatin [26–29]. However, the methods with most similar nucleotide processing protocols to triplex-seq are RNA And DNA Interacting Complexes Ligated and sequenced (RADICL-seq) [30] and RedC [31]. These methods collectively identify specific interactions between transcripts and regions of the genome through ligation of RNA and DNA via a linker sequence in a proximity-based manner. RADICL-seq and RedC provide rich sources of data on RNA–DNA interaction but also remain relatively novel and experimentally complex. The undertaking of such experiments across a range of steady-state and differential conditions is therefore not feasible at this juncture. Consequently, the most widely applicable use for these data may be as input to machine learning algorithms, which could permit the prediction of RNA–DNA interactions in a condition of interest.

Here, we present a method for the prediction of RNA–DNA interactions based on RNA–DNA binding probabilities learned by expectation-maximization from triplex-forming sequences identified in triplexDNA-seq and triplexRNA-seq. Applying these binding rule sets as substitution matrices in local alignment permitted more accurate recall of RNA–DNA interactions identified from RADICL-seq and RedC when compared with previously published tools. Experimentally validated triplex formation between transcripts and genomic loci could also be recapitulated. A predicted interaction was also subjected to biophysical validation, where triplex formation could be experimentally verified *ex vivo*.

Materials and methods

Triplex sequencing data processing

Publicly accessible triplexDNA-seq data and triplexRNA-seq data [25] (NCBI Short Read Archive accessions SRR7965691, SRR7965692, SRR7965693, SRR7965694, SRR7965701, SRR7965702, SRR7965703) were downloaded and aligned against the *hg38* genome and transcriptome, respectively, using *Bowtie2* (v2.4.4) with default parameters [32]. Peaks were called from alignments using *HOMER findPeaks* (v4.11.1) with default parameters [33]. Sequences underlying identified peaks were then extracted from the *hg38* genome or transcriptome using *bedtools getfasta* (v2.27.1) with the peak coordinates per sample as input [34].

Motif enrichment and processing

Peak sequences were used as input to motif enrichment using *MEME-ChIP* (v5.0.5) [35]. Shuffled peak sequences were supplied as negative sequence input, with the maximum number of enriched *MEME* motifs restricted to 16, with a maximum motif length of 32 nucleotides. Significantly enriched motifs were considered to be those with an $E < 0.05$. Enriched motifs were subjected to motif comparison between samples using *Tomtom* (also part of *MEME Suite*), and matches were considered to be present when $P < 0.05$. *FIMO*, another tool contained within *MEME Suite*, was used to compute the occurrences of enriched triplex motifs across the breadths of triplex DNA and RNA peaks. Triplex motif occurrence was also computed separately for peaks lying in distinct genomic features as defined by the *TxDb.Hsapiens.UCSC.hg38.knownGene* annotation package for *R*, maintained by *Bioconductor* [36–38]. Triplex RNA motif occurrence was computed per transcript biotype, as defined by *EnsDb.Hsapiens.v86* and normalized to transcript length [39, 40]. Enriched triplex DNA and RNA motifs were compared against the *JASPAR 2020* [41] and *ATtRACT* [42] motif databases, respectively, in order to remove any motifs which were identical to transcription factor-binding or RNA-binding protein motifs.

An Expectation-Maximization-based method for RNA–DNA code optimization

Nomenclature

Assume, we are given a DNA motif D of length l, which is a matrix D^{4xl}, over $\Sigma = \{A, C, G, T\}$ denoting the set of characters in the DNA alphabet. Also, we have RNA motif matrix R^{4xl}, for simplicity we assume that the RNA alphabet has been translated to the DNA alphabet by exchanging U→T. For simplicity, all motifs considered here have the same length l. **Note** that a difference in length between an RNA and DNA matrix can easily be accounted for by testing all possible shifts of the smaller matrix against the larger matrix.

We assume that there is a set of DNA motifs $\mathcal{D} = \{D_1, \ldots, D_n\}$ and equivalently a set of RNA motifs $\mathcal{R} = \{R_1, \ldots, R_n\}$. For notational simplicity, we assume that they have the same number of elements n, although in practice this may change.

We assume that there exists a mapping code C^{4x4}, which is a matrix that maps nucleotides from RNA to DNA nucleotides. For example, $C_{A,A}$ denotes the probability to map an *A* RNA nucleotide to an A DNA nucleotide. Entries in the row of the matrix sum to one, and thus, it holds that

$$\forall r \in \Sigma, \sum_{d \in \Sigma} C_{r,d} = 1 \,. \tag{1}$$

The interest in this formulation is to learn the code *C* that is behind a given set of motifs $\mathcal{R}$ and $\mathcal{D}$.

Code objective value

We define the average column-wise mapping error between a DNA motif *D* and an RNA motif *R*–termed code objective value–given a defined code matrix C as

$$objective(R, D, C) = \frac{\sum_{i \in \Sigma, j \in 1,\dots,l} abs(D_{i,j} - \hat{D}_{i,j})}{l} \,, \tag{2}$$

where $\hat{D}$ is the projected DNA motif after conversion of *R* using C

$$\hat{D}_{i,j}, = \sum_{a \in \Sigma} R_{a,j} \cdot C_{a,i} \,, \tag{3}$$

where $i \in \Sigma$, $j \in 1, \dots, l$.

Obtaining the best code using quadratic programming

Given the two sets of input motifs from DNA $\mathcal{D}$ and RNA $\mathcal{R}$ motifs, we are looking for an optimal code C that describes the conversion of an RNA motif to a DNA motif, as would be done when a subsequence of an RNA is aligned to a subsequence in a DNA sequence, a triplex match.

Assume that we had a known pairing *P* of the RNA to DNA motifs, then we would be looking for the code matrix *C* that minimizes the code objective value (Eq. (2)) for the pairing of RNA and DNA matrices

$$\underset{C}{\textbf{argmin}} = objective(\mathcal{D}, \mathcal{R}, \mathcal{P}, C), C \in \mathcal{C} \,, \tag{4}$$

where *objective*$(\mathcal{D}, \mathcal{R}, \mathcal{P}, \mathcal{C})$ denotes the sum of code objective values for all defined pairs using Eq. (2) and $\mathcal{C}$ denotes the space of all possible code matrices. The term *error* in Eq. (4) may be interpreted as the error of a learned code given the pairings of motifs it describes. Luckily, we can obtain the code matrix C that minimizes the code objective value using quadratic programming efficiently.

An Expectation-Maximization algorithm for finding optimal code sets

While it is straightforward to obtain a code matrix C that minimizes the code objective value for a given pairing of RNA and DNA motifs, in practice, the true pairing is not known. Furthermore, it is unknown whether the triplex binding of all RNA–DNA pairs follows the same code and the possibility of several code matrices needs to be considered.

Therefore, we have designed an Expectation-Maximization algorithm for finding a set of code matrices starting from a given set of RNA and DNA motifs for which the correct pairing is unknown.

Conceptually, the algorithm performs the three following steps to find *k* many code matrices

```
Input: $\mathcal{D}, \mathcal{R}, k$
generate $k$ random code matrices $C^*_1, \dots, C^*_k$
1. $C_1 = C^*_1, \dots, C_k = C^*_k$
2. obtain the best pairings for elements in $\mathcal{D}$ and $\mathcal{R}$ using one of $C_1, \dots, C_k$
3. for each: i=1,...,k
    $C^*_i$ = minimize code of all paired DNA and RNA motifs that used $C_i$
repeat at 1. if $(C_1 \neq C^*_1, \dots, C_k \neq C^*_k)$
Output: final code matrices $C^*_1, \dots, C^*_k$, pairing $P$
```

The second step listed above–where the best motif pairings are obtained–refers to obtaining the best pair for each RNA motif in $\mathcal{R}$ with a DNA motif in $\mathcal{D}$ testing all *k* code matrices. The best pair are the indices i, j, h where $error(\mathcal{R}_i, \mathcal{D}_j, \mathcal{C}_h)$ is minimal. In this process, it is allowed that several elements from $\mathcal{R}$ are paired with the same element in $\mathcal{D}$ and *vice versa*.

In summary, the above EM procedure determines the best pairing between RNA and DNA motifs and determines *k* code matrices as a result of the process. Results are output upon convergence of the algorithm, when all code objective values have been minimized and the code matrices are unchanged. Applications using both simulated and real motif pairs in the course of this work have shown that the algorithm converges in a small number of iterations in practice. However, the solution only constitutes a local minimum; therefore, we run the algorithm many times with the same input data, e.g. 11 270 times with the real triplex input motifs.

The code for the steps described above is publicly available at https://github.com/SchulzLab/Codefinder.

Code processing and annotation

Learned code models, which were output from the Expectation-Maximization algorithm, were stratified by their objective values and the total number of motifs incorporated into the model. These metrics were also used to subset the models and identify the most promising candidates for further study (*objective* < 0.75, *total motifs* > 50%). *In vitro* RNA·DNA:DNA triple helix base triplet stability data [43] were used to compare the affinities of learned code models versus a size-matched set of random code models. In short, the normalized dissociation constant of the RNA:DNA interaction as reported in [43] was multiplied by the probabilities of nucleotide interaction contained within the code matrix, and then summed. The relative affinity values for each code model with *total motifs* > 10% were also linearly regressed against the objective values returned from Expectation-Maximization. Following this, high-scoring code models were subjected to hierarchical clustering with Euclidean distances and Ward's method in order to control for redundancy [44]. The resulting dendrogram was then cut to produce eight clusters, and the mean code model for each cluster was computed.

Formulation of TriplexAligner

To be implemented in *TriplexAligner*, probabilistic code model values were converted to log odds scores according to [45]. Subsequent score distributions were computed for each code model with *Biostrings::pairwiseAlignment* [46], using simulated DNA and RNA sequences with matching nucleotide proportions relative to human promoter sequences and known triplex-forming transcript sequences, respectively. Arising scores were fitted using a generalized extreme value distribution [47] with *EnvStats::egevd* [48], using maximum likelihood estimation. The parameter values *K* and λ could then be identified for each code model. Using these parameters, bit scores (S') and corresponding E values could be calculated from local alignment scores (S) of respective code models according to the following formulas:

$$S' = \frac{\lambda S - ln(K)}{ln2}\,, \tag{5}$$

$$E = mn\,2^{-S'}\,. \tag{6}$$

TriplexAligner computes local alignment scores, bit scores and E values between supplied DNA and RNA sequences for each code model using *Biostrings::pairwiseAlignment* and the above formulae, with the log odds code model supplied as the *substitutionMatrix* parameter.

The code for *TriplexAligner* is publicly available at https://github.com/SchulzLab/TriplexAligner, where it is formalized as an *R* package which may be downloaded, installed and used by interested parties.

Computational validation of TriplexAligner using global RNA–DNA interactions

RNA–DNA interactions arising from either RedC (GSE136141) or RADICL-seq (GSE132192) were used to benchmark the performance of *TriplexAligner* compared with the previously published tools *LongTarget* and *Triplexator*. For RedC data, interactions between RNAs and 5 kb genomic bins were used for validation if they were present in two separate replicates. For the RADICL-seq interactions, significant interactions between RNAs and 5 kb genomic bins were identified according to [30]. Interactions between RNAs and genomic bins were expanded to include all possible transcripts of the involved RNA gene, as annotated in *TxDb.Hsapiens.UCSC.hg38.knownGene* [49] or *TxDb.Mmusculus.UCSC.mm10.knownGene* [50] annotation packages for *R* [38]. RADICL-seq interactions were limited to those involving transcripts expressed in accompanying nuclear RNA-seq data, quantified using *Salmon* (v 1.6.0) [51]. For each interaction, involved RNA and DNA sequences were subjected to RNA·DNA:DNA triple helix prediction using *TriplexAligner*, *LongTarget* and *Triplexator*. A corresponding negative dataset was constructed via shuffling of the transcript sequences. *LongTarget* was run with default parameters, and *Triplexator* with *-e* 20 *-l* 5. Maximum metrics (*TriplexAligner*: $-\log_{10}(E)$; *LongTarget*: MeanStability; *Triplexator*: Score) were identified per gene-bin interaction and used as predictive values in subsequent analyses with *pROC* and *ROCR* [52, 53]. Receiver operating characteristic curves were computed for each method with binomial smoothing and statistically compared by bootstrapping ($n = 2000$).

Electrophoretic mobility shift assay

All hybridization steps were performed in 25 mM HEPES, pH 7.4, 50 mM NaCl and 10 mM MgCl2. DNA oligos, spanning the predicted triplex DNA sequence (20 pmol), were hybridized to DNA duplex in a thermocycler by heating up for 5 min to 95°C followed by a cool down to 24°C with a rate of 1°C/ 30 s. For triplex formation, 10 eq of ssRNA (200 pmol), containing the predicted triplex RNA sequence, was added to the DNA duplex followed by incubation at 60°C for 1 h and a cool down to 24°C with a rate of 1°C/ 30 s. RNase H digestion was performed by adding RNase H to a final concentration of 375 mU/μL to a triplex sample and incubate it for 30 min at 37°C. RNase A was added to a triplex sample with a final concentration of 5 ng/μL and incubated similarly to RNase H. Samples were applied on a native 15% Polyacrylamide gel in a running buffer containing 40 mM Tris-Ac pH 8.3 supplemented with 3 mM magnesium acetate. The gels ran for 6 h at room temperature with 160 V.

CD spectroscopy and melting curve analysis

Circular dichroism spectra were acquired on a Jasco J-810 spectropolarimeter. The measurements were recorded from 210 to 320 nm at 25°C using 1 cm path length quartz cuvette. CD spectra were recorded on 8 μM samples of each DNA duplex, DNA:RNA heteroduplex and DNA:DNA:RNA-triplex (10 equivalents of RNA (80 μM)) in 25 mM HEPES, 50 mM NaCl and 10 mM $MgCl_2$ (pH 7.4). Spectra were acquired with 8 scans and the data were smoothed with Savitzky–Golay filters. Observed ellipticities recorded in millidegree (mdeg) were converted to molar ellipticity $[\Theta] = deg \times cm^2 \times dmol^{-1}$. Melting curves were acquired at constant wavelength using a temperature rate of 1°C/min in a range from 5 to 95°C. All data were evaluated using SigmaPlot 12.5. All melting temperature data were converted to normalised ellipticity and evaluated by the following equation: $f = a/(1 + exp(-(x - x0)/b)) + c/(1 + exp(-(x - x2)/d))$.

Results

Prediction of RNA·DNA:DNA triplex interactions from captured triplex sequences

In order to predict interactions between DNA and RNA mediated by triplex formation (Figure 1A), we developed *TriplexAligner*. *TriplexAligner* is capable of predicting RNA–DNA interactions with high accuracy, surpassing currently available methods. Development of *TriplexAligner* (Figure 1B, Supplementary Figure 1) encompassed multiple stages and included multiple next-generation sequencing datasets, alongside machine learning and biophysical methods. Initially, key sequence elements of triplex-forming RNA and DNA sequences needed to be identified. For this purpose, triplexRNA-seq and triplexDNA-seq data from HeLa cells [25] were analysed and triplex-forming regions were identified by peak calling. RNA and DNA components of the published triplex interaction between the *MEG3* transcript and the gene locus of *COL15A1* [13] could be clearly identified in the dataset (Figure 1C). This satisfied the requirement that the input data sufficiently capture triplex formation taking place in the cellular context.

Identification of short sequences underpinning triplex formation

The next step in development of *TriplexAligner* was the identification of triplex-enriched DNA and RNA regions, along with associated sequences. Peak calling on triplexDNA-seq data identified regions most often in intronic areas of the genome (Figure 2A), although promoter regions were most enriched for peak occurrence relative to the proportion of the genome covered **(Supplementary Figure 2A)**. When calling peaks from triplexRNA-seq data, most peaks were detected in protein-coding transcripts (Figure 2B). Transcripts with retained introns, followed by antisense transcripts, were most enriched for triplexRNA-peaks relative to transcript length **(Supplementary Figure 2B)**.

To identify sequences underpinning triplex-seq peaks, motif enrichment analysis was performed using *MEME-ChIP* [35] on sequences underlying peaks in each sample. Between 22 and 36 significantly enriched motifs were identified per triplexDNA-seq sample (Figure 2C, left). More motifs were identified in the RNA samples, which each returned more than 125 enriched motifs (Figure 2C, right). To investigate the reproducibility of the enriched

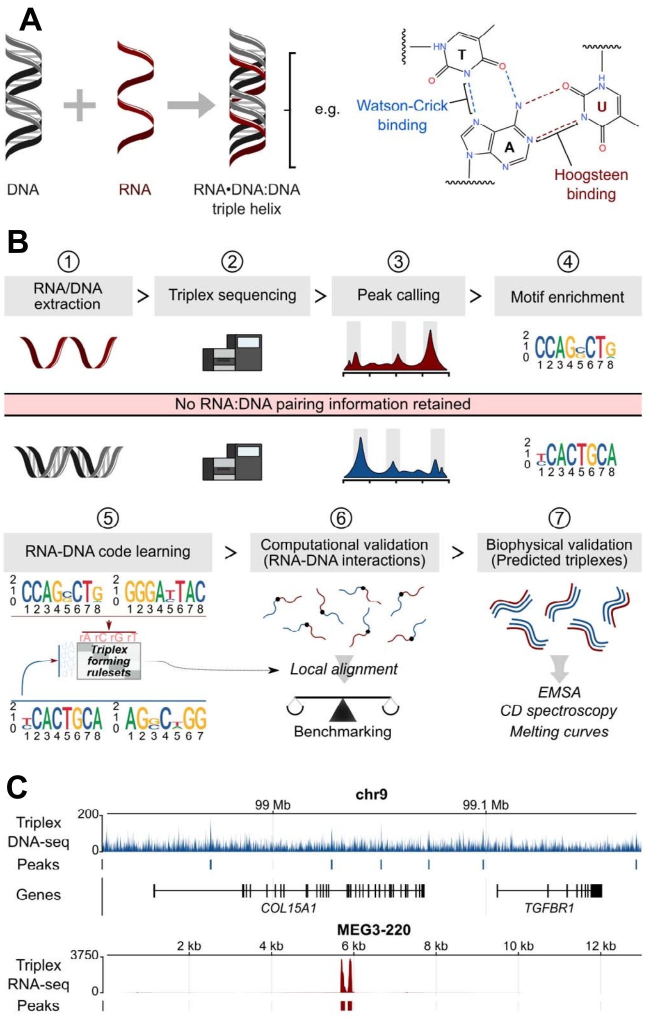


Figure 1. Overview of RNA·DNA:DNA triple helix formation and the development of TriplexAligner from triplex-seq data. **(A)** Schematic of RNA·DNA:DNA triple helix formation and effects on gene expression. **(B)** Overview of the development of TriplexAligner. **(C)** Peak calling on triplexDNA-seq (blue) and triplexRNA-seq data (red). The displayed regions reflect the published RNA·DNA:DNA triple helix interaction between *MEG3* and a DNA site in the locus of *COL15A1*, which results in the regulation of the downstream gene *TGFBR1*.

sequences, motifs were compared between samples using *Tomtom* [35]. Both RNA and DNA samples displayed high degrees of reproducibility, with a minimum of 76% of DNA motifs per sample having similar motifs in another DNA sample (Figure 2D, top). RNA motifs were also reproducible, with the minimum percentage of similar motifs between any two samples being 66% (Figure 2D, bottom). Individual RNA motifs also exhibited more significant enrichment than DNA motifs, observable when examining the five most enriched motifs of each molecule (Figure 2E). The most enriched RNA motif had an *E* value of 8.6×10^{-1382}, in comparison to an *E* value of 4.2×10^{-95} for the most enriched DNA motif.

In order to establish whether enriched triplex motifs reflect putative regulatory functions of triplex formation, motif occurrence in different features and transcripts was compared. Enriched triplex DNA motifs occurred at higher rates in triplex DNA peaks residing in promoters, intergenic regions and introns relative to exonic regions (Figure 2F). Triplex RNA motifs appeared more often in transcripts lacking open-reading frames–specifically lincRNAs, antisense transcripts and transcripts with retained introns–relative to protein coding transcripts (Figure 2G). Taken together, these findings indicate that triplex formation between non-coding RNA and non-coding regions of the genome is best described by reproducible sequence elements. In a positional sense, triplex motifs did not show any specific pattern of occurrence within peak regions **(Supplementary Figure 2C)**.

To select triplex motifs to take forward to the next stage of development, the enriched RNA and DNA motifs were subjected to several stratification steps. Identical motifs between samples were removed from the analysis, and motifs previously implicated in either transcription factor or RNA-binding protein interactions were excluded in an attempt to isolate sequences of most importance for triplex formation. Complementary DNA motifs were also added (Figure 2H), owing to the non-stranded nature of the analysis. The outcome of these steps were two sets of triplex motifs, consisting of 192 DNA motifs and 324 RNA motifs, which were used in the development of *TriplexAligner*.

Expectation-maximization to learn triplex formation rules

To learn the putative nucleotide pairing rules which might govern triplex formation, an Expectation-Maximization (EM) algorithm was used to compute triplex nucleotide pairing probabilities from triplex motifs (Figure 3A). The expectation portion of the algorithm was formed by pairings of RNA and DNA triplex motifs. From these pairs, probabilistic models were computed by quadratic programming, averaged across the pairings and evaluated for their error per motif pair (here termed the code objective value). When the objective value was minimized, motif pairings and probabilistic triplex mapping codes were returned. Initially, simulated motif sets were used to test the algorithm. The algorithm was capable of accurately learning nucleotide pairing probabilities, with an example shown in Figure 3B of the learning of Watson–Crick base pairing rules from 100 simulated motif pairs.

The algorithm was subsequently implemented with enriched triplex motifs as input, across 11 270 separate random initiations (Figure 3C). Results were then probed for several metrics, including the final objective values and proportion of total motifs included in the final motif pairings. Results with objective values less than 0.75 and containing more than 50% of total triplex motifs represented the most promising results and were subset, resulting in 801 putative codes. These subset conditions reflect the intention to find a balance between incorporating as much of the input data as possible, along with discarding low-scoring codes which may reflect aberrant starting points of the algorithm (**Supplementary Figure 3**). High-scoring codes resulting from these steps were annotated with published *in vitro* triplex nucleotide dissociation equilibrium constants [43]. The relative binding strengths were calculated for each code and its reverse complement, with the maximum then being taken as the value for that code. The high-scoring mapping codes presented significantly ($W = 415474$, $P < 2.2 \times 10^{-16}$, Mann–Whitney U test) greater relative binding strengths than an identically sized set of randomly generated codes (Figure 3D). When regressing code objective values from all results containing more than 10% of all motifs versus relative *in vitro* binding strengths, a negative correlation could be seen ($R^2 = 0.394$, $P < 2.2 \times 10^{-16}$, Figure 3E). This correlation suggested that the code objective values reflect experimental data, making them appropriate to stratify the codes by. Given that the algorithm used here minimizes the code objective value, a negative correlation between this value and *in vitro* stability directly supports the approach. Collectively, these data suggest that the probabilistic

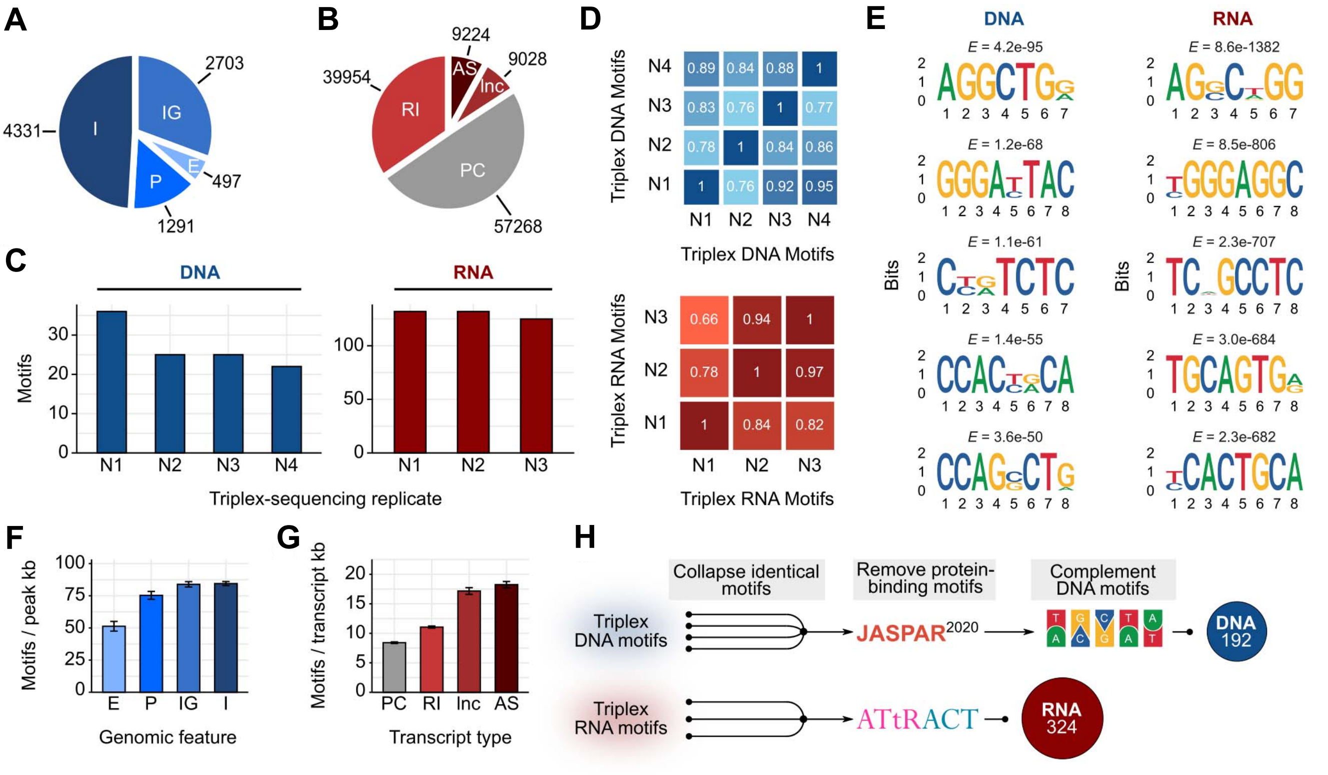


Figure 2. Identification of enriched and reproducible RNA·DNA:DNA triple helix-forming motifs. **(A)** Distribution of triplexDNA-seq peaks across intronic regions (I), intergenic regions (IG), exonic regions (E) and promoter regions (p) as annotated in the hg38 genome build by NCBI. **(B)** Distribution of triplexRNA-seq peaks across antisense transcripts (AS), long non-coding RNAs (lnc), protein-coding transcripts (PC) and transcripts with retained introns (RI). **(C)** Total significantly enriched ($E < 0.01$) triplexDNA and triplexRNA motifs identified per replicate of triplexDNA-seq and triplexRNA-seq. **(D)** Proportions of motifs per replicate with similar ($P < 0.05$, *Tomtom*) motifs in accompanying replicates of triplexDNA-seq (blue) or triplexRNA-seq (red). **(E)** The five most enriched motifs across all replicates of triplexDNA-seq (left) and triplexRNA-seq (right). **(F)** Occurrence of triplexDNA motifs per kilobase of triplexDNA-seq peaks appearing in exonic (E), promoter (P), intergenic (IG) and intronic (I) genomic regions. **(G)** Occurence of triplexRNA motifs per kilobase of protein-coding (PC), retained intron (RI), long non-coding (lnc) and antisense (AS) transcripts. **(H)** Schematic of motif processing steps, including removal of identical motifs, removal of known protein-binding motifs and inclusion of reverse-complement triplexDNA motifs, which resulted in the final sets of triplexRNA (red) and triplexDNA (blue) motifs.

codes learned using an expectation-maximization algorithm from triplex motifs have biophysical relevance in triplex formation.

In order to stratify the codes to be taken forward and implemented in *TriplexAligner*, high-scoring results (*objective value* < 0.75, *total motifs utilised* > 50%) output from the expectation-maximization algorithm were hierarchically clustered and subjected to tree cutting, resulting in eight distinct code clusters (Figure 3F–G). Amongst these were partially redundant pairs of codes (clusters 1 and 2, clusters 3 and 4), resulting from the probabilistic nature of the outputs. Complementary codes (clusters 5 and 7) were also present, reflecting the non-stranded nature of the input sequencing data. Amongst these codes were canonical Hoogsteen base pairing rules (C·G:C, U·A:T, G·G:C, A·A:T) [21] along with previously unreported RNA·DNA:DNA base pairings. The learning of potentially novel base pairings is a key point of the naive, unbiased approach taken herein. To determine the potential worth of these base pairings, an assessment of the returned codes in the prediction of published RNA–DNA interactions was carried out.

In order to perform the validation of the stratified codes returned by the expectation-maximization algorithm, the codes were implemented as scoring matrices in a new software called *TriplexAligner*. *TriplexAligner* is a local alignment program which uses Karlin–Altschul statistics [54] to determine subsequences of triplex formation between RNA and DNA.

TriplexAligner recalls RNA–DNA interactions and known triplexes

Genome-wide validation of *TriplexAligner* was carried out using published RNA–DNA interactions as detected by RADICL-seq [30] and RedC [31]. Significant interactions between transcripts and 5 kb genomic bins were decomposed to RNA and DNA sequences and constituted the positive data set. RADICL-seq interactions were further refined using accompanying nuclear RNA-seq data [30]. Transcript sequences were shuffled in order to generate a negative interaction set whilst maintaining nucleotide frequencies. Triplex formation between sequences was then predicted using *TriplexAligner*, *Triplexator* [20] and *Long Target* [22]. Maximum scores for each method were computed per RNA–DNA interaction (Figure 4A) and used as predictive values. All three tools were able to positively classify RADICL-seq and RedC interactions, with *TriplexAligner* returning the greatest area under the receiver operating characteristic (ROC) curve for both assays (Figure 4B and C, Supplementary Figure 4A). In both cases, the area under the ROC curve was significantly greater for *TriplexAligner* compared with the other tools ($P < 0.05$, bootstrapping $n = 2000$) (Figure 4D).

Upon assessing the performance of each of the individual mapping codes, which constitute *TriplexAligner*, it became clear that code performance varied between both individual codes and assays. Notably, Code 5 displayed the lowest performance on both RADICL-seq and RedC data and was only marginally better

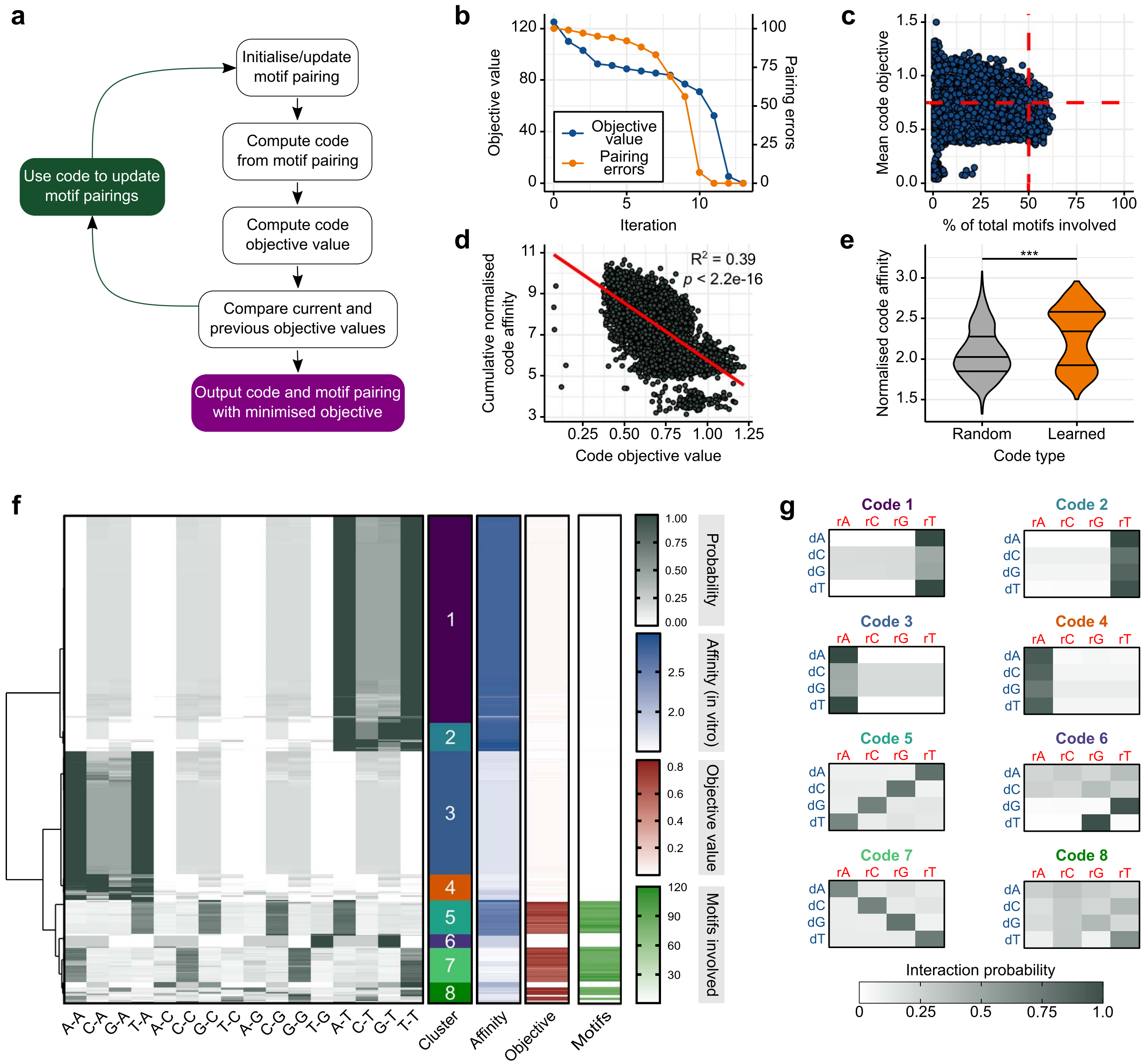


Figure 3. Learning RNA·DNA:DNA triple helix nucleotide pairing rules from motifs using expectation-maximization. **(A)** Schematic of the expectation-maximization algorithm used to learn RNA·DNA:DNA triple helix base pairing probabilities from pairings of enriched triplexRNA and triplexDNA motifs. **(B)** Example use-case of the expectation-maximization algorithm on simulated sets of motifs ($n = 100$) which were paired by Watson–Crick base pairing rules, with corresponding objective values and number of incorrect motif pairs displayed per iteration of the algorithm. **(C)** Output from the expectation-maximization algorithm when run on enriched triplexDNA and triplexRNA motifs identified from triplex-seq, displaying the mean objective values across all code models learned per initiation of the algorithm and the corresponding proportion of motifs included. **(D)** Correlation between code model objective values and *in vitro* RNA·DNA:DNA binding affinities as reported in [43]. Objective values and affinities were subjected to linear regression, with corresponding coefficient of determination (R^2) and P-value displayed on the plot. **(E)** Comparison in code model affinities between high-scoring subset (*objective value* < 0.75, *total motifs* > 50%) expectation-maximization results and a size-matched set of randomly generated code models ($P < 0.001$, Wilcoxon signed-rank test). **(F)** High-scoring expectation-maximization results subjected to hierarchical clustering and tree-cutting ($k = 8$), with corresponding clusters, code model affinities, objective values and total motifs assigned displayed. **(G)** Mean probabilistic code models per cluster of expectation-maximization results.

than random code performance (Figure 4E, Supplementary Figure 4B). Different transcripts also showed different preferences for codes. For instance, the two most prevalent lncRNAs in the RedC dataset–*MALAT1* and *NEAT1*–showed completely different code preferences. RedC interactions involving *MALAT1* were most often best-predicted by code 3, compared with interactions of *NEAT1* which were more heterogeneously predicted, with a tendency towards code 7 (Supplementary Figure 4C).

Where the above results showcase the ability of *TriplexAligner* to recall genome-wide RNA–DNA interactions, we also sought to demonstrate that *TriplexAligner* could predict previously published triplex interactions. Triplex interactions between the lncRNA *SARRAH* and a number of cardiac gene promoters (*ITPR2*, *PARP8*, *PDE3A*, *SSBP2* and *GPC6*) have been reported [55]. When using *TriplexAligner* to predict the triplex formation between *SARRAH* and these genomic regions, triplex formation at the cardiac promoters was predicted with greater $-\log_{10}(E)$ values compared with the control promoter used in the experiment (*GAPDH*) (Figure 4G). Of the promoters considered, the best alignment returned by *TriplexAligner* was between *SARRAH* and *ITPR2* **(Supplementary Figure 4D)**. Other published triplex interactions, between *HOTAIR* and the promoter region of *PCDH7* [56] (Figure 4H, left), as well as the triplex formed between *NEAT1* and the *CYP4F22* promoter [25] (Figure 4H, right) returned $-\log_{10}(E)$ values of 5.84 and 18.0, respectively.

If implemented in a genome-wide manner, *TriplexAligner* could also be used to identify regulatory regions of RNAs which are important to triplex formation. Here, triplex formation between an exemplary RNA (*Neat1*) and promoter sequences of genes differentially expressed after *Neat1* knockout [57] was predicted using *TriplexAligner*. It was evident that a specific region of the *Neat1* transcript was implicated in predicted triplex formation (Figure 4I), indicating a region of putative regulatory importance in the transcript.

TriplexAligner code models are biophysically valid

To assess the biophysical validity of the code models used in *TriplexAligner*, maximal RADICL-seq pair alignments of each code were computed. The alignment with the greatest $-\log_{10}(E)$ value was that implementing code 7. RNA and DNA oligonucleotides representing maximally scoring subsequences of this interaction (Figure 5A) were submitted for analysis by electrophoretic mobility shift assays (EMSA), circular dichroism (CD) spectroscopy and melting curve analysis.

When the double-stranded DNA was incubated with single-stranded RNA and subjected to an EMSA, an RNaseH-resistant band could be observed in the gel separate from the double-stranded DNA alone (Figure 5B). RNaseH resistance indicates that the formed structure was not an R-loop [58] and could therefore be an RNA·DNA:DNA triple helix. When subjected to CD spectroscopy, a distinct negative peak at 230 nm was present when RNA and DNA were mixed, along with a shifted and prominent main peak at 270 nm (Figure 5C). These shifts were not visible when double-stranded DNA alone or mixed single-stranded DNA and single-stranded RNA (heteroduplex) were tested. In melting assays, two melting points could be assigned to the curve obtained from the mixed double-stranded DNA and single-stranded RNA (Figure 5D). In contrast, only single melting points could be assigned to double-stranded DNA alone and heteroduplex inputs.

These results indicate that RNA–DNA interactions positively predicted by *TriplexAligner* have the potential to be biophysically valid triplexes, even when only a small portion of the predicted triplex is tested.

Discussion

Unlike previously published tools for the prediction of triplex formation, *TriplexAligner* uses probabilistic nucleotide pairing models learned from sequencing of triplex-forming DNA and RNA to predict triplexes. Compared with discrete and canonical Hoogsteen base pairing rules, this resulted in the improved recall of all-to-all RNA-DNA interactions. This demonstrates that formation of RNA–DNA interactions is more complex than simple base pairing rules, and therefore, prediction of such interactions requires more malleable models such as those proposed here.

The nature of the input data originating from triplexRNA-seq and triplexDNA-seq [25], and the fact that we wanted to integrate triplex codes into local alignment methods, has led us to our formulation using RNA-to-DNA matrices. We showed that optimization of the code matrices and the corresponding motif pairs can be formulated as a code mixture problem using an expectation-maximization algorithm [59]. The use of this algorithm permitted the estimation of triplex motif pairings, and subsequently RNA–DNA base pairing probabilities whose error (objective) could be minimized. In practice, the algorithm converged in few iterations. The drawbacks of expectation-maximization [60] either had a minimal impact (quick convergence) or could be effectively overcome (many initiations to cover the search space). Thus, the algorithm provided useful information for implementation in *TriplexAligner*. Notably, if more complex formulations of RNA-to-DNA interaction would be considered, the optimization would likely become more challenging and a straightforward integration into existing local alignment approaches may become impossible. Here, we were able to use the established Karlin–Altschul statistic to directly generate *E*-values for the triplex alignments, which is a novel contribution compared with established tools. Nevertheless, more complex formulations of RNA-to-DNA interaction would require alternative machine learning approaches, such as neural network-based approaches.

The naive, unbiased manner in which the expectation-maximization algorithm used herein was formulated meant that the returned code models are a heterogeneous mix of known, canonical Hoogsteen base pairings [21] along with previously unreported RNA–DNA base pairings. Due to the probabilistic nature of the code models, it was initially challenging to determine whether these putative novel base pairings were genuinely interesting, or artefacts. In the validation carried out in the course of this work, it could be shown that these unconventional pairings are also good predictors of RNA–DNA interaction. The question remains, however, whether these base pairings are involved in true biophysical interaction between RNA and DNA molecules, or whether their roles are more complex. For example, seeing as the roles of proteins in triplex formation remain unknown, it could be that these non-canonical base pairings are important for the recruitment of co-factors important for stabilization of the triplex structure.

Interestingly, it could be demonstrated that different transcripts may have different preferences for the code with which their interactions were predicted. The examples given here–*MALAT1* and *NEAT1*–are both lncRNAs which have been previously shown to interact with chromatin [61]. However, when their RNA–DNA interactions are predicted by *TriplexAligner*, they are best-predicted with different codes. These two lncRNAs have been proposed to carry out functions at similar genomic loci–namely actively transcribed genes–but to bind at distinct sites in these regions. Therefore, interactions underpinned by different codes would facilitate this process. Inferring functional roles of different types of triplex formation would constitute an exciting and important future research area.

When investigating the relative performance of each of the codes learned by the expectation-maximization algorithm described herein, we could observe differences in performance of codes between datasets. For example, code 1 was the highest performing code in prediction of RADICL-seq RNA–DNA interactions but was outperformed by codes 4, 2 and 3 in recall of RedC interactions. Given the distinct nature of the methods used to identify these RNA–DNA interactions, it is difficult to establish whether these are true differences in performance or merely reflect the subtle differences between the wet-lab methodologies. For instance, it may be that one of the methods enriches *trans*

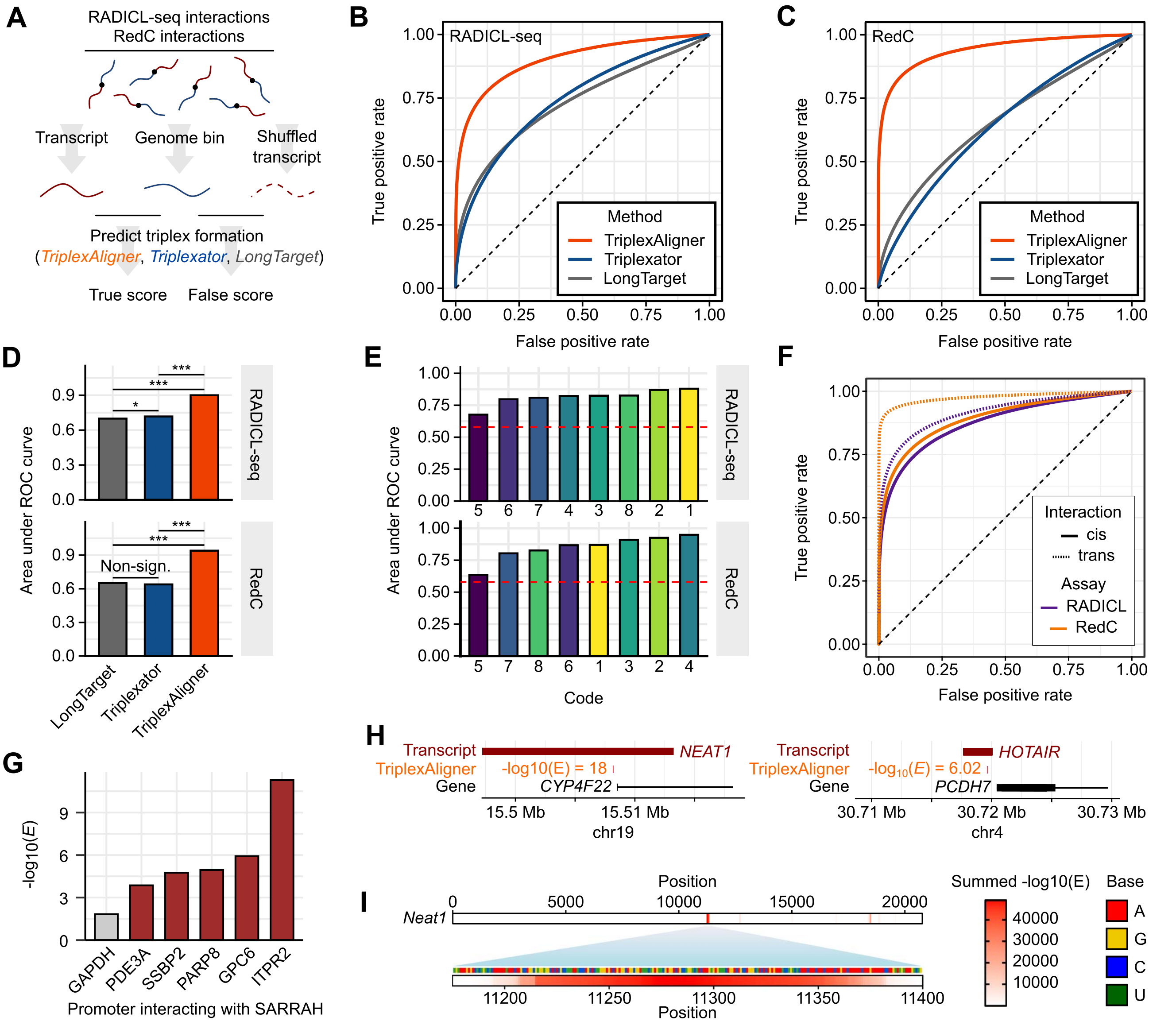


Figure 4. Computational validation of TriplexAligner using RNA–DNA interaction data and published RNA·DNA:DNA triple helix interactions. **(A)** Schematic outlining the computational validation of *TriplexAligner*, using global RNA–DNA interactions identified by either RADICL-seq or RedC and subjecting the corresponding RNA and DNA sequences to prediction of RNA·DNA:DNA triplex formation with *TriplexAligner*, *Triplexator* and *LongTarget*. Negative interaction data were generated by shuffling of RNA sequences. **(B)** ROC curves summarizing performance of *TriplexAligner* (orange), *Triplexator* (blue) and *LongTarget* (grey) in prediction of RADICL-seq RNA–DNA interactions. **(C)** ROC curves summarizing performance of *TriplexAligner* (orange), *Triplexator* (blue) and *LongTarget* (grey) in prediction of RedC RNA–DNA interactions. **(D)** Comparison of area under the ROC curves displayed in B and C (Non-sign. $P > 0.05$, $^{*}P < 0,05$, $^{***}P < 0.001$, bootstrapping ($n = 2000$)). **(E)** Area under the ROC curves of individual *TriplexAligner* code models for RADICL-seq and RedC RNA–DNA interactions. **(F)***TriplexAligner* ROC curves for *cis* (RNA gene locus and interaction site on the same chromosome, solid line) and *trans* (RNA gene locus and interaction site on different chromosomes, dashed line) RNA–DNA interactions arising from RADICL-seq (purple) and RedC (orange) data. **(G)***TriplexAligner* -log10(E) values for predicted interactions between lncRNA *SARRAH* and published interacting promoters *ITPR2*, *PARP8*, *PDE3A*, *SSBP2* and *GPC6*, in comparison to the negative control promoter of *GAPDH*. **(H)***TriplexAligner* predictions of published RNA·DNA:DNA triplex helix formation between the lncRNAs *NEAT1* and *HOTAIR* and the promoter regions of *CYP4F22* and *PCDH7*, respectively. **(I)** Schematic of the lncRNA *Neat1* showing most commonly predicted sites of RNA·DNA:DNA triple helix formation in the lncRNA against multiple gene promoters dysregulated after *Neat1* knockout.

interactions more so than the other or fails to effectively remove interactions arising from nascent transcription. Another aspect to consider is species differences, given that the RADICL-seq data used here arise from murine cells, and the RedC from human material. Naively, we would not expect this to make a difference. However, so little is known about the conditions which facilitate triplex formation; there could be unknown species-specific co-factors which favour different triplex base pairings.

When compared with *Triplexator*, the most widely used tool for prediction of triplex formation, outputs from *TriplexAligner* differ in a number of aspects. Most notable is that triplexes predicted by *TriplexAligner* tend to be far broader than those predicted by *Triplexator*. There are several potential reasons for this observation, all of which are figurative, given the lack of wet-lab data. From a technical perspective, the implementation of *TriplexAligner* as a local aligner using Karlin–Altschul statistics [54] means that the score metric is highly dependent on the width of the alignment, and thus, broad alignments are more likely to be reported as interacting regions. Triplexes predicted by *Triplexator* are - by default - a minimum length of 20 base pairs. During

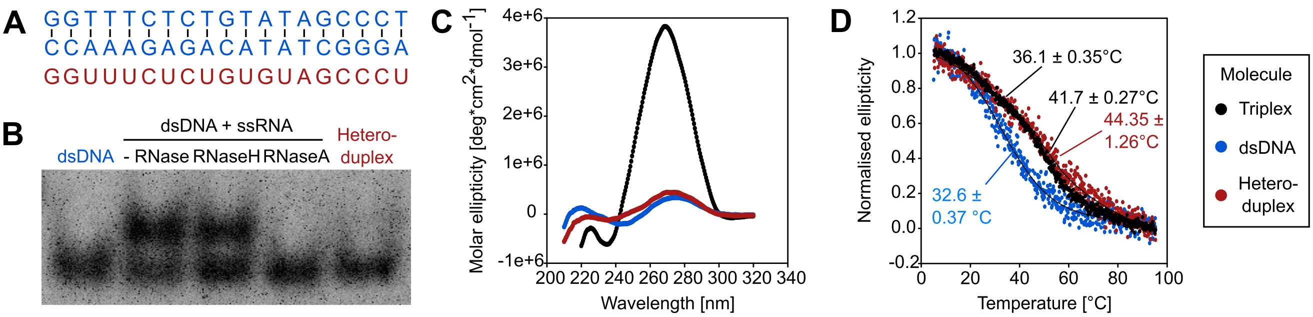


Figure 5. Biophysical validation of interacting DNA and RNA sequences as predicted by TriplexAligner. **(A)** Maximal scoring DNA (blue) and RNA (red) subsequences across RADICL-seq interactions as predicted by *TriplexAligner*, which were synthesized *in vitro* and used in subsequent biophysical experiments investigating RNA·DNA:DNA triple helix formation. **(B)** EMSA using combinations of DNA and RNA (shown in A), as either double-stranded DNA (dsDNA), double-stranded DNA and single-stranded RNA (dsDNA + ssRNA) and single-stranded DNA in combination with single-stranded RNA (heteroduplex). RNA·DNA:DNA triple helix formation was investigated in RNase-free conditions (- RNase), in combination with RNaseH or in combination with RNaseA. **(C)** CD spectroscopy of double-stranded DNA and single-stranded RNA (Triplex, black), double-stranded DNA (dsDNA, blue) and single-stranded DNA with single-stranded RNA (Heteroduplex, red). **(D)** Melting analysis DNA and RNA molecules (described in C), with melting points labelled and annotated.

prediction of RADICL-seq and RedC interactions, the widths of *Triplexator*-predicted interactions did not exceed 30 base pairs. In comparison, the alignments reported by *TriplexAligner* exceeded 100 base pairs on a number of occasions. Due to technical and financial restraints, it is challenging to experimentally determine whether these alignments are reflected in biological systems. However, longer tracts of triplex formation could permit increased specificity of interactions between transcripts and genomic loci, thereby mediating more precise regulatory relationships between RNAs and target genes. Alongside this, longer tracts of interaction could result in increased stability, increasing the robustness of the regulatory mechanism. Alternatively, the long tracts of interaction predicted by *TriplexAligner* may provide the conditions for RNA–DNA interaction to take place along the length of the tract in a dynamic manner. This would entail that the entire length is not interacting at any one time, and instead, sub-tracts of RNA and DNA interact when spatio-temporal conditions are favourable.

Whilst *TriplexAligner* recalls RNA–DNA interactions more accurately than previously published tools, it remains imperfect. There exist a variety of reasons for this. In *TriplexAligner*, the assumption is made that triplex formation takes place between two linear molecules, consequently disregarding the influence of higher order structures. Whilst *TriplexAligner* does not consider chromatin state, it was previously shown that triplexDNA-seq data isare enriched in regions of open chromatin [25]. It is therefore likely that motifs used to develop *TriplexAligner* arose from open chromatin, in spite of the previously reported repressive functions of triplex formation [62–64]. Consequently, triplexDNA-seq and therefore *TriplexAligner* could be biased towards triplex formation with activatory functions. Validating the effects of chromatin conformation on triplex formation would require non-steady-state data, where both differential chromatin states and differential triplexDNA-seq regions could be identified, and these data do not currently exist.

Beyond chromatin conformation, it is also possible that triplex formation is influenced by more complex 3D structures of both RNA and DNA. Sites of predicted triplex formation are correlated with 3D genome structure [65], but it is unclear when triplexes form relative to the establishment of 3D genomic structures. It is also likely that the secondary structures of triplex-forming transcripts affect on the formation of triplexes. Transcripts can fold into complex structures, resulting in regions with divergent accessibility [66]. This would restrict regions of RNA which are free to interact with DNA, alongside forming new interfaces which are irrecoverable from linear molecules. Integration of features beyond linear RNA and DNA sequences therefore represents an important future research topic. Experimental methods to examine high-resolution 3D structures of nucleic acids, such as RNA SHAPE [67], remain complex and challenging. However, progress in this field would provide further insight into the 3D requirements for RNA–DNA interactions to successfully form. Here again, the roles of proteins in the facilitation of RNA–DNA interactions are unclear, but likely highly important. The incorporation of 2D RNA structure prediction, such as tools available from the ViennaRNA package [68], into *TriplexAligner* could be a useful addition to the tool. In addition, the use of RNA aligners which consider secondary and tertiary structures [69–71], could be considered as a downstream analysis option, in order to compare triplex-forming RNA structures and identify important regulatory structures, which may facilitate triplex formation.

TriplexAligner was developed with the aim of providing researchers with a method of predicting RNA–DNA interactions which is grounded in data. By leveraging of triplex-forming sequences captured in next-generation sequencing experiments, *TriplexAligner* reports predictions with a basis in data, and which extend beyond canonical and discrete base pairing rules. As such, *TriplexAligner* is a unique tool with the potential to direct wet-lab research on regulatory RNA networks and thereby further clarify the role of RNA–DNA interactions in epigenetics.

Key Points

- Short, reproducible sequence motifs can be identified from triplexRNA- and triplexDNA-seq data.
- Expectation-Maximization can be used to learn RNA–DNA base pairing rules from enriched motifs.
- Both canonical Hoogsteen base pairing rules and previously unreported base pairing rules were identified by this approach and could be positively correlated with previous *in vitro* work.
- Implementation of the learned RNA–DNA base pairings in a local alignment program permitted the more accurate prediction of RNA–DNA interactions than previously published gold-standard tools.

- An RNA–DNA interaction predicted in the course of this analysis could be shown to be a biophysically valid RNA·DNA:DNA triple helix *in vitro*.

Data availability

Accession numbers for published data used in this study are detailed in *Materials and Methods*.

Code availability

The R package for *TriplexAligner* is available at https://github.com/SchulzLab/TriplexAligner. The code used in the learning of probabilistic RNA-DNA mapping codes is available at https://github.com/SchulzLab/Codefinder.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Authors' contributions

T.W., R.P.B., M.H.S. and M.S.L. designed the study and wrote the manuscript. T.W. and M.H.S. formulated and wrote the algorithms and carried out computational experiments. S.S., N.M.K., J.K.B. and H.S. provided instruments, performed experiments and analysed data. L.A., A.D., S.E., J.A.O. and A.B. provided important data and advice. All authors read and commented on the manuscript.

Funding

Goethe University Frankfurt am Main, the German Centre for Cardiovascular Research (DZHK), the DFG excellence cluster EXS2026 (Cardio-Pulmonary Institute); Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 403584255 - TRR 267. Work at BMRZ is supported by the state of Hesse.

References

1. Holoch D, Moazed D. RNA-mediated epigenetic regulation of gene expression. *Nat Rev Genet* 2015; **16**(2): 71–84.
2. Oo JA, Brandes RP, Leisegang MS. Long non-coding RNAs: novel regulators of cellular physiology and function. *Pflügers Archiv-European Journal of Physiology* 2021;1–14.
3. Takemata N, Ohta K. Role of non-coding RNA transcription around gene regulatory elements in transcription factor recruitment. *RNA Biol* 2017; **14**(1): 1–5.
4. Andric V, Nevers A, Hazra D, et al. A scaffold lncRNA shapes the mitosis to meiosis switch. *Nat Commun* 2021; **12**(1): 1–12.
5. Shimada Y, Mohn F, Bühler M. The RNA-induced transcriptional silencing complex targets chromatin exclusively via interacting with nascent transcripts. *Genes Dev* 2016; **30**(23): 2571–80.
6. Csankovszki G, Nagy A, Jaenisch R. Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *J Cell Biol* 2001; **153**(4): 773–84.
7. Zhang J-Y, Zheng K-w, Xiao S, et al. Mechanism and manipulation of DNA: RNA hybrid G-quadruplex formation in transcription of G-rich DNA. *J Am Chem Soc* 2014; **136**(4): 1381–90.
8. García-Muse T, Aguilera A. R loops: from physiological to pathological roles. *Cell* 2019; **179**(3): 604–18.
9. Santos-Pereira JM, Aguilera A. R loops: new modulators of genome dynamics and function. *Nat Rev Genet* 2015; **16**(10): 583–97.
10. Crossley MP, Bocek M, Cimprich KA. R-loops as cellular regulators and genomic threats. *Mol Cell* 2019; **73**(3): 398–411.
11. Felsenfeld G, Davies DR, Rich A. Formation of a three-stranded polynucleotide molecule. *J Am Chem Soc* 1957; **79**(8): 2023–4.
12. Escudeé C, Francçois J-C, Sun J-S, et al. Stability of triple helices containing RNA and DNA strands: experimental and molecular modeling studies. *Nucleic Acids Res* 1993; **21**(24): 5547–53.
13. Mondal T, Subhash S, Vaid R, et al. MEG3 long noncoding RNA regulates the tgf-β pathway genes through formation of RNA–DNA triplex structures. *Nat Commun* 2015; **6**(1): 1–17.
14. Chu C, Kun Q, Zhong FL, et al. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell* 2011; **44**(4): 667–78.
15. Vrba L, Futscher BW. Epigenetic silencing of lncRNA MORT in 16 TCGA cancer types. *F1000Research* 2018;7.
16. Jiang M-C, Ni J-J, Cui W-Y, et al. Emerging roles of lncRNA in cancer and therapeutic opportunities. *Am J Cancer Res* 2019; **9**(7): 1354.
17. Zhou Z, Giles KE, Felsenfeld G. DNA. RNA triple helix formation can function as a cis-acting regulatory mechanism at the human β-globin locus. *Proc Natl Acad Sci* 2019; **116**(13): 6130–9.
18. Garratt H, Ashburn R, Sopić M, et al. Long non-coding RNA regulation of epigenetics in vascular cells. *Non-coding RNA* 2021; **7**(4): 62.
19. Maldonado R, Filarsky M, Grummt I, et al. Purine–and pyrimidine–triple-helix-forming oligonucleotides recognize qualitatively different target sites at the ribosomal DNA locus. *RNA* 2018; **24**(3): 371–80.
20. Buske FA, Bauer DC, Mattick JS, et al. Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome Res* 2012; **22**(7): 1372–81.
21. Kuo C-C, Hänzelmann S, Cetin NS, et al. Detection of RNA–DNA binding sites in long noncoding RNAs. *Nucleic Acids Res* 2019; **47**(6): e32–2.
22. He S, Zhang H, Liu H, et al. LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis. *Bioinformatics* 2015; **31**(2): 178–86.
23. Antonov IV, Mazurov E, Borodovsky M, et al. Prediction of lncRNAs and their interactions with nucleic acids: benchmarking bioinformatics tools. *Brief Bioinform* 2019; **20**(2): 551–64.
24. Zhang Y, Long Y, Kwoh CK. Deep learning based dna: Rna triplex forming potential prediction. *BMC bioinformatics* 2020; **21**(1): 1–13.
25. Cetin NS, Kuo C-C, Ribarska T, et al. Isolation and genome-wide characterization of cellular DNA: RNA triplex structures. *Nucleic Acids Res* 2019; **47**(5): 2306–21.
26. Zhou B, Li X, Luo D, et al. GRID-seq for comprehensive analysis of global RNA–chromatin interactions. *Nat Protoc* 2019; **14**(7): 2036–68.
27. Zhong S, Sridhar B, Rivas-Astroza M, et al. Mapping RNA-chromatin interactions. *FASEB J* 2018; **32**:525–2.
28. Wu W, Yan Z, Nguyen TC, et al. Mapping RNA–chromatin interactions by sequencing with iMARGI. *Nat Protoc* 2019; **14**(11): 3243–72.
29. Bell JC, Jukam D, Teran NA, et al. Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. *Elife* 2018; **7**:e27024.

30. Bonetti A, Agostini F, Suzuki AM, et al. RADICL-seq identifies general and cell type–specific principles of genome-wide RNA-chromatin interactions. *Nat Commun* 2020; **11**(1): 1–14.
31. Gavrilov AA, Zharikova AA, Galitsyna AA, et al. Studying RNA–DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics. *Nucleic Acids Res* 2020; **48**(12): 6699–714.
32. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**(4): 357–9.
33. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010; **38**(4): 576–89.
34. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**(6): 841–2.
35. Bailey TL, Johnson J, Grant CE, et al. The MEME suite. *Nucleic Acids Res* 2015; **43**(W1): W39–49.
36. Bioconductor Core Team and Bioconductor Package Maintainer. *TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb object(s)*, 2021, R package version 3.13.0.
37. Morgan M. *BiocManager: Access the Bioconductor Project Package Repository*, 2021, R package version 1.30.16.
38. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021.
39. Rainer J. *EnsDb.Hsapiens.v86: Ensembl based annotation package*, 2017, R package version 2.99.0.
40. Howe KL, Achuthan P, Allen J, et al. Ensembl 2021. *Nucleic Acids Res* 2021; **49**(D1): D884–91.
41. Fornes O, Castro-Mondragon JA, Khan A, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2020; **48**(D1): D87–92.
42. Giudice G, Sánchez-Cabo F, Torroja C, et al. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database* 2016; **2016**.
43. Kunkler CN, Hulewicz JP, Hickman SC, et al. Stability of an RNA. DNA–DNA triple helix depends on base triplet composition and length of the RNA third strand. *Nucleic Acids Res* 2019; **47**(14): 7213–22.
44. Murtagh F, Legendre P. *Ward's hierarchical clustering method: clustering criterion and agglomerative algorithm* arXiv preprint arXiv:1111.6285, 2011.
45. Altschul SF, Wootton JC, Zaslavsky E, et al. The construction and use of log-odds substitution scores for multiple sequence alignment. *PLoS Comput Biol* 2010; **6**(7):e1000852.
46. Pagès H, Aboyoun P, Gentleman R, et al. *Biostrings: Efficient manipulation of biological strings*, 2021, R package version 2.60.2.
47. Altschul SF, Bundschuh R, Olsen R, et al. The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res* 2001; **29**(2): 351–61.
48. Millard SP. *EnvStats: An R Package for Environmental Statistics*. New York: Springer, 2013.
49. Bioconductor Core Team and Bioconductor Package Maintainer. *TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb object(s)*, 2021, R package version 3.13.0.
50. Bioconductor Core Team and Bioconductor Package Maintainer. *TxDb.Mmusculus.UCSC.mm10.knownGene: Annotation package for TxDb object(s)*, 2019, R package version 3.10.0.
51. Patro R, Duggal G, Love MI, et al. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017; **14**(4): 417–9.
52. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; **12**:77.
53. Sing T, Sander O, Beerenwinkel N, et al. ROCR: visualizing classifier performance in R. *Bioinformatics* 2005; **21**(20): 7881.
54. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci* 1990; **87**(6): 2264–8.
55. Trembinski DJ, Bink DI, Theodorou K, et al. Aging-regulated anti-apoptotic long non-coding RNA sarrah augments recovery from acute myocardial infarction. *Nat Commun* 2020; **11**(1): 1–14.
56. Kalwa M, Hänzelmann S, Otto S, et al. The lncRNA HOTAIR impacts on mesenchymal stem cells via triple helix formation. *Nucleic Acids Res* 2016; **44**(22): 10631–43.
57. Mello SS, Sinow C, Raj N, et al. Neat1 is a p53-inducible lincRNA essential for transformation suppression. *Genes Dev* 2017; **31**(11): 1095–108.
58. Amon JD, Koshland D. RNase H enables efficient repair of R-loop induced DNA damage. *Elife* 2016; **5**:e20533.
59. Do CB, Batzoglou S. What is the expectation maximization algorithm? *Nat Biotechnol* 2008; **26**(8): 897–9.
60. Guo Y, Schuurmans D. Convex relaxations of latent variable training. *Advances in Neural Information Processing Systems* 2007; **20**.
61. West JA, Davis CP, Sunwoo H, et al. The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol Cell* 2014; **55**(5): 791–802.
62. O'Leary VB, Ovsepian SV, Carrascosa LG, et al. PARTICLE, a triplex-forming long ncRNA, regulates locus-specific methylation in response to low-dose irradiation. *Cell Rep* 2015; **11**(3): 474–85.
63. O'Leary VB, Ovsepian SV, Smida J, et al. PARTICLE- The RNA podium for genomic silencers. *J Cell Physiol* 2019; **234**(11): 19464–70.
64. Schmitz K-M, Mayer C, Postepska A, et al. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev* 2010; **24**(20): 2264–9.
65. Farabella I, Di Stefano M, Soler-Vila P, et al. Three-dimensional genome organization via triplex-forming RNAs. *Nat Struct Mol Biol* 2021; **28**(11): 945–54.
66. Sato K, Akiyama M, Sakakibara Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat Commun* 2021; **12**(1): 1–9.
67. Spitale RC, Crisalli P, Flynn RA, et al. RNA SHAPE analysis in living cells. *Nat Chem Biol* 2013; **9**(1): 18–20.
68. Gruber AR, Lorenz R, Bernhart SH, et al. The vienna RNA websuite. *Nucleic Acids Res* 2008; **36**(suppl_2): W70–4.
69. Siebert S, Backofen R. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics* 2005; **21**(16): 3352–9.
70. Techa-Angkoon P, Sun Y. glu-RNA: aliGn highLy strUctured ncRNAs using only sequence similarity. In: *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, 2013, 508–17.
71. Gong S, Zhang C, Zhang Y. RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. *Bioinformatics* 2019; **35**(21): 4459–61.

5.3 **Research article:** Conformational switch in the ribosomal protein S1 guides unfolding of structured RNAs for translation initiation

Nusrat Shahin Qureshi, Jasleen Kaur Bains, Sridhar Sreeramulu, Harald Schwalbe and Boris Fürtig; *Nucleic Acids Research* **46**, 10917-10929 (2018)

In this work, the interaction of the ribosomal protein S1 (rS1) with the 5'-UTR of the adenine sensing riboswitch (Asw) from the human pathogenic bacterium *V. vulnificus* was characterized by solution NMR spectroscopy and biological activity assays. The rS1 domains D3 and D4 are responsible for the mRNA binding and domain D5 for chaperone activity. Thus, the RNA chaperone character induces translation initiation by destabilizing the secondary structure of the riboswitch. The *in vitro* transcription-translation data demonstrated the importance of the mRNA-binding domains D3-D5 for efficient translation, as protein expression decreased in the absence of each domain. The presence of rS1 mediated a population shift of Asw from the hairpin conformation (14% at 35°C) to the single-stranded conformation (86% at 35°C).

EMSA, CD, and NMR experiments on the interaction of the rS1 and Asw were performed by N. S. Qureshi. The project was designed, and the manuscript was written by N. S. Qureshi, H. Schwalbe, and B. Fürtig. The author of this thesis expressed and purified the full-length rS1 and its truncated mutants, which were further investigated with the coupled *in vitro* transcription-translation assay.

https://doi.org/10.1093/nar/gky746

Published online 16 August 2018 *Nucleic Acids Research, 2018, Vol. 46, No. 20* ***10917–10929***
doi: 10.1093/nar/gky746

Conformational switch in the ribosomal protein S1 guides unfolding of structured RNAs for translation initiation

Nusrat Shahin Qureshi, Jasleen Kaur Bains, Sridhar Sreeramulu, Harald Schwalbe* **and Boris Fürtig***

Institute for Organic Chemistry and Chemical Biology, Center for Biomolecular Magnetic Resonance (BMRZ), Johann Wolfgang Goethe-Universität, Frankfurt am Main, Hessen 60438, Germany

Received May 23, 2018; Revised July 31, 2018; Editorial Decision August 02, 2018; Accepted August 03, 2018

ABSTRACT

Initiation of bacterial translation requires that the ribosome-binding site in mRNAs adopts single-stranded conformations. In Gram-negative bacteria the ribosomal protein S1 (rS1) is a key player in resolving of structured elements in mRNAs. However, the exact mechanism of how rS1 unfolds persistent secondary structures in the translation initiation region (TIR) is still unknown. Here, we show by NMR spectroscopy that *Vibrio vulnificus* rS1 displays a unique architecture of its mRNA-binding domains, where domains D3 and D4 provide the mRNA-binding platform and cover the nucleotide binding length of the full-length rS1. D5 significantly increases rS1's chaperone activity, although it displays structural heterogeneity both in isolation and in presence of the other domains, albeit to varying degrees. The heterogeneity is induced by the switch between the two equilibrium conformations and is triggered by an order-to-order transition of two mutually exclusive secondary structures (β-strand-to-α-helix) of the 'AERERI' sequence. The conformational switching is exploited for melting of structured 5′-UTR's, as the conformational heterogeneity of D5 can compensate the entropic penalty of complex formation. Our data thus provides a detailed understanding of the intricate coupling of protein and RNA folding dynamics enabling translation initiation of structured mRNAs.

INTRODUCTION

RNA chaperones form a sequentially and structurally diverse class of proteins. They are important for promoting many RNA-regulated cellular processes (1). By destabilizing secondary structures of RNA (2,3) they accelerate RNA refolding and thus resolve RNAs trapped in misfolded, nonfunctional conformations. In contrast to RNA helicases, RNA chaperones do not require an external energy source. Early on, it was suggested that RNA chaperones decrease energy barriers between non-native RNA structures and thus assist the conformational search of RNAs for their native fold (4,5). The driving force for the unfolding reaction, however, has remained elusive. An 'entropy transfer' model was proposed, in which disordered regions—as frequently found in RNA and in protein chaperones – serve as energy reservoir. According to this model, binding induces a disorder-to-order transition within the chaperone to compensate the order-to-disorder transition of the substrate (6).

RNA chaperone activity is abundant in all organisms and necessary for many RNA-binding proteins required for RNA function, transcription and decay. One prime example is the ribosomal protein S1 (rS1) that exhibits RNA chaperone activity. It is associated to the small ribosomal subunit (30S), where it locates within the cleft between the head and platform on the solvent accessible side of the ribosome (7–9). Particularly, it is close to the 3′-end of the 16S rRNA and thus spatially close to the anti-Shine-Dalgarno (anti-SD) sequence (10). Here, 5′-untranslated regions (5′-UTRs) with their complementary SD-sequence are positioned during translation initiation (11). During translation initiation, proper positioning of the start codon AUG is crucial for setting the correct reading frame. Typically, the positioning for the correct reading frame is guided by ribosomal RNA provided that a single-stranded and strong SD-sequence is present in the mRNA transcript (12–14). However, productive binding to the 16S rRNA is often hampered, as either the SD-sequence found in the mRNA displays poor complementarity or is sequestered in intra-molecular base pairing interactions (15). In these cases, the rS1 is pivotal for translation initiation (16,17).

rS1 has a modular structure. In Gram-negative bacteria it generally consists of altogether six imperfect oligonu-

[*] To whom correspondence should be addressed. Email: fuertig@nmr.uni-frankfurt.de
Correspondence may also be addressed to Harald Schwalbe. Tel: +49 69 7982 9157; Fax: +49 69 7982 9515; Email: schwalbe@nmr.uni-frankfurt.de

cleotide binding fold (OB fold) motifs (16,18). These OB fold motifs constitute a five-stranded antiparallel β-barrel composed of ∼75 amino acids and preferentially bind to single-stranded RNAs (19,20). They are connected via linkers of 10–15 amino acids that provide the rS1 with the flexibility and the adaptability needed for recognition of miscellaneous mRNAs. Furthermore, the motifs vary in sequence and thereby tune the biological function of each domain. The two N-terminal domains (D1-D2) function as platform for ribosome-binding through protein-protein interactions with the ribosomal protein S2 (8,21), while the four C-terminal domains (D3-D6) harbor the mRNA-binding site (16). Thus, rS1 contains a protein- and an RNA-binding site in a single protein, mediating interactions between proteins and RNAs. In line with this ability, rS1 is known to activate the RegB endonuclease of the T4 bacteriophage (22,23) and also acts as a host factor during infection with *Qβ* phage as an integral part of the replication machinery (24,25). Moreover, as the majority of initiation in prokaryotes requires accessibility of both, the SD sequence and the AUG start codon to the ribosome (26,27), the central role of rS1 is to recruit single-stranded regions within the 5′-UTR of mRNA transcripts. Recently, it was shown that rS1 is also critically needed in the establishment of translation by polysomes (28).

Generally, mRNA-based regulation of translation in bacteria can proceed via three different mechanisms: masking of the ribosomal binding site (RBS), preventing accommodation of the mRNA into the decoding channel or indirect competition due to steric clashes (29). These three mechanisms have in common that the initiation region of the mRNA is involved in molecular interaction either *in trans* with other RNAs and proteins or *in cis* with sequences upstream from the initiation codon (30–32). To drive translation of highly structured mRNAs that either mask the RBS or prevent accommodation, rS1's ability to melt local secondary structures is exploited (33). This function is essential as rS1 deletion is lethal to prokaryotic cells (34).

Here, we characterize the interaction of the rS1 protein from the human pathogenic bacterium *Vibrio vulnificus* with the structured 5′-UTR of the adenine-sensing riboswitch (Asw) from the same organism (35). The riboswitch controls the translation of the adenosine deaminase from the *add* gene and modulates the expression levels by a novel three-state-mechanism enabling a temperature-compensated regulation of translation by the ligand adenine (36–41). The regulation involves the switch between sequestration and liberation of the ribosome-binding site.

Here, we show for rS1 from *V. vulnificus* by activity assays and solution NMR that the core region for the mRNA interaction is composed of domains D3, D4 and D5 and that these domains contribute differently in the interplay with structured mRNAs. Domains D3 and D4 adopt stable OB folds and provide a platform for mRNA-binding. In sharp contrast, domain D5 adopts an intrinsically bistable fold that exhibits a conformational equilibrium between a predominantly structured ($D5^{OB}$) and a predominantly unstructured ($D5^{res\text{-}\alpha}$) state. We find that conformational switching is triggered by a β-strand-to-α-helix transition, thus involves an order-to-order transition on the secondary structure level, leading to a domain-wide order-to-disorder transition. Our study provides insights into how the rS1 exploits its modular domain architecture to enable translation of structured RNAs, by locally melting RNA secondary structure.

MATERIALS AND METHODS

Expression and purification of rS1 constructs

All proteins were cloned with NcoI and BamHI as restriction sites into an in-house modified pKM260 vector (pKM-TX). They were expressed as fusion proteins in BL21(DE3) cells carrying a N-terminal polyhistidine (His_6) and thioredoxin (trx) tag (His_6-trx-TEV-rS1construct, see also Supplementary Table S1). For isotope labeling cells were grown in M9 medium supplemented with 1 g/l $^{15}NH_4Cl$ and either 2 g/l ^{13}C-glucose or 4–10 g/l ^{12}C-glucose. Expression was induced at $OD_{600} = 0.6–0.8$ with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). Proteins were expressed over night at 16–20°C and were purified via HisTrap HP (GE Healthcare) columns using 50 mM Tris/HCl (pH 8), 300 mM NaCl, 10 mM imidazol and 10 mM β-mercaptoethanol as lysis buffer. The tag was cleaved using TEV (tabacco etch virus) protease and removed via HisTrap HP column. All proteins were further purified via size exclusion chromatography (Superdex 75 or Superdex 200 columns, 26/60, GE Healthcare) with the NMR buffer 25 mM potassium phosphate (pH 7.2) 150 mM KCl, 5 mM DTT. In case of rS1-D4 low yields were obtained in the soluble fraction, hence the size exclusion chromatography was omitted and instead the rS1-D4 was transferred into NMR buffer using PD-10 columns (GE Healthcare).

RNA preparation

The short RNA fragments Asw-6 to Asw-14 were purchased from Dharmacon carrying 2′-ACE protecting groups. They were deprotected and desalted using the manufacturers' protocol. Lyophilized RNA was reconstituted with NMR buffer (25 mM potassium phosphate (pH 7.2) 150 mM KCl, 5 mM DTT).

Full-length Asw (112 nt) was transcribed with a 5′-hammerhead ribozyme from PCR template via *in vitro* transcription as described in (39). Asw-42 was transcribed without ribozyme and with an additional G at the 5′-end to enhance transcription yield. The sequences are listed in Supplementary Table S2. For both RNAs 3′-homogeneity was achieved following the instructions in (42). The RNAs were purified either by HPLC or PAGE and were refolded by thermal denaturation at 95°C, subsequent tenfold dilution with ice cold water and incubation on ice for 1 h. They were exchanged into NMR buffer (25 mM potassium phosphate (pH 7.2), 150 mM KCl, 5 mM DTT) using centrifugal concentrators (Sartorius AG).

Electrophoretic mobility shift assay (EMSA)

The EMSAs were performed at constant RNA concentration (4 μM) and increasing amounts of protein (0–25 equivalents). All samples were prepared in EMSA buffer (25 mM potassium phosphate (pH 7.2), 75 mM KCl, 5 mM DTT) and were incubated for 30 minutes on ice. Bands were separated on discontinuous gels (6%/12% PAGE) at low powers (<500 mW, 3.5–5.0 V/cm) in a Tris-acetate buffer (50 mM Tris (pH 8.3), 100 mM sodium acetate). The band separation was optimized for each rS1-construct in order to allow migration of the RNP and the free RNA into the gel (rS1: 500 mW, 14 h; rS1-D3456: 250 mW, 14 h; rS1-D345: 250 mW, 14 h; rS1-D34: 500 mW, 7 h; rS1-D45: 500 mW, 7 h). The RNA was visualized upon staining with GelRed® (Biotium) and subsequent excitation of the bands at 254 nm.

CD RNA chaperone assay

RNA chaperone assays were performed on a JASCO J-810 spectrophotometer using a 1 mm path length cuvette screening a range from 220 to 320 nm. All CD experiments were acquired in NMR buffer (25 mM potassium phosphate (pH 7.2), 150 mM KCl, 5 mM DTT) at a Asw-42 concentration of 10 μM. The assays were performed at 10°C. 0–4 equivalents of the respective protein were stepwise added to the RNA. Each titration point was measured with ten scans and the data was smoothed. In case of rS1-D345 the chaperone assay was additionally performed at 25°C and 35°C. CD melting curve of Asw-42 was measured at a wavelength of 264 nm in a range from 10 to 95°C. Temperature dependent CD spectra of free Asw-42 were acquired in a range from 220 to 320 nm using three scans per temperature interval. The temperature ranged from 10°C to 65°C using 5°C intervals. The CD spectra were smoothed and evaluated using Microsoft Excel 2010 and SigmaPlot 11.0.

In vitro transcription-translation assay

The coupled transcription-translation assays were performed as described previously (43) using the shifted green fluorescent protein to monitor expression levels. The protein was synthesized from plasmid (pIVEX 2.3d) containing T7 regulatory elements, a strong RBS and the sequence of sGFP. To study the influence of rS1-constructs on translation efficiency, the assays were performed as duplicates with rS1-depleted ribosomes reconstituted in S100 extract. The respective rS1-proteins were added in equimolar ratios to the ribosome concentration. 25 μl reactions were performed in 96-well V-shape microplate (Greiner Bio-One, Frickenhausen, Germany) at 30°C for 1 h. 20 μl of the reaction was transferred in μClear® 384-well flat bottom microplates (Greiner Bio-One, Frickenhausen, Germany). The fluorescence measurements were performed with an Inifinite® 200 PRO plate reader (Tecan, Männedorf, Switzerland) using an excitation wavelength of 484 nm and emission wavelength of 510 nm.

NMR spectroscopy

All samples were prepared in NMR buffer (25 mM potassium phosphate (pH 7.2), 150 mM KCl, 5 mM DTT and 5–7% D_2O). Trimethylsilylpropanoic acid (TSP) was used as chemical shift standard.

NMR experiments were performed on Bruker spectrometers (either 600, 800, 900 or 950 MHz; Rheinstetten, Germany) equipped with cryoprobes. Standard Bruker pulse sequences as distributed with Topspin 3.5 were used (44–51). Data was processed using Topspin 3.5 (Bruker, Rheinstetten, Germany) and analyzed with Sparky 3.114 (52). Assigned chemical shifts for rS1-D5 and rS1-D345 are deposited under BMRB accession codes 27489 and 27490, respectively.

All NMR titrations were performed at 35°C with protein concentrations of 100 μM. For each titration point ^{1}H–^{15}N-BEST-TROSY was acquired keeping all acquisition parameters constant. For rS1-D34 the RNAs were added stepwise in molar ratios of [RNA]:[protein]: 0, 0.1, 0.2, 0.3, 0.6, 1, 2, 4, 5.5 and the titration experiments were performed at 600 MHz. For rS1-D345 the titration was performed at 950 MHz with stepwise adding Asw-42 in molar ratios of [RNA]:[protein]: 0, 0.1, 0.2, 0.3, 0.6, 1, 2.

Chemical shift perturbations were calculated using $\Delta\delta = \sqrt{(0.1 \cdot \delta_N)^2 + (\Delta\delta_H)^2}$. Dissociation constants were determined as described in (53) by plotting largest CSP against the RNA concentration and fitting with $\Delta \quad \delta_{obs} = \Delta\delta_{max}\{([P]_0 + [R]_0 + [K_D]) - \sqrt{([P]_0 + [R]_0 + [K_D])^2 - 4[P]_0[R]_0}\}/2[P]_0$.

RESULTS

Delineation of mRNA-interaction-core of rS1 from *V. vulnificus*

In order to dissect the function of the individual mRNA-binding domains, we investigated several fragments of the binding region (D3–D6, Figure 1A) and studied their interaction with the adenine-sensing riboswitch. We used the 112 nucleotide long full-length *add* Asw and in addition a 42 nucleotide long fragment of this riboswitch ranging from G85 to U125 (from here on referred as Asw-42, see also Figure 1C). Asw-42 contains the translation initiation region (TIR) of the full-length Asw including the helix P5 followed by the SD sequence (GAA) and the start codon (AUG). This RNA provides a single-stranded 3′-end and an AU-rich sequence within the structured TIR of the riboswitch as constitutive binding platform for rS1 (54).

We probed RNA-binding ability for eight rS1 constructs (Figure 1A and B) with the full-length *add* Asw as RNA substrate in electrophoretic mobility shift assays (EMSAs). In line with previously reported results (16), the full-length rS1 binds the RNA with an apparent $K_D{}^{rS1}$ of 3 ± 1 μM. Deletion of the first two domains D1 and D2, responsible for ribosome-binding, results in proteins that retain the full RNA-binding ability. Removal of domain D6 does not significantly reduce the affinity towards RNA. However, deletion of D5 considerably decreases the RNA-binding affinity of rS1 (Figure 1B). Nevertheless, as RNA-binding to the two-domain protein rS1-D34 is detected this construct represents the minimal unit for mRNA-binding. In sharp contrast, no RNA-binding for rS1-D45 (Supplementary Figure S1) or for the single-domain fragments (D3, D4, D5) could be detected within the tested concentration range.

In order to evaluate the RNA-melting properties of the rS1 constructs, we performed CD titration experiments with Asw-42. Thermal unfolding was used as a benchmark for the single-stranded RNA state (Figure 1D and E). The ob-

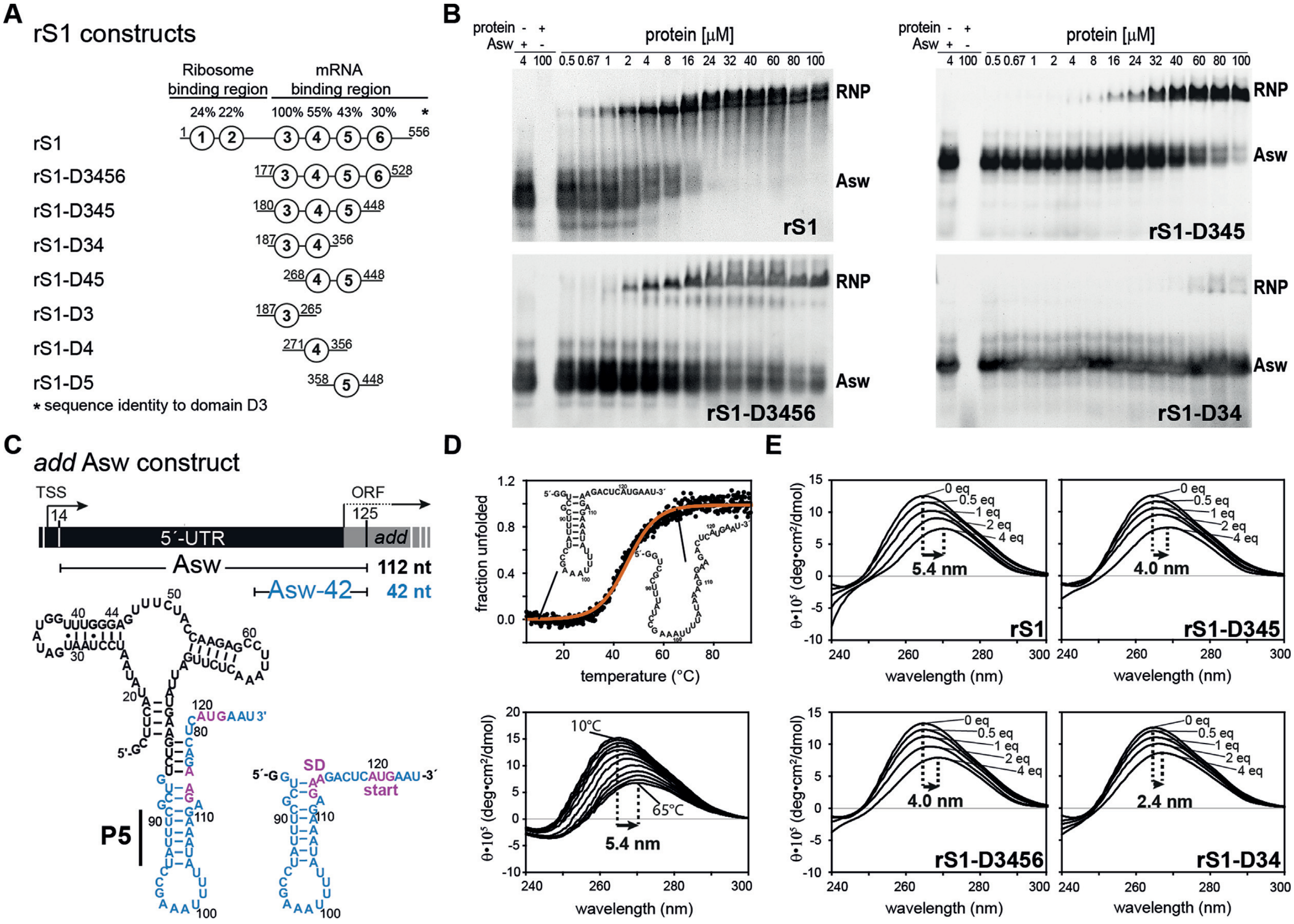


Figure 1. RNA-binding and –chaperone activity of rS1 and its truncated versions. (**A**) Schematic overview of the rS1 constructs used in this study. Fragments are aligned according to their amino acid sequence. Sequence identities compared to domain D3 are displayed on the top. (**B**) Electrophoretic mobility shift assays (EMSAs) of selected rS1 constructs, using full-length *add* Asw (4 μM) as RNA substrate. Increasing protein concentrations are depicted on the top of the gels. The EMSAs were performed on discontinuous gels and only the resolving phases are displayed. Bands reporting on complex formation or free RNA are labeled with RNP or Asw, respectively. Two bands are observed for the free RNA, as the ligand-free riboswitch adopts two conformations (apoA and apoB) (38). In case of the full-length rS1 these states are also resolved for the RNP, since the EMSA was performed at higher powers in order to ensure migration of the complex into the resolving phase of the gel. In addition for the full-length rS1 a high-molecular species was observed (<5%) that did not migrate into the resolving phase of the discontinuous gel (not displayed). (**C**) Schematic overview of *add* Asw constructs. Asw-42 is highlighted in blue and the location of the P5 helix is indicated. The *add* gene from *V. vulnificus* including the transcription start site (TSS) and the open reading frame (ORF) is displayed on the top. (**D**) Thermal unfolding of Asw-42 (10 μM) to visualize the red-shift of CD maxima upon RNA melting. Temperature-dependent CD spectra of Asw-42 are shown. They were used as a benchmark for CD titration experiments reporting on RNA melting. (**E**) rS1-induced unfolding of Asw-42 (10 μM) as measured by CD spectra. All CD titration experiments were carried out at 10°C. Molar ratios of [protein]:[RNA] are depicted in the plots. The maximal shifts are displayed.

served red-shift and ellipticity decrease of the CD spectrum of Asw-42 in the presence of four equivalents of rS1 reports on considerable distortions of base pairing within the RNA und thus a protein-induced unfolding. This indicates purely single-stranded conformation of the RNA in the complex. The red-shift is of the same magnitude as observed during thermal unfolding of the RNA. The deletion of the ribosome-binding domains reduces the chaperone activity by 26%. Further deletion of domain D6 leads to no additional reduction in unwinding capacity. In line with the binding affinity, the two-domain construct rS1-D34 is still able to unwind the RNA but with significantly decreased capability. Therefore, the three-domain protein rS1-D345 contains the full chaperone activity of the mRNA-binding region. For this construct, the chaperone activity was also studied as a function of temperature (Supplementary Figure S2). We find that the stability of the RNA is decreased in the presence of four equivalents of rS1-D345 at all temperatures, where the largest destabilization is observed at 10°C ($\Delta\Delta G \approx -3.7$ kcal/mol). This further corroborates that the rS1 protein actively melts the RNA.

Consistent with reported results, our data show that even upon deletion of domain D6 and the ribosome-binding domains D1 and D2 mRNA-binding of rS1 is essentially retained (23). This observation emphasizes the exceptional importance of domains D3-D5 for the interplay with mRNAs; they constitute the mRNA-interaction core of rS1. Additional deletions within this core region lead to a severe loss of RNA-binding and chaperone activity. We find that deletion from the N-terminal end has a more dramatic effect, since removal of D3 completely diminishes the RNA-binding ability (Supplementary Figure S1). Furthermore, deletion of domain D5 from the core region significantly reduces the RNA-binding and chaperone activity, suggesting that domain D5 is also crucial in establishing the full magnitude of rS1–mRNA interaction. The binding behavior for *V. vulnificus* rS1 is therefore in stark contrast to that of *E. coli* rS1, where activity assays show that domain D4 and D5 are sufficient for mRNA-binding (23).

Delineation of the functional core of rS1

We further investigated the functional importance of rS1-domains for translation by use of *in vitro* transcription-translation assays (Supplementary Figure S3). Therefore, we studied the influence of the full-length rS1 and also its truncated mutants increasingly lacking the mRNA-binding domains on the expression of sGFP. In line with the EMSAs, the *in vitro* assays show that deletion of domain D6 does not significantly reduce the expression levels of sGFP, whereas additional deletion of domain D5 reduces the expression level to ∼77%. Further deletion of D4 reduces the sGFP expression to a basal level. Complete lack of mRNA-binding domains leads to inhibition of translation. These findings suggest that the mRNA-binding domains D3–D5 are necessary for efficient translation. Again a significant effect of D5 deletion on rS1 function is observed. Moreover, in case of *Vibrio* rS1 domains D3 and D4 need to act in tandem, since deletion of D4 diminishes the capability of rS1 to promote translation. In contrast, rS1-D12 inhibits translation probably by preventing ribosome-binding of residual *E. coli* rS1 that is present in the cell-free reaction mix.

NMR spectroscopic investigation of the mRNA-interaction core

The different rS1 constructs were structurally characterized by solution NMR spectroscopy (Figure 2A, see also Supplementary Figure S4), where the overall protein fold was probed by ^{1}H–^{15}N-BEST-TROSY experiments. Such ^{1}H–^{15}N correlation spectra serve as a fingerprint of the protein fold. A large dispersion of amide signals reports on well-folded proteins. Particularly, for proteins with high beta sheet contents the resonances are generally distributed over a wide ppm range

Here, we find that the multi-domain constructs (rS1-D345, rS1-D34 and rS1-D45) dominantly exhibit well-dispersed resonances and the stepwise removal of domains does not substantially perturb the overall fold. The distribution of the amide signals indicates OB fold motifs. However, the D5 containing multi-domain proteins display severe resonance overlap around 8.5 ppm, pointing towards structural disorder within the respective constructs (Supplementary Figure S5). This crowding of signals can be addressed to domain D5. In isolation, this domain exhibits poorly dispersed resonances that are in addition broadened, indicative of a rather dynamic unstructured state. In stark contrast, the spectra of the isolated domains D3 and D4 display the characteristics of OB-fold motifs. Given the high homology of domains D3, D4 and D5 and further the fact that these are forming together the mRNA-binding core (Figures 1 and 2B), this finding was unexpected.

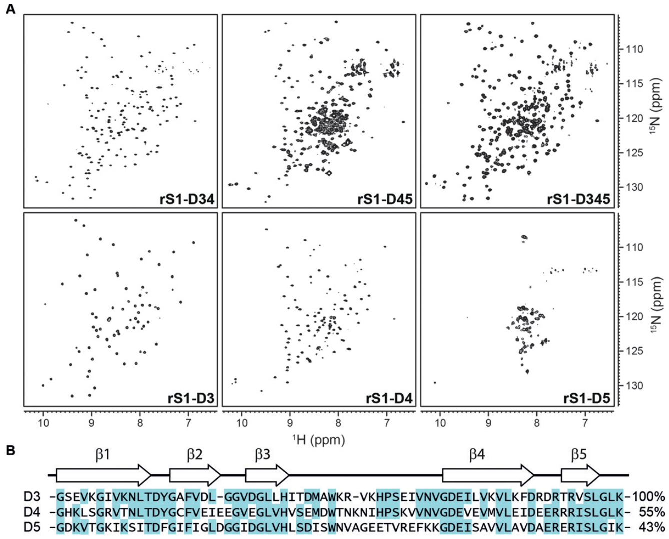


Figure 2. Comparison of the mRNA-binding core domains by NMR and sequence. (**A**) ^{1}H–^{15}N-BEST-TROSY spectra of rS1 constructs as indicated. The spectrum of rS1-D345 was acquired at 35°C and 950 MHz. All other spectra were acquired at 30°C and 600 MHz. (**B**) Domain sequences were aligned using the Multiple Sequence Alignment tool from Clustal Omega (68–70). Sequence identities in comparison to domain D3 are displayed on the right. Secondary structure elements were computed using SWISS-MODEL (71–74) and are displayed on the top. Identical residues are highlighted.

Further, the resonances of D3 and D4 can be unambiguously assigned and the chemical shifts can subsequently be interpreted as structural reporters of the respective protein structure. The superposition of the individual spectra results overall in the spectrum of the two-domain protein rS1-D34.

The mRNA-interaction-core of rS1 is thus organized in two regions. D3 and D4 together are capable to bind and unwind RNA, whereas domain D5 lacks a single well-defined structure. Nevertheless, its presence significantly increases the activity of the preceding domains. We thus decided to study both regions (rS1-D34 and rS1-D5) in separation, in order to gain insight into each functional contribution to the rS1–mRNA-interaction.

rS1-D34 harbors the mRNA-binding platform

First, we elucidated the RNA-binding behavior of rS1-D34 by solution NMR. We acquired ^{1}H–^{15}N-BEST-TROSY experiments of ^{15}N-labeled rS1-D34 in the presence of increasing amounts of RNA. For our interaction studies we used six different RNAs of increasing length originating from the expression platform of the riboswitch (Table 1).

In addition to the structured Asw-42, we also included RNAs containing only the single-stranded 3′-end of the riboswitch (Asw-6 to Asw-14). They were used for probing the minimal nucleotide length within the RNA-binding site of rS1-D34.

We find that rS1-D34 binds the Asw-42 and Asw-14 with an apparent K_D of 12.5 ± 1.5 μM, comparable to the affini-

Table 1. RNA-binding affinities of rS1-D34 for 5′-truncated Asw constructs obtained by NMR

	Sequence (5′-3′)	RNA length (nt)	K_D[a] (μM)	$\Delta G^{\text{free-complex}}$[b] (kcal/mol)
Asw-42	see Figure 1C	42	13 ± 1	−6.9 ± 0.1
Asw-14	112-GAA GA CUC AUG AAU	14	12 ± 1	−6.9 ± 0.1
Asw-12	114-A GA CUC AUG AAU	12	27 ± 3	−6.4 ± 0.1
Asw-10	116-A CUC AUG AAU	10	57 ± 6	−5.9 ± 0.1
Asw-8	118-UC AUG AAU	8	197± 50	−5.2 ± 0.2
Asw-6	120-AUG AAU	6	>1000	>−4.2

[a] The K_D values for the corresponding RNA lengths are listed and they were obtained by fitting largest CSPs with $\Delta\,\delta_{obs} = \Delta\delta_{max}\{([P]_0 + [R]_0 + [K_D]) - \sqrt{([P]_0 + [R]_0 + [K_D])^2 - 4[P]_0[R]_0}\}/2[P]_0$ (see also, Supplementary Figure S6). Errors were obtained from fitting procedure (53).

[b] Free energy of complex formation was calculated as $\Delta G^{free-complex} = -RT \cdot \ln(\frac{1}{K_D})$. Errors were calculated using Gaussian error propagation.

ties of full-length rS1. Furthermore, from our NMR titrations a minimal length for productive binding of 14 nucleotides can be inferred, as we do not observe an increase in binding affinity for the 42 nt long Asw-42 (see also, Supplementary Figure S6). Further, we find that rS1 preferentially binds to the single-stranded TIR of the riboswitch (Supplementary Figure S7).

Upon addition of increasing amounts of Asw-14 to rS1-D34 we observe large chemical shift perturbations (CSPs) affecting residues that are conserved within both domains. Most of the amide resonances display population-averaged resonances upon addition of the RNA (e.g. H230, R341), where the observed peak position shifts on a line between the free and the complexed protein state. Here, the exchange rate is larger than the frequency difference between the two states ($k_{\text{ex}} > \Delta\nu$), reporting on a predominantly fast exchanging system. However, the largest shifts (CSP > 0.2 ppm) are accompanied by severe line-broadening of the respective resonances (e.g. Y205, I220, R252, Y290, L303, V306, see also Figure 3A). For those amino acids, the amide signal is not detectable in the early course of titration, as here the larger frequency difference ranges in the same order of magnitude as the exchange rate ($k_{\text{ex}} \sim \Delta\nu$). This is indicative of an intermediate exchange process between the free and bound state of the protein. Most of the resonances can be recovered at higher amounts of RNA, as increasing ligand concentrations shift the overall exchange rates towards 'fast exchange' conditions ($k_{ex} = k_{on}\,[L] + k_{off}$) (55). From this, we can estimate a lower limit for the exchange rate between the free and Asw-14 bound rS1-D34 of $k_{\text{ex}} \sim 200\ \text{s}^{-1}$. In case of L217, T221, V228, F248, D251, V255, S307 and N352 the resonances remain in intermediate exchange (Figure 3C).

In line with reported results, our data show that the RNA-binding site involves strands β2, β3 and β5 and contains several key residues characterized by aromaticity or positive charges (1,20,56). In particular, Y205, R252, R254 in D3 and Y290, R341 in D4 display large perturbations in the presence of RNA (Figure 3A). The basic side chains enable charge compensation of the sugar-phosphate backbone, facilitating RNP formation directed by electrostatic interactions. The aromatic side chains of H230 and of Y205 and Y290 can engage in interactions with RNA nucleobases. Presumably, they interfere with base pairing interactions in the RNA and hence maintain residual chaperone activity within rS1-D34, as observed in the CD chaperoning assay. Furthermore, many amide resonances (e.g. I220, V306) distributed within the primary contact site experience large CSPs, due to RNA-induced structural rearrangements of the binding interface.

To investigate structural rearrangements and to probe for changes in the dynamical properties within rS1-D34 further, we acquired and analyzed heteronuclear NOEs, longitudinal and transverse ^{15}N relaxation rates for rS1-D34 in its RNA free form and at saturating RNA concentrations (Supplementary Figure S8). In the absence of RNA both domains tumble together, as they exhibit similar global dynamics. The increase in R_2 as well as the decrease in R_1 rates in the presence of Asw-14 are indicative of a slower tumbling system and report on complex formation.

Taken together, in rS1-D34 both domains are structurally linked. They present a continuous surface for RNAs, where both domains equally contribute to RNA-binding, accommodating a stretch of 14 nucleotides. This finding is very conspicuous, as an RNA-binding length of 10–15 nucleotides has been reported for the full-length rS1 from *E. coli* (16,57). In other words, rS1-D34 already covers the size of RNA that the full-length rS1 is known to bind. This strongly supports the idea that domain D5 plays a different role in mRNA-interaction in *V. vulnificus* rS1.

rS1-D5 from *V. vulnificus* populates two distinct states

Domains D5 from *V. vulnificus* and *E. coli* rS1 display a sequence identity of 90% (Supplementary Figure S9). Both domains contain highly conserved amino acids, where altogether 64 out of 71 amino acids are identical and additional five are homologous. Aliprandi *et al.* state that even in isolation, the *E. coli* D5 homolog is well-folded, where the large dispersion of the resonances is in line with the predicted OB-fold motif (56). However, their NMR data evidence for the isolated domain a certain degree of heterogeneity. Further we now show, that the D5 from *V. vulnificus* exhibits structural disorder, not only present in the isolated domain but also observable in the respective multi-domain rS1 constructs. In order to evaluate the origins of the structural heterogeneity of domain D5, we analyzed rS1-D5 from *V. vulnificus* in depth by NMR.

We first acquired temperature-dependent ^{1}H–^{15}N-HSQC spectra of rS1-D5, screening a range from 8°C to 35°C (Figure 4, see also Supplementary Figure S10). Intriguingly, low temperatures revealed that for rS1-D5 two distinct conformational states are detectable. The major state (D5$^{\text{res-}\alpha}$) is

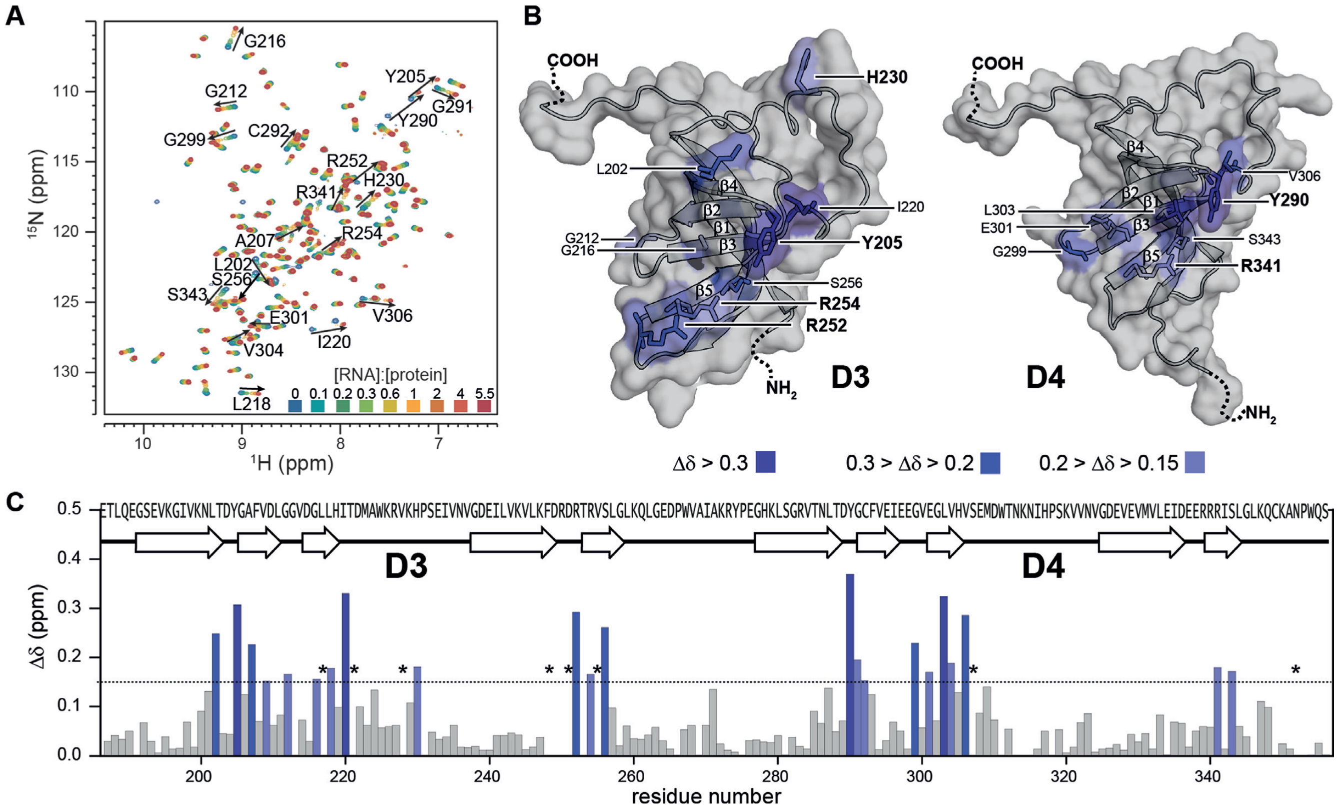


Figure 3. Interaction of rS1-D34 with Asw-14. (**A**) Superposition of ^{1}H-^{15}N-BEST-TROSY acquired at 35°C and 600 MHz. NMR titration was performed with 100 μM ^{15}N-labeled rS1-D34. Unlabeled Asw-14 was added stepwise with ratios ranging from 0 to 5.5 equivalents. The molar ratio ([RNA]:[protein]) is color coded within the superimposed spectra. (**B**) Cartoon representations of D3 and D4 model structures that were generated in SWISS-MODEL (71–74) using the PDB entry 2KHI (18) as template for homology modeling. Observed chemical shift perturbations of binding site reporters are plotted and color coded according to CSP values. The surface is displayed and solvent exposed residues are additionally shown as sticks. Surface binding site reporters are annotated and basic and aromatic residues are highlighted. (**C**) Chemical shift perturbations, calculated as described in the method section, between free and bound rS1-D34 are plotted as function of rS1-D34 sequence. Horizontal line indicates threshold value that was used to identify binding site reporters (53). Asterisks mark residues that are undetectable due to RNA-induced exchange broadening. Missing values represent either prolines or undetectable residues (G213, R250, W311, N313, N315 are exchange broadened at 35°C. D249, T253, K314, L346, A269 and C349 could not be assigned as the amide resonances were absent from ^{1}H–^{15}N-correlation spectra).

populated to 92% (Supplementary Table S3) and displays the poor signal dispersion of a predominantly unstructured state. In contrast, the minor state, D5^{OB}, exhibits well-dispersed resonances (Figure 4A) and is in overall agreement with the reported *E. coli* D5 backbone amide chemical shifts (56). Both D5 states interconvert on the millisecond timescale, causing the observed peak broadening at elevated temperatures.

Conformational switch of sequence element induces domain heterogeneity

We analyzed the D5$^{res\text{-}\alpha}$ state of rS1-D5 by multidimensional NMR. For rS1-D5 altogether 92 resonances in the ^{1}H-^{15}N-HSQC are expected, but at 12°C 154 backbone signals were visible. Using triple resonance experiments, we were able to assign 91 residues to the D5$^{res\text{-}\alpha}$ state (Figure 4B). Under native conditions S373 to L383, D388 to S397 and additionally G385 and R407 were not assignable due to the lack of cross peaks in the 3D experiments. Nevertheless, we were able to transfer the respective assignments from non-native conditions using 8 M urea as denaturing agent. The assignment of the D5^{OB} state could not be assured, since this state is at maximum populated to 8%. The signal pattern, however, indicates OB-fold formation.

The temperature series and urea titration experiments reveal residual secondary structure in the D5$^{res\text{-}\alpha}$ state of rS1-D5 within residues D423 to G432 (Supplementary Figure S10). In order to identify the fold of proteins, chemical shifts of backbone atoms are utilized, as their values strongly correlate with local structure. Particularly, the deviations of the observed chemical shifts from random coil values are a powerful tool for identification of secondary structure from experimental data. We determined the residual structure in the predominantly unstructured D5$^{res\text{-}\alpha}$ state by use of ^{1}H, ^{15}N and ^{13}C secondary chemical shifts (58) and TALOS-N (59) (Figure 5). Our data identifies the existence of an α-helix spanning from A424 to I429. Within the predominantly folded state D5^{OB} the fifth β-strand (E427 to L431) is located exactly in this region. This finding implies that the particular amino acid sequence ‘AERERI’ is capable of switching between two secondary structure elements, a β-strand and an α-helix. Notably, within the OB-fold motif the β5-strand is the closing strand that binds to both, β3 and β4 (Supplementary Figure S11). It potentially locks the

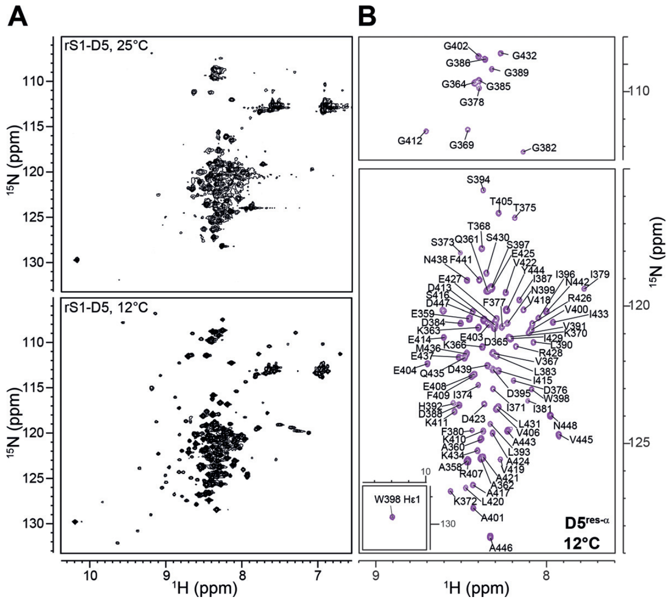


Figure 4. The two conformational states of rS1-D5. (**A**) ^{1}H–^{15}N-HSQC spectra were acquired at the displayed temperatures and at 900 MHz. (**B**) Annotated resonance assignment for D5$^{\text{res-}\alpha}$. The spectrum (acquired at 12°C and 900 MHz) is identical to the spectrum in panel (A) but to exclusively display the predominantly unstructured D5 state, the ^{1}H–^{15}N-HSQC is plotted at higher contour levels.

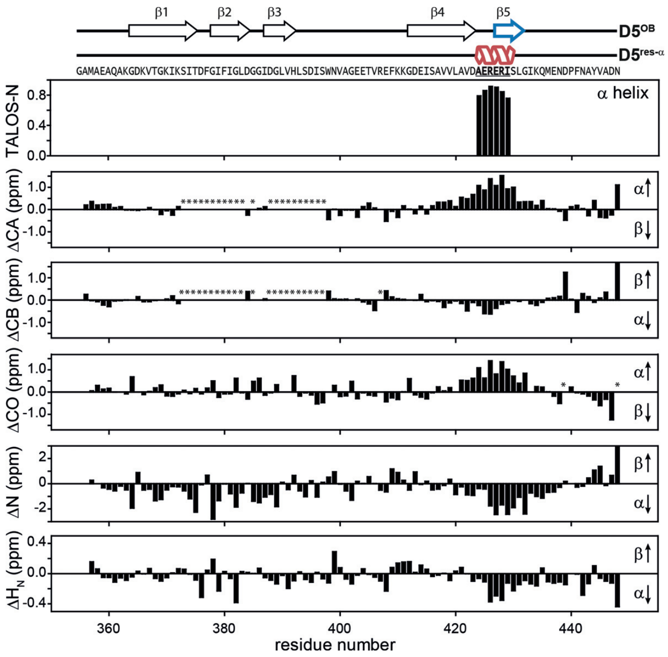


Figure 5. Structural characteristics of the D5$^{\text{res-}\alpha}$ state. The chemical shifts of backbone atoms are sensitive for local structure. Their deviations from random coil values can be used for determination of secondary structure based on experimental data. In the upper plot the fractional secondary structure as calculated from TALOS-N is plotted against the protein sequence (59), where an α-helix is found between A424-I429. The lower plots display ^{1}H, ^{15}N and ^{13}C secondary chemical shifts as a function of residue number. They were calculated as $\Delta = \delta_{obs} - \delta_{rc}$ (58), where δ_{obs} are the observed chemical shifts and δ_{rc} are the random coil chemical shifts. δ_{rc} were generated using the Javascript provided by Alex Maltsev, on the website of the University of Copenhagen (http://www1.bio.ku.dk/english/research/bms/research/sbinlab/groups/mak/randomcoil/script; 18 April 2018). The expected secondary structure element is indicated within the plot with arrows. All secondary chemical shift values indicate α-helical structure of the sequence stretch A421-G432. The largest values coincide with the AERERI sequence, strongly pointing towards α-helical structure of this particular sequence. In case of ΔCB and ΔN the secondary chemical shifts for N448 are truncated and have a value of 1.9 and 5.7, respectively. Asterisks mark not observed resonances. For clarity primary sequence and secondary structure elements of both rS1-D5 states are displayed on the very top of the figure.

anti-parallel β-sheet, following the pattern β3–β2–β1–β4–β5, into a β-barrel structure by forming a parallel β-sheet between strand β3 and β5. The β-strand-to-α-helix transition of this closing strand will inevitably have an impact on the structural integrity of the whole OB-fold motif.

We wondered whether the switching sequence exhibits an intrinsic preference for α-helical structures and thereby promotes the observed order-to-order transition from β-strand to α-helix. Therefore, we analyzed the secondary structure propensities of the switching sequence 'AERERI' that is highly conserved amongst *gammaproteobacteria* (Supplementary Figure S12). For comparison, the corresponding regions of D3 ('RDRTRV') and of D4 ('EERRRI') were also included into the analysis. We performed sequence-based secondary structure predictions and modeled peptide structures for all core domains (Supplementary Figures S13 and S14). Both prediction methods indicate a preference for α-helical structures for the switching regions of domains D3 to D5. Further, we compared the switching sequences of all core domains with protein structures deposited in the PDB. We were able to identify the respective sequences also in other proteins (right hand side of Supplementary Figure S13), where the peptide sequences dominantly adopt α-helical structures, supporting a general preference for this secondary structure element.

Taken together, the switching regions of all core domains display a tendency towards α-helical structures and this propensity seems to be more pronounced within D4 and D5. The transition to such an α-helix would inevitably destroy the integrity of the OB-fold, yet the bistability is only observed for D5.

Conformational heterogeneity of D5 is preserved in multidomain constructs

We determined the populations of the two D5 states within the multi-domain rS1 constructs using different reporter signals, well resolved in both states. We chose G378, G389, G412 and G432 as reporters for structural order, as these glycines are located either directly within β-strands or in the direct vicinity. Their assignments within D5$^{\text{OB}}$ were achieved by homology, as these glycines are conserved also in D3 and D4. Furthermore, W354 and W398 indole resonances were used as indicators for domain-domain interactions. The unstructured states of both indole resonances could be unambiguously assigned, as W398 is the sole tryptophan within D5 (Figure 4) and W354 is located close to the C-terminus of rS1-D34 and hence is unstructured. Both resonances could be additionally correlated with those stemming from the folded state via an N_{zz}-exchange experiment using rS1-D345 (Supplementary Figure S15). Partic-

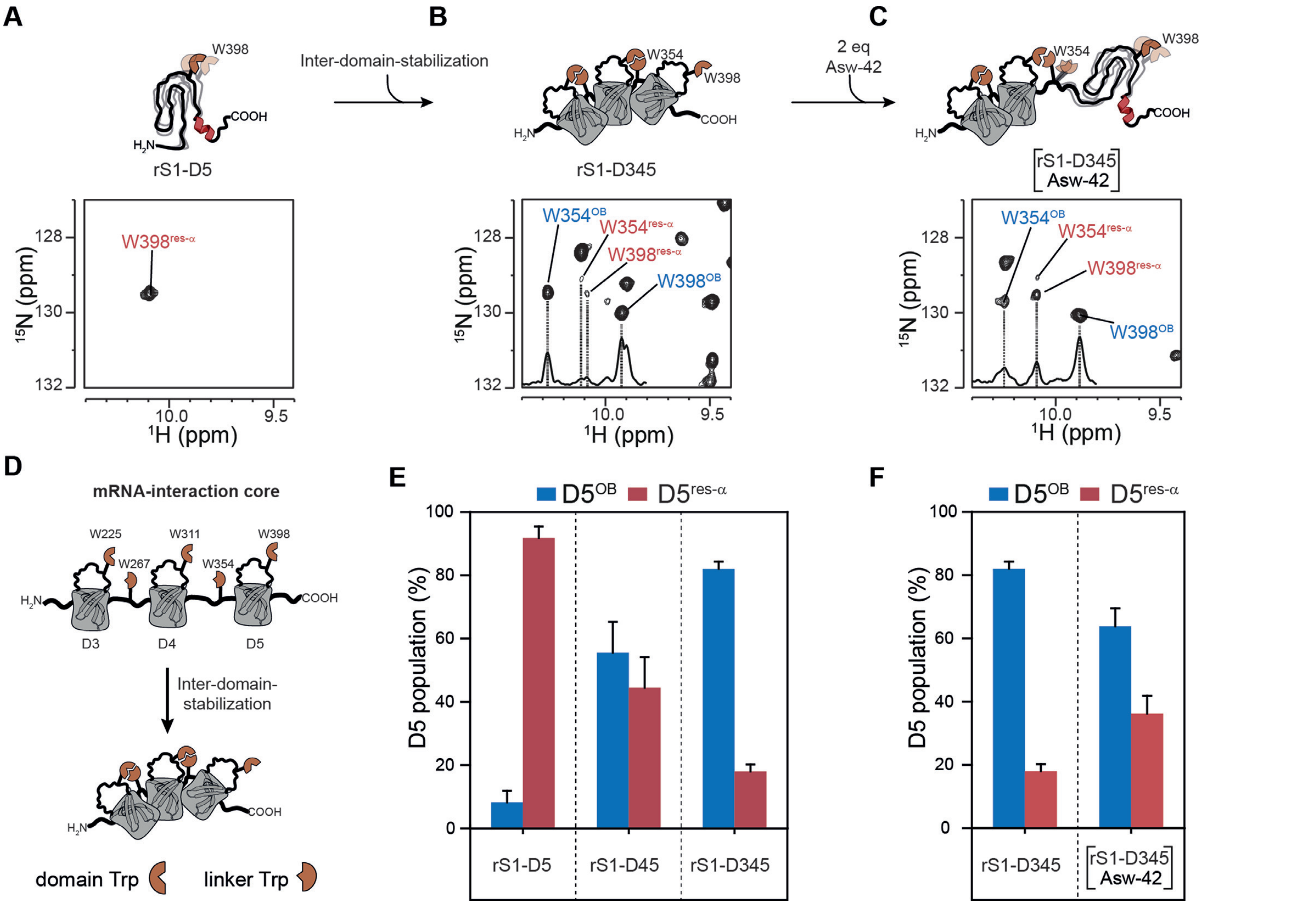


Figure 6. Structural heterogeneity of D5 in presence of preceding domains and RNA. (**A**) The ^{1}H–^{15}N-HSQC of rS1-D5 was acquired at 35°C and 950 MHz and an excerpt of tryptophan indole region is displayed to illustrate the res-α state. (**B**) and (**C**) displays excerpts of tryptophan indole region of rS1-D345 in the absence and presence of two equivalents Asw-42, respectively. For the NMR titration experiments, ^{1}H–^{15}N-BEST-TROSY (NS = 96) spectra were acquired at 35°C and 950 MHz. Unlabeled Asw-42 was added stepwise with ratios ranging from 0 to 2 equivalents. In the presence of RNA a huge intensity decrease of rS1-D345 resonances was observed. Hence, the titration end point was additionally acquired with more number of scans (NS = 256). This spectrum was used for determination of D5-populations and is displayed in panel (C). Positive 1D projections are displayed within the spectra. W-Hε1 resonances of D5$^{\text{res-}\alpha}$are annotated in red and of D5$^{\text{OB}}$ in blue. (**D**) Schematic overview of tryptophan-mediated domain stabilization. (**E**) The populations of the two D5 states are plotted against rS1 constructs. Populations of each reporter signal were determined from their intensities as $P(D5^{OB}) = 100 \cdot \frac{I^{OB}}{I^{OB}+I^{res-\alpha}}$ and *vice versa*. The populations were averaged over all reporter peaks for each rS1-construct. Error bars represent standard deviation of the average. (**F**) D5 population of rS1-D345 is plotted in the absence and presence (2 equivalents) of Asw-42. The rS1-D345-induced melting of Asw-42 was also monitored by NMR and is displayed in Supplementary Figure S7.

ularly, W354 and W398 side chain signals are sensitive reporters of D5 heterogeneity as they exhibit large signal intensities allowing identification of low populations of the D5$^{\text{res-}\alpha}$ state even in the context of multi-domain constructs.

We find that the presence of preceding domains stepwise increases the population of the folded D5$^{\text{OB}}$ state. While in isolation domain D5 populates the D5$^{\text{OB}}$ state only by 8%, the presence of D4 in rS1-D45 leads to a population ratio close to equality (Figure 6E and Supplementary Table S3). Furthermore, the presence of both, D3 and D4 drives domain D5 almost exclusively into the D5$^{\text{OB}}$ state (82%). The stabilization of the D5$^{\text{OB}}$ state within the respective multi-domain constructs is probably induced by the order promoting properties of tryptophan side chain (60) as the indole moiety can participate in hydrophobic interactions and in hydrogen bonding (61).

We classify the five tryptophans of rS1-D345 into two groups. Three 'domain-tryptophans' are conserved within domains D3 (W225), D4 (W311) and D5 (W398) and they are located in the internal loop between the β3-strand and the β4-strand. Two 'linker-tryptophans' (W267 and W354) are conserved within inter-domain linkers between D3–D4 or D4–D5, respectively (Figure 6D). The linker-tryptophans are part of a PWX-motif. This motif is also found in other nucleic acid-binding proteins, in which the PWX-motif packs against the side chain of a phenylalanine that is conserved in the protein core (62). In line, we find evidence that the linker-tryptophans as found here, interact with the preceding domain-tryptophans. In case of the W225–W267 pair we find direct cross peaks in the ^{1}H-^{15}N-NOESY-HSQC of rS1-D34 between the two tryptophan side chains and also between mutual neighboring amino acids (Supplementary Figure S16). Due to signal overlap, such direct NOE evidence for tryptophan-tryptophan interaction could not easily be detected for the larger constructs. However, the interaction can also be detected from chemical shift analysis. Comparison of rS1 constructs of different length reveals that succeeding domains (D4 in rS1-D34 and D5 in rS1-D345) lead to chemical shift perturbations within the internal loop between β3- and β4-strands of the preceding domain (D3 in rS1-D34 and D4 in rS1-D345). Largest CSPs cluster around the indole resonances of domain-tryptophans (Supplementary Figures S17 and S18).

Taken together, domain-tryptophans are engaged in interactions with linker-tryptophans forming hydrophobic clusters and thus guiding and positioning adjacent domains in close proximity. The formation of these two-tryptophan centered hydrophobic clusters are important for the protein-architecture, as they mediate the stabilization of rS1's fold. The domain-domain contacts induce stabilization of the succeeding domain. Domain D3 is the most stable domain within rS1-D345 and we observe that it increases the stability of domain D4. Both domains together stabilize the OB-fold of domain D5. We do not observe any cooperative influence between domains D5 and D6. In line, domain D6 lacks the stability promoting tryptophan residues both in the linker connecting D5 and D6 as well as within the domain D6.

RNA-binding reintroduces conformational heterogeneity of D5

We investigated the properties of the extended RNA-binding core rS1-D345 upon interaction with Asw-42 by NMR titration experiments (Figure 6). We acquired a series of ^{1}H-^{15}N-BEST-TROSY spectra of rS1-D345 by stepwise increasing the RNA molar ratio from 0 to 2 equivalents. Already after the first titration point the average signal intensities decline at least by a factor of two and remain below 50% during the course of titration. We conclude that the interaction between rS1-D345 and Asw-42 is within the intermediate exchange regime. Therefore, the RNA-binding reporters (Y205, H230, Y290, R341) as found in rS1-D34 cannot be recovered even at higher RNA concentrations. Nevertheless, the overall chemical shift perturbations in domains D3 and D4 (Supplementary Figure S19) agree between the extended three-domain (rS1-D345) and the essential two-domain (rS1-D34) RNA-binding core.

The intensity ratios between the free and bound rS1-D345 state show that particularly for D3 and D4 the signal loss is most prominent (Supplementary Figure S19, left panel). With two equivalents of Asw-42 the average signal intensities of these two domains decrease to 37%. D5 resonances display average intensity ratios between the free and bound state >80%. Notably, the remaining D5 resonances do not display significant CSPs (Supplementary Figure S19, right panel). We observe that the presence of RNA unfolds the OB-fold of D5, as increased population of the res-α state is detected at saturating RNA concentrations (Figure 6C). This shift of the population ratio is most likely driven by the intrinsic property of the 'AERERI' sequence to adopt α-helical structures. It is presumably a direct consequence of RNA-binding that induces dissociation of domain D5 by competing out the domain-domain contacts between D4 and D5 leading to the β5-to-α transition as observed for the isolated D5 domain (Figure 6F).

DISCUSSION

We have investigated the structural basis of the RNA-binding and chaperone behavior of the rS1 protein from *Vibrio vulnificus*. The full-length protein composed of six homologues domains is able to melt RNA secondary structures upon binding. The minimal functional core is composed of domains D3 and D4, covering a stretch of 14 single-stranded nucleotides. We dissected the contribution of each of the four C-terminal domains to binding affinity and chaperone activity. While domain D6 is nearly dispensable, deletion of D5 renders both affinity and activity to a minimum level. Domain D3 and domain D4 form the protein's functional core and operate in tandem. In contrast, for the rS1 from *E. coli* it was proposed that individual domains consecutively bind the RNA and mediate melting by capturing single-stranded conformations in multiple small substeps (33).

Although, the rS1 from *V. vulnificus* displays high degree of sequence identity (~86%) to its homolog from *E. coli* we find that the *V. vulnificus* rS1 protein displays remarkable structural differences. For all *E. coli* rS1 RNA-binding domains it is reported that they adopt structured OB-fold motifs. For the single-domains D4, D5 and D6 NMR structures were reported (18). Whereas the structures of D4 and D6 exhibit a tight bundle of conformations, the structure of D5 is rather loosely defined. Furthermore, it was found that within a fragment containing domains D3 to D5, D4 and D5 are in close contact forming a continuous interaction surface, while D3 is mainly dissociated from the other domains (18,56). Further, the minimal fragment for RegB activation is formed by D4 and D5 (23).

In contrast for the *V. vulnificus* rS1 only domains D3 and D4 exhibit a single stable and well-ordered conformation. For domain D5 an unusual structural behavior is found that is key to modulate RNA-binding. This domain can populate two conformational states. In one of the two equilibrium conformations D5 exhibits a well-structured OB-fold. In the second conformation, the local refolding of the sequence 'AERERI' from the OB-fold closing β5-strand that is essentially required to stabilize the OB-fold to α-helix leads to an overall unstructured state with local residual α-helical secondary structure. Therefore, D5 contains a switching sequence that by forming either the mutual exclusive α-helix or β-strand conformation triggers a full change of the domain's structure. This α-helix to β-strand conformational switch is reminiscent to the changes in transformer proteins, but with the exception that those are bistable systems of two fully folded conformations (63). As the switching sequence is highly conserved in γ-proteobacteria it might be a functional feature of rS1 proteins allowing domain 5 to act as dynamic joint between RNA binding domains 3 and 4 and regulatory domain 6. As the population of the two states will be affected by further interactions throughout the domain, the second conformation could be easily overlooked in proteins from organism other than *V. vulnificus* where the equilibrium is shifted more to the D5^{OB} state.

The RNA chaperone activity of *V. vulnificus* rS1 is manifested in its ability to destabilize the secondary structure of RNAs. Whereas in the absence of the protein the ex-

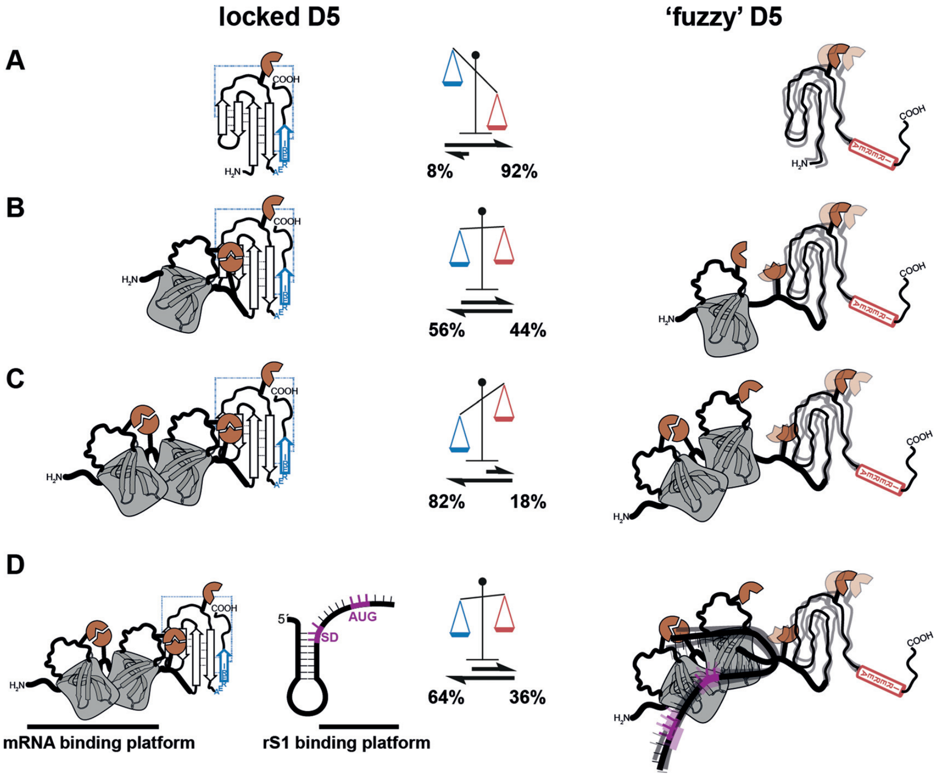


Figure 7. Intrinsic structural heterogeneity of domain D5. Schematic representation of the $D5^{OB}$ and $D5^{res\text{-}\alpha}$ conformational states within the rS1 constructs (**A**) rS1-D5, (**B**) rS1-D45 and (**C**) rS1-D345. The populations are displayed in the scheme. (**D**) D5 populations in the absence and presence of Asw-42.

pression platform module ASW-42 adopts predominantly its hairpin conformation (14% at 35°C), presence of rS1 induces a significant shift in population towards the single-stranded conformation (86% at 35°C) (Supplementary Figure S2). Delineated from the above described CD RNA chaperone assay, this can be expressed in terms of free energy between the hairpin and single-strand conformation that is changed by interaction with rS1 from $\Delta G^{\text{hp-ss}}(35°\text{C}, 0 \text{ eq rS1-D345}) = 1.1$ kcal/mol to $\Delta G^{\text{hp-ss}}(35°\text{C}, 4 \text{ eq rS1-D345}) = -1.0$ kcal/mol. Notably, the change towards the single-stranded conformation is strongly dependent on the stoichiometry (Figure S7) and the temperature (Figure S2). Whereas the supra-stoichiometric amounts of rS1 can lead to shifts in the equilibrium of $\Delta\Delta G(4 \text{ eq}) = -2$ kcal/mol, sub-stoichiometric amounts only contribute to $\Delta\Delta G(0.5 \text{ eq}) = -0.7$ kcal/mol. However, both changes significantly indicate the RNA melting capacity of rS1.

For the rS1 from *V. v*ulnificus the RNA chaperone activity is intimately linked to the structural dynamics of the three domains of the mRNA interaction core. Upon RNA-binding domain D5 is released from its associated state, reflected in a twofold increase of the D5 res-α population from 18% to 36% (Figure 7). The displacement of D5 upon interaction is presumably of allosteric nature as the binding sites for the RNA and D5 onto the mRNA-binding platform D3–D4 only minimally overlap. As the release of D5 increases the structural heterogeneity in the protein, it can compensate the loss of entropy upon complex formation and promote the exothermic ($\Delta G^{\text{free-complex}} < -7$ kcal/mol) binding reaction. Due to the conformational heterogeneity of D5, the resulting RNA-protein complex is a 'fuzzy' complex. The positive influence on affinity towards mRNA and the increased ability to stabilize single-stranded conformations (Supplementary Figure S7) in presence of D5 can be attributed to its dynamic disordered behavior in the res-α state. Such a remote effect of complex stabilization has already been documented for several protein-protein complexes. In these cases, the direct interaction is mediated by structured domains but dynamic disordered parts that do not contribute directly increase the affinity of complex formation manifold. A comparable example is the formation of a dynamic 'fuzzy' complex of the two protein domains KID and KIX, containing flanking disordered regions close to the interaction site (64). The kinase-inducible domain (KID) binds its interaction domain (KIX) by a segment of 29 amino-acids, while the rest of the protein stays disordered and does not take part in the direct interaction. However, deletion of the flanking disordered region decreases the affinity 5-fold (65).

From an overall functional point of view, the protein rS1, as an RNA chaperone, shall be responsible to mediate the translation initiation by melting structured TIRs. Following Sabatier's principle for catalysts (66) it therefore needs to bind, unwind and subsequently release the mRNA chain without falling into an over stabilized conformational state (67). The intricate linkage of nearly equally stable protein and RNA conformational states is exactly what is needed for this function to be fulfilled.

DATA AVAILABILITY

Assigned chemical shifts for rS1-D5 and rS1-D345 are deposited under BMRB accession codes 27489 and 27490, respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Hendrik R.A. Jonker and Christian Richter for help and insightful discussions. We further like to thank Elke Stirnal for excellent technical assistance and Jens Wöhnert for critical reading of the manuscript. Further, we would like to thank the research group of Volker Dötsch for kindly providing the pIVEX-sGFP plasmid.

FUNDING

German funding agency (DFG) in Collaborative Research Center 902: Molecular principles of RNA-based regulation. Work at BMRZ was supported by the state of Hesse and by iNEXT (to S. Sreeramulu). Funding for open access charge: DFG [SFB902].

Conflict of interest statement. None declared.

REFERENCES

1. Cristofari,G. and Darlix,J.-L. (2002) The ubiquitous nature of RNA chaperone proteins. *Prog. Nucleic Acid Res. Mol. Biol.*, **72**, 223–268.
2. Rajkowitsch,L., Chen,D., Stampfl,S., Semrad,K., Waldsich,C., Mayer,O., Jantsch,M.F., Konrat,R., Bläsi,U. and Schroeder,R. (2007) RNA chaperones, RNA annealers and RNA helicases. *RNA Biol.*, **4**, 118–130.
3. Doetsch,M., Fürtig,B., Gstrein,T., Stampfl,S. and Schroeder,R. (2011) The RNA annealing mechanism of the HIV-1 Tat peptide: Conversion of the RNA into an annealing-competent conformation. *Nucleic Acids Res.*, **39**, 4405–4418.
4. Herschlag,D. (1995) RNA chaperones and the RNA folding problem. *J. Biol. Chem.*, **270**, 20871–20874.
5. Woodson,S.A. (2010) Taming free energy landscapes with RNA chaperones. *RNA Biol.*, **7**, 677–686.
6. Tompa,P. and Csermely,P. (2004) The role of structural disorder in the function of RNA and protein chaperones. *FASEB J.*, **18**, 1169–1175.
7. Sengupta,J., Agrawal,R.K. and Frank,J. (2001) Visualization of protein S1 within the 30S ribosomal subunit and its interaction with messenger RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 11991–11996.
8. Byrgazov,K., Grishkovskaya,I., Arenz,S., Coudevylle,N., Temmel,H., Wilson,D.N., Djinovic-Carugo,K. and Moll,I. (2015) Structural basis for the interaction of protein S1 with the Escherichia coli ribosome. *Nucleic Acids Res.*, **43**, 661–673.
9. Wilson,D.N. and Nierhaus,K.H. (2005) Ribosomal proteins in the spotlight. *Crit. Rev. Biochem. Mol. Biol.*, **40**, 243–267.
10. Shine,J. and Dalgarno,L. (1974) The 3′-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. U.S.A.*, **71**, 1342–1346.
11. Yusupova,G.Z., Yusupov,M.M., Cate,J.H.D. and Noller,H.F. (2001) The path of messenger RNA through the ribosome. *Cell*, **106**, 233–241.
12. Kaminishi,T., Wilson,D.N., Takemoto,C., Harms,J.M., Kawazoe,M., Schluenzen,F., Hanawa-Suetsugu,K., Shirouzu,M., Fucini,P. and Yokoyama,S. (2007) A Snapshot of the 30S ribosomal subunit capturing mRNA via the Shine-Dalgarno interaction. *Structure*, **15**, 289–297.
13. Yusupova,G., Jenner,L., Rees,B., Moras,D. and Yusupov,M. (2006) Structural basis for messenger RNA movement on the ribosome. *Nature*, **444**, 391–394.
14. Steitz,J.A. and Jakes,K. (1975) How ribosomes select initiator regions in mRNA: base pair formation between the 3′ terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 4734–4738.
15. Studer,S.M. and Joseph,S. (2006) Unfolding of mRNA secondary structure by the bacterial translation initiation complex. *Mol. Cell*, **22**, 105–115.
16. Subramanian,A.R. (1983) Structure and functions of ribosomal protein S1. *Prog. Nucleic Acid Res. Mol. Biol.*, **28**, 101–142.
17. Kolb,A., Hermoso,J.M., Thomas,J.O. and Szer,W. (1977) Nucleic acid helix-unwinding properties of ribosomal protein S1 and the role of S1 in mRNA binding to ribosomes. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 2379–2383.
18. Salah,P., Bisaglia,M., Aliprandi,P., Uzan,M., Sizun,C. and Bontems,F. (2009) Probing the relationship between Gram-negative and Gram-positive S1 proteins by sequence analysis. *Nucleic Acids Res.*, **37**, 5578–5588.
19. Murzin,A.G. (1993) OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO J.*, **12**, 861–867.
20. Draper,D.E. and Reynaldo,L.P. (1999) RNA binding strategies of ribosomal proteins. *Nucleic Acids Res.*, **27**, 381–388.
21. Loveland,A.B. and Korostelev,A.A. (2017) Structural dynamics of protein S1 on the 70S ribosome visualized by ensemble cryo-EM. *Methods*, **2**, 1–12.
22. Ruckman,J., Ringquist,S., Brody,E. and Gold,L. (1994) The bacteriophage T4 regB ribonuclease. Stimulation of the purified enzyme by ribosomal protein S1. *J. Biol. Chem.*, **269**, 26655–26662.
23. Bisaglia,M., Laalami,S., Uzan,M. and Bontems,F. (2003) Activation of the RegB endoribonuclease by the S1 ribosomal protein is due to cooperation between the S1 four C-terminal modules in a substrate-dependant manner. *J. Biol. Chem.*, **278**, 15261–15271.
24. Wahba,A.J., Miller,M.J., Niveleau,A., Landers,T.A., Carmichael,G.G., Weber,K., Hawley,D.A. and Slobin,L.I. (1974) Subunit I of G beta replicase and 30 S ribosomal protein S1 of Escherichia coli. Evidence for the identity of the two proteins. *J. Biol. Chem.*, **249**, 3314–3316.
25. Takeshita,D., Yamashita,S. and Tomita,K. (2014) Molecular insights into replication initiation by Qβ replicase using ribosomal protein S1. *Nucleic Acids Res.*, **42**, 10809–10822.
26. de Smit,M.H. and van Duin,J. (1994) Control of translation by mRNA secondary structure in Escherichia coli. A quantitative analysis of literature data. *J. Mol. Biol.*, **244**, 144–150.
27. de Smit,M.H. and van Duin,J. (1994) Translational initiation on structured messengers. Another role for the Shine-Dalgarno interaction. *J. Mol. Biol.*, **235**, 173–184.
28. Andreeva,I., Belardinelli,R. and Rodnina,M. V. (2018) Translation initiation in bacterial polysomes through ribosome loading on a standby site on a highly translated mRNA. *Proc. Natl. Acad. Sci. U.S.A.*, **0**, 201718029.
29. Duval,M., Simonetti,A., Caldelari,I. and Marzi,S. (2015) Multiple ways to regulate translation initiation in bacteria: mechanisms, regulatory circuits, dynamics. *Biochimie.*, **114**, 18–29.
30. Serganov,A. and Nudler,E. (2013) A decade of riboswitches. *Cell*, **152**, 17–24.
31. Kortmann,J. and Narberhaus,F. (2012) Bacterial RNA thermometers: molecular zippers and switches. *Nat. Rev. Microbiol.*, **10**, 255–265.
32. Nahvi,A., Sudarsan,N., Ebert,M.S., Zou,X., Brown,K.L. and Breaker,R.R. (2002) Genetic control by a metabolite binding mRNA. *Chem. Biol.*, **9**, 1043–1049.
33. Qu,X., Lancaster,L., Noller,H.F., Bustamante,C. and Tinoco,I. (2012) Ribosomal protein S1 unwinds double-stranded RNA in multiple steps. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 14458–14463.
34. Duval,M., Korepanov,A., Fuchsbauer,O., Fechter,P., Haller,A., Fabbretti,A., Choulier,L., Micura,R., Klaholz,B.P., Romby,P. *et al.* (2013) Escherichia coli ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation. *PLoS Biol.*, **11**, 12–14.
35. Mandal,M. and Breaker,R.R. (2004) Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nat. Struct. Mol. Biol.*, **11**, 29–35.

36. Rieder,R., Lang,K., Graber,D. and Micura,R. (2007) Ligand-induced folding of the adenosine deaminase A-riboswitch and implications on riboswitch translational control. *ChemBioChem.*, **8**, 896–902.
37. Neupane,K., Yu,H., Foster,D.A.N., Wang,F. and Woodside,M.T. (2011) Single-molecule force spectroscopy of the add adenine riboswitch relates folding to regulatory mechanism. *Nucleic Acids Res.*, **39**, 7677–7687.
38. Reining,A., Nozinovic,S., Schlepckow,K., Buhr,F., Fürtig,B. and Schwalbe,H. (2013) Three-state mechanism couples ligand and temperature sensing in riboswitches. *Nature*, **499**, 355–359.
39. Warhaut,S., Mertinkus,K.R., Hollthaler,P., Furtig,B., Heilemann,M., Hengesbach,M. and Schwalbe,H. (2017) Ligand-modulated folding of the full-length adenine riboswitch probed by NMR and single-molecule FRET spectroscopy. *Nucleic Acids Res.*, **45**, 5512–5522.
40. Lemay,J.F., Desnoyers,G., Blouin,S., Heppell,B., Bastet,L., St-Pierre,P., Massé,E. and Lafontaine,D.A. (2011) Comparative study between transcriptionally- and translationally-acting adenine riboswitches reveals key differences in riboswitch regulatory mechanisms. *PLoS Genet.*, **7**, e1001278.
41. Tian,S., Kladwang,W. and Das,R. (2018) Allosteric mechanism of the V. vulnificus adenine riboswitch resolved by four-dimensional chemical mapping. *Elife*, **7**, 1–36.
42. Helmling,C., Keyhani,S., Sochor,F., Fürtig,B., Hengesbach,M. and Schwalbe,H. (2015) Rapid NMR screening of RNA secondary structure and binding. *J. Biomol. NMR*, **63**, 67–76.
43. Kai,L., Dötsch,V., Kaldenhoff,R. and Bernhard,F. (2013) Artificial environments for the Co-translational stabilization of cell-free expressed proteins. *PLoS One*, **8**, e56637.
44. Mori,S., Abeygunawardana,C., Johnson,M.O. and van Zijl,P.C. (1995) Improved sensitivity of HSQC spectra of exchanging protons at short interscan delays using a new fast HSQC (FHSQC) detection scheme that avoids water saturation. *J. Magn. Reson. B*, **108**, 94–98.
45. Favier,A. and Brutscher,B. (2011) Recovering lost magnetization: polarization enhancement in biomolecular NMR. *J. Biomol. NMR*, **49**, 9–15.
46. Schulte-Herbrüggen,T. and Sørensen,O.W. (2000) Clean TROSY: compensation for Relaxation-Induced Artifacts. *J. Magn. Reson.*, **144**, 123–128.
47. Solyom,Z., Schwarten,M., Geist,L., Konrat,R., Willbold,D. and Brutscher,B. (2013) BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. *J. Biomol. NMR*, **55**, 311–321.
48. Bodenhausen,G. and Ruben,D.J. (1980) Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chem. Phys. Lett.*, **69**, 185–189.
49. Piotto,M., Saudek,V. and Sklenár,V. (1992) Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR*, **2**, 661–665.
50. Sklenar,V., Piotto,M., Leppik,R. and Saudek,V. (1993) Gradient-tailored water suppression for 1H-15N HSQC experiments optimized to retain full sensitivity. *J. Magn. Reson. Ser. A*, **102**, 241–245.
51. Lescop,E., Schanda,P. and Brutscher,B. (2007) A set of BEST triple-resonance experiments for time-optimized protein resonance assignment. *J. Magn. Reson.*, **187**, 163–169.
52. Lee,W., Tonelli,M. and Markley,J.L. (2015) NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics*, **31**, 1325–1327.
53. Williamson,M.P. (2013) Using chemical shift perturbation to characterise ligand binding. *Prog. Nucl. Magn. Reson. Spectrosc.*, **73**, 1–16.
54. Hajnsdorf,E. and Boni,I.V. (2012) Multiple activities of RNA-binding proteins S1 and Hfq. *Biochimie.*, **94**, 1544–1553.
55. Kovrigin,E.L. (2012) NMR line shapes and multi-state binding equilibria. *J. Biomol. NMR*, **53**, 257–270.
56. Aliprandi,P., Sizun,C., Perez,J., Mareuil,F., Caputo,S., Leroy,J.L., Odaert,B., Laalami,S., Uzan,M. and Bontems,F. (2008) S1 ribosomal protein functions in translation initiation and ribonuclease RegB activation are mediated by similar RNA-protein interactions: an NMR and SAXS analysis. *J. Biol. Chem.*, **283**, 13289–13301.
57. Lipecky,R., Kohlschein,J. and Gassen,H.G. (1977) Complex formation between ribosomal protein S1, oligo-and polynucleotides: chain length dependence and base specificity. *Nucleic Acids Res.*, **4**, 3627–3642.
58. Mielke,S.P. and Krishnan,V.V. (2009) Characterization of protein secondary structure from NMR chemical shifts. *Prog. Nucl. Magn. Reson. Spectrosc.*, **54**, 141–165.
59. Shen,Y. and Bax,A. (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J. Biomol. NMR*, **56**, 227–241.
60. Uversky,V.N. (2013) Unusual biophysics of intrinsically disordered proteins. *Biochim. Biophys. Acta - Proteins Proteomics*, **1834**, 932–951.
61. Samanta,U., Pal,D. and Chakrabarti,P. (2000) Environment of tryptophan side chains in proteins. *Proteins Struct. Funct. Genet.*, **38**, 288–300.
62. Szymczyna,B.R., Bowman,J., McCracken,S., Pineda-Lucena,A., Lu,Y., Cox,B., Lambermon,M., Graveley,B.R., Arrowsmith,C.H. and Blencowe,B.J. (2003) Structure and function of the PWI motif: a novel nucleic acid-binding domain that facilitates pre-MRNA processing. *Genes Dev.*, **17**, 461–475.
63. Burmann,B.M., Knauer,S.H., Sevostyanova,A., Schweimer,K., Mooney,R.A., Landick,R., Artsimovitch,I. and Rösch,P. (2012) An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell*, **150**, 291–303.
64. Tompa,P. and Fuxreiter,M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.*, **33**, 2–8.
65. Zor,T., Mayr,B.M., Jane Dyson,H., Montminy,M.R. and Wright,P.E. (2002) Roles of phosphorylation and helix propensity in the binding of the KIX domain of CREB-binding protein by constitutive (c-Myb) and inducible (CREB) activators. *J. Biol. Chem.*, **277**, 42241–42248.
66. Roduner,E. (2014) Understanding catalysis. *Chem. Soc. Rev.*, **43**, 8226–8239.
67. Espah Borujeni,A., Channarasappa,A.S. and Salis,H.M. (2014) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res.*, **42**, 2646–2659.
68. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
69. McWilliam,H., Li,W., Uludag,M., Squizzato,S., Park,Y.M., Buso,N., Cowley,A.P. and Lopez,R. (2013) Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res.*, **41**, W597–W600.
70. Li,W., Cowley,A., Uludag,M., Gur,T., McWilliam,H., Squizzato,S., Park,Y.M., Buso,N. and Lopez,R. (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.*, **43**, W580–W584.
71. Biasini,M., Bienert,S., Waterhouse,A., Arnold,K., Studer,G., Schmidt,T., Kiefer,F., Cassarino,T.G., Bertoni,M., Bordoli,L. *et al.* (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.*, **42**, W252–W258.
72. Benkert,P., Biasini,M. and Schwede,T. (2011) Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, **27**, 343–350.
73. Guex,N., Peitsch,M.C. and Schwede,T. (2009) Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. *Electrophoresis*, **30**, S162–S173.
74. Bienert,S., Waterhouse,A., de Beer,T.A.P., Tauriello,G., Studer,G., Bordoli,L. and Schwede,T. (2017) The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Res.*, **45**, D313–D319.

5.4 **Research article:** Switching at the ribosome: riboswitches need rProteins as modulators to regulate translation

Vanessa de Jesus, Nusrat S. Qureshi, Sven Warhaut, Jasleen Kaur Bains, Marina S. Dietz, Mike Heilemann, Harald Schwalbe and Boris Fürtig; *Nature Communications* **12**, 4723 (2021)

In this article, the dynamic network between the translational adenine-sensing riboswitch of *V. vulnificus*, the ribosomal protein S1 and the ribosome was investigated by NMR spectroscopy as well as *in vivo* and *in vitro* assays. It was shown that the addition of the ligand adenine is not sufficient to unwind the secondary structure of the riboswitch to its translational on-state. The chaperone activity of rS1, which is present in the 30S ribosomal subunit, allows the melting of the riboswitch expression platform and releases the ribosome binding site for translation initiation.

V. de Jesus performed the *in vivo* assays, electromobility shift assays, microscale thermophoresis as well as NMR experiments and wrote the manuscript. The author of this thesis designed and prepared the adenine-sensing riboswitch constructs used in the cell-free expression system. The rS1 protein was expressed and purified. The cell-free system was established with a riboswitch-modulated sGFP sequence to test the influence of adenine. The data were analyzed and the corresponding paragraph in the manuscript was prepared.

https://doi.org/10.1038/s41467-021-25024-5

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-25024-5 OPEN

Check for updates

Switching at the ribosome: riboswitches need rProteins as modulators to regulate translation

Vanessa de Jesus[1], Nusrat S. Qureshi[1], Sven Warhaut[1], Jasleen K. Bains[1], Marina S. Dietz[2], Mike Heilemann[2], Harald Schwalbe[1] & Boris Fürtig[1]✉

Translational riboswitches are cis-acting RNA regulators that modulate the expression of genes during translation initiation. Their mechanism is considered as an RNA-only gene-regulatory system inducing a ligand-dependent shift of the population of functional ON- and OFF-states. The interaction of riboswitches with the translation machinery remained unexplored. For the adenine-sensing riboswitch from *Vibrio vulnificus* we show that ligand binding alone is not sufficient for switching to a translational ON-state but the interaction of the riboswitch with the 30S ribosome is indispensable. Only the synergy of binding of adenine and of 30S ribosome, in particular protein rS1, induces complete opening of the translation initiation region. Our investigation thus unravels the intricate dynamic network involving RNA regulator, ligand inducer and ribosome protein modulator during translation initiation.

[1] Institute of Organic Chemistry and Chemical Biology - Center for Biomolecular Magnetic Resonance (BMRZ), Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue Straße 7, Frankfurt am Main, Germany. [2] Institute of Physical and Theoretical Chemistry, Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue Straße 7, Frankfurt am Main, Germany. ✉email: fuertig@nmr.uni-frankfurt.de

During translation initiation, messenger RNA (mRNA) and the 30S ribosome form a complex that is stabilized by interactions between mRNA, ribosomal RNA (rRNA) and ribosomal proteins (rproteins). The mRNA contains a ribosome binding site (RBS) of approx. 30 nucleotides (nt) around the start codon that includes the evolutionary conserved Shine-Dalgarno (SD) sequence 5'-GGAGGA-3'[1,2]. Translation initiation requires the coupling of the opening of mRNA secondary structure in the translation initiation region and binding of the 30S ribosome[3]. The opening of mRNA secondary structure is transient and often too short for recruiting the 30S ribosome[4]. To circumvent this short opening, the 30S is already bound to an mRNA standby site[4–7] and shifts into place to form a specific complex with the SD-mRNA and the anti-SD-rRNA interaction only if the SD-mRNA site becomes accessible.

Sensing of the accessibility of the SD sequence thus is one mechanism to regulate translation initiation[8]. In fact, translational riboswitches, cis-acting regulation elements that bind to a specific inducer ligand found in bacteria, modulate this accessibility[9,10]. Up to now, it is assumed that binding of an inducer ligand is sufficient for translational riboswitches to interfere with the dynamic shift between non-specific standby binding and specific binding of the 30S[9,11]. The adenine-sensing *add* riboswitch (ASW) from the human pathogenic bacterium *Vibrio vulnificus* regulates translation in an inducer ligand-dependent (characterized by the dissociation constant K_D) and temperature-compensated (characterized by the equilibrium constant K_pre) manner and functions over a broad temperature range[12,13]. In the absence of adenine, ASW populates two secondary structures (apoB and apoA) (Fig. 1, Supplementary Fig. 1). Both, apoB and apoA are functional OFF-states as they sequester the SD in stable base pair interactions that preclude translation initiation. The *add* riboswitch can bind adenine post-transcriptionally and operates under thermodynamic control, as shown by in vitro and in vivo studies[14]. Gene regulation of the *add* riboswitch is independent from the coupling of transcription and translation. It is clear that the kinetics of ligand binding and structural switching between apoA and apoB has to occur on a timescale faster than translation initiation[12,15]. It was shown that the rate limiting step in regulation is dependent on the RNA refolding rate, which is independent from adenine concentration[12]. Little is known about the interaction of translational riboswitches and the ribosome, since X-ray crystallography or cryo-EM studies are difficult as the regulatory relevant interactions are transient. However, NMR studies capable of detecting such transient interactions are challenging because of the size of the investigated system of more than 800 kDa, which contains the riboswitch mRNA and the 30S ribosome. Here, we succeed in applying solution NMR spectroscopy of various reconstituted adenine-sensing riboswitch-30S ribosome complexes to investigate both the influence of the inducer ligand adenine and the ribosomal modulator protein rS1 and unravel a synergistic regulatory switch to a functional ON-state.

Results and discussion

We investigated the riboswitch (ASW$^{14-140}$) containing 18 nucleotides within the coding region of the mRNA as a standby site for ribosome binding (Fig. 1a). The inducer ligand adenine binds to ASW with a K_D of 0.6 μM at 25 °C[12]. The helix P4 of ASW$^{14-140}$ was only partially opened up to 20% at 25 °C, and complete opening was never reached even at saturating concentrations of the inducer ligand (Fig. 1b, Supplementary Fig. 1).

The footprint of bound 30S ribosome comprises 30 nucleotides that start 15 nt upstream of the start codon (position −15)[16]. The adenine-induced opening of RNA structure in ASW$^{14-140}$, however, only affected nucleotides up to position −9, nucleotide A111 (Supplementary Fig. 1). This ligand-induced partial opening of P4 in ASW$^{14-140}$ was not sufficient to initiate translation. In other words, the interaction of adenine alone with ASW$^{14-140}$ was insufficient to enable the complete accommodation of the mRNA into the ribosome, as the neighboring helix P5 ultimately impeded binding. Therefore, we here determined the additional factors

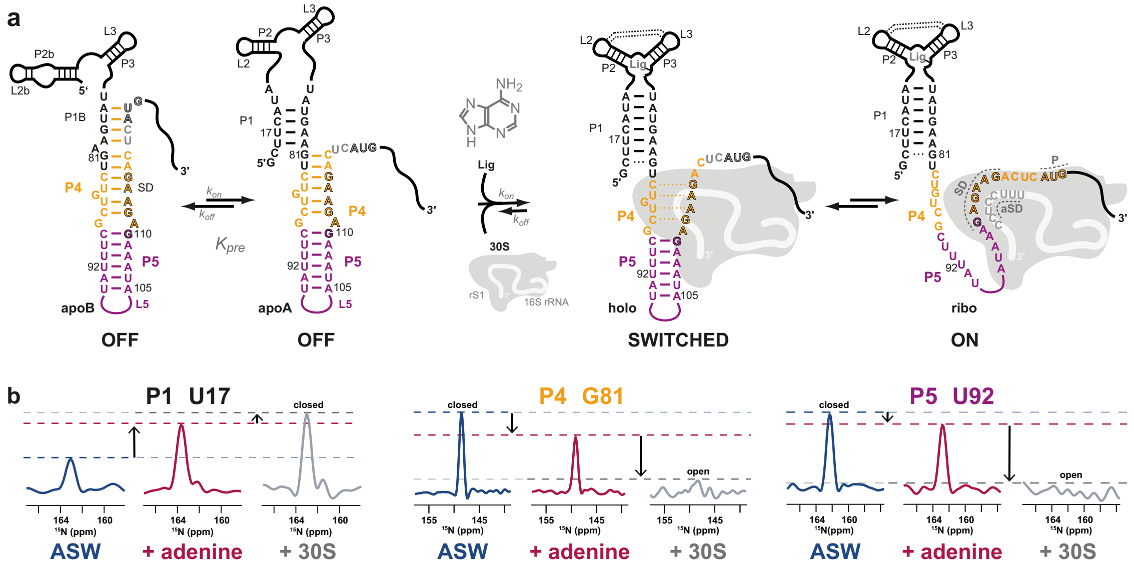


Fig. 1 Adenine-sensing riboswitch (ASW) regulating translation initiation in the context of adenine and 30S ribosome binding. a The two functional OFF-conformations apoB and apoA are in a temperature-dependent equilibrium linked through K_pre. Adenine binding to the aptamer region leads to a compact structure with tertiary interactions such as the kissing loop interaction between L2 and L3. 30S ribosome binding to the standby site 3′-terminal to the AUG start codon of the mRNA results in to melting of the Shine-Dalgarno (SD) sequence in helix P4 of the riboswitch (switched). Further, a stable SD-aSD interaction forms and the mRNA is incorporated into the mRNA tunnel when both P4 and P5 are opened. **b** 1D projections of the ^{15}N dimension of 2D [^{1}H-^{15}N]-BEST-TROSY experiments at 25 °C shown for peaks in helices P1, P4 and P5 for ASW$^{14-140}$ without (apo, blue), with 1.4 eq. adenine (holo, magenta) and in complex with the ribosome (30S + holo, gray).

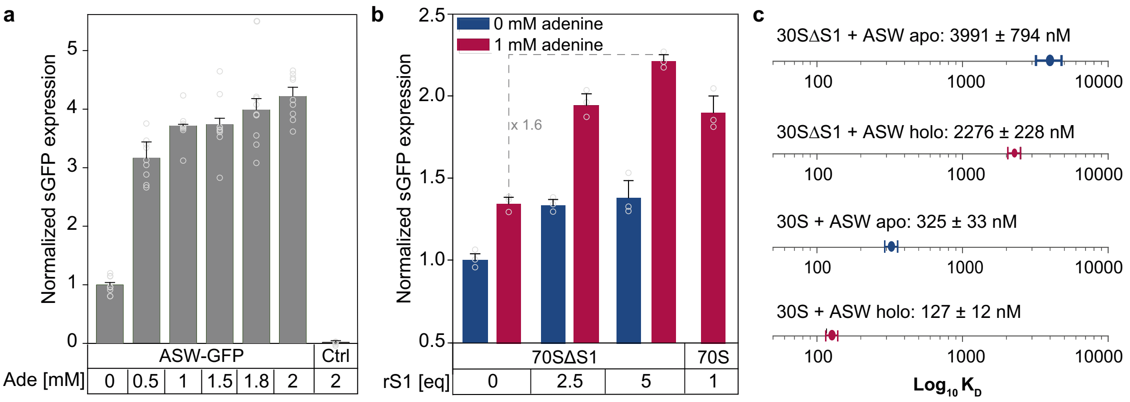


Fig. 2 Biochemical probing of ASW function in vivo and in vitro. a Adenine-dependent riboswitch function in vivo. Translation efficiency was monitored by quantifying sGFP expression levels under the control of the riboswitch. For error estimation, three biological independent samples were measured in triplicates (mean ± s.d.). **b** Riboswitch function in a fully reconstituted coupled transcription-translation assay in dependence of adenine and rS1. For error estimation, one biological sample was measured in triplicates (mean ± s.d.). **c** Binding affinities of mRNA-ribosome-complexes determined by microscale thermophoresis in the absence and presence of rS1 and adenine. Source data are provided as a Source data file.

that interact with the riboswitch to enable full accommodation in the ribosomal tunnel.

rS1 modulates riboswitch activity. When inserted prior to the shifted green fluorescent protein (sGFP) gene, the riboswitch retained its regulatory function in vivo. Expression of sGFP under the control of the ASW led to an adenine-dependent increase of translated sGFP levels (Fig. 2a). Adenine concentrations in the range between 0.5 mM and 2 mM increased sGFP expression by ~3- to ~4-fold, respectively.

Since, the inducer ligand adenine alone could not account for a structural transition to the functional ON-state, the results of the in vivo experiments suggested that there must be additional factors of the bacterial translation machinery that positively interfered with the riboswitch. Several ribosomal proteins exhibit ATP-independent RNA chaperone activity that accelerates opening of RNA secondary structure[17,18]. Chaperone activity has been reported for rproteins rS12[17,19], and in particular rS1[20]. rS1 is essential for translation initiation in Gram-negative bacteria[21]. Bound to the 30S ribosome[22–25], it recruits mRNAs for translation initiation by binding to AU-rich sequences. We thus investigated the functional role of rS1 in assisting riboswitch-mediated gene regulation. rS1 concentrations cannot be easily manipulated in vivo[26], as knock-out of rS1 is lethal[27]. Thus, we investigated its role in a reconstituted coupled transcription-translation assay (Fig. 2b) to determine rS1's function in the riboswitch-mediated regulation of gene expression. In line with the in vivo experiments, addition of adenine resulted in an increase of sGFP expression in the presence of rS1. However, the regulation efficiency diminished to basal levels in the absence of rS1, decreasing by a factor of 1.6 upon removal of rS1. The switching behavior could be restored upon addition of rS1.

Native composite gel electrophoreses revealed complex formation of ASW$^{14–140}$ with 30 S ribosomes (Supplementary Fig. 2). As the cognate ribosome from *V. vulnificus* is not accessible for practical and biosafety reasons, the interaction is studied here with the homologous *E. coli* ribosome. The 3′-end of the 16S rRNA in *E. coli* which harbors the aSD sequence is nearly identical (identity = 98%) to that of *V. vulnificus* (Supplementary Fig. 11) supporting the validity of using *E. coli* ribosomes to study the molecular basis of translation initiation for the *add* riboswitch from *V. vulnificus*. Binding of ASW$^{14–140}$ to the 30S ribosome was dependent on the presence of rS1 as shown by microscale thermophoresis (MST) (Fig. 2c, Supplementary Fig. 3). The binding constants of apo and holo ASW$^{14–140}$ to 30SΔS1 ribosomes were K_D = 4000 ± 800 nM and 2200 ± 200 nM, respectively. By stark contrast, binding of ASW$^{14–140}$ to 30S ribosomes was substantially increased (K_D = 320 nM (apo) and 130 nM (holo)). The affinity increase was nearly four-fold larger in the holo-state hinting at synergistic structural effects of the rS1 protein and the inducer ligand adenine (Supplementary Fig. 4). Notably, the affinity of the riboswitch was further increased to $K_D^{30S\Delta S1+vvrS1}$ = 60 nM (holo) when the rS1 protein from *E. coli* was exchanged to rS1 from *V. vulnificus* (Supplementary Fig. 3b). To show that rS1 acts within the 30S complex, the K_D of *V. vulnificus* rS1 towards 30S was determined to be 170 nM and further NMR experiments reported on a tight binding to the 30S ribosome (Supplementary Fig. 5). The observation that adenine binds more weakly to the ASW than the ribosome hints at a structural and functional synergistic effect of the inducer ligand and the ribosome.

30S ribosome-bound rS1 opens expression platform in holo state. To delineate the structural basis for the increased affinity, we characterized conformational changes induced by stoichiometric amounts of rS1 and the 30S ribosome using solution NMR spectroscopy to map the base pairing interactions within ASW$^{14–140}$ in eight different experimental setups (Fig. 3a, Supplementary Fig. 6). In the absence of inducer ligand adenine (apo state), the conformations apoB and apoA were found to be in equilibrium. In both conformations, helices P4 and P5 were closed. This structural behavior was generally maintained irrespective of the presence of rS1 alone or in the presence of the ribosomal particle without rS1 (30SΔS1). In a titration with rS1, minor effects towards a destabilization of base pairs in the helices P4 and P5 in the translation initiation region were detected at saturating levels of rS1 (Supplementary Fig. 7). Although, in absence of inducer ligand the expression platform was unresponsive against rS1 alone and 30SΔS1, it was slightly responsive against the full ribosomal particle, however the closed conformational state was still the dominant conformation (Fig. 3b). In the presence of adenine (holo state), the equilibrium was shifted away from apoB towards the holo conformation with partial opening of helix P4 and closed helix P5. By stark contrast to the apo state, the holo state of the riboswitch was much more susceptible to conformational changes upon interaction with rS1 and 30S. At 1:1 stoichiometry, addition of rS1 led to an opening of P4 and P5 by 50%. However, addition of 30S containing rS1 led

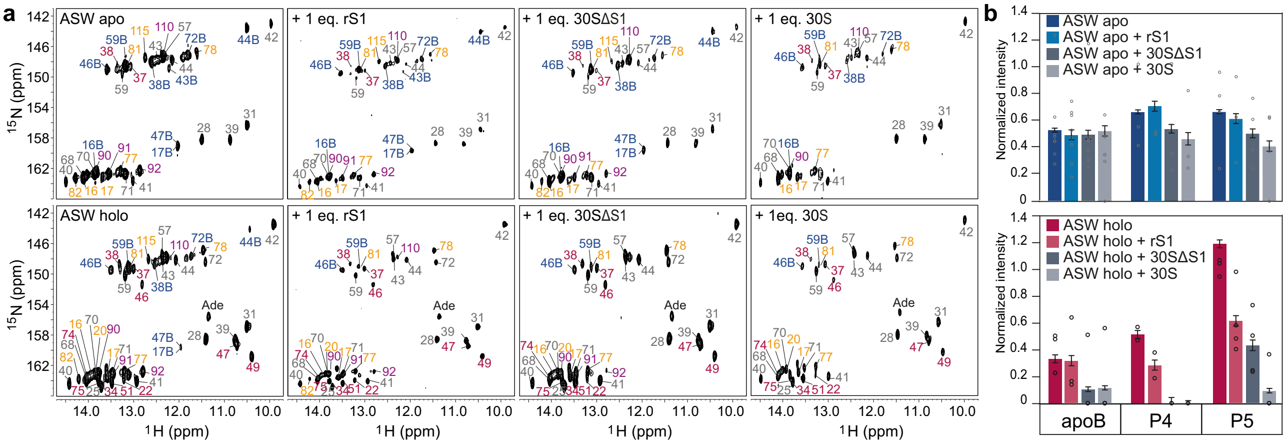


Fig. 3 NMR spectroscopic investigations of mRNA-ribosome complexes. a [^{1}H-^{15}N]-BEST-TROSY experiments of ASW, ASW-rS1, ASW-30SΔS1 and ASW-30S in absence and presence of 1.4 eq. adenine. Peaks are color coded according to Fig. 1. **b** Normalized intensities of peaks according to the secondary structure elements. Peaks were normalized to helix P3 (see Supplementary Fig. 6, Source data are provided as a Source data file). Error bars represent the standard deviation of the average weighted by the signal-to-noise ratio of the individual peaks.

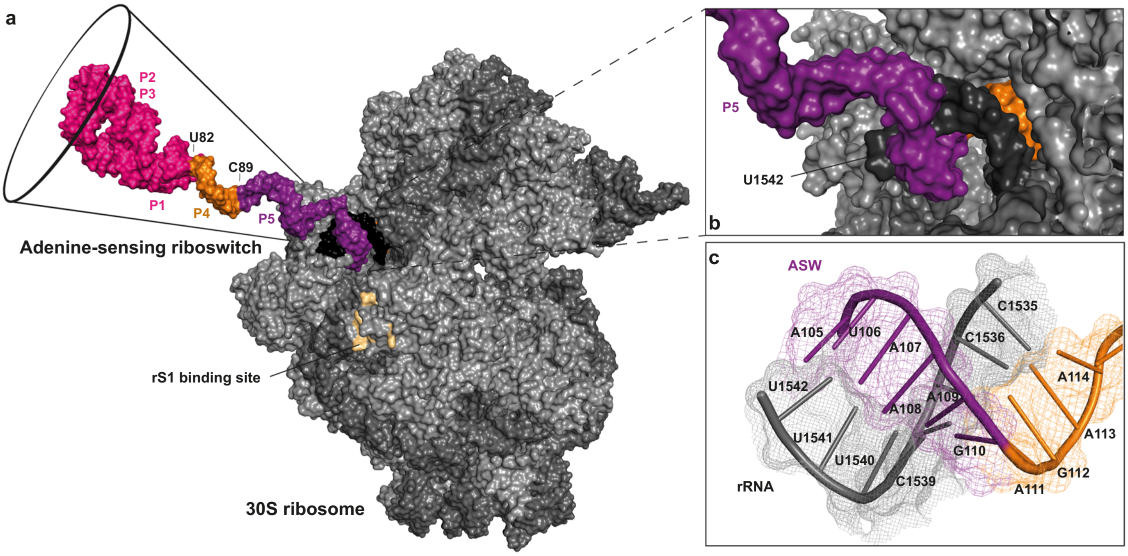


Fig. 4 Structural model of the adenine-sensing riboswitch (ASW) in complex with 30S ribosome from *E. coli*. a Overall depiction of the initiating riboswitch with the 3′-end of the 16S rRNA which is shown in dark gray. The structural modules of the ASW are color-coded according to Fig. 1: aptamer (P1, P2, P3) in magenta, 5′-strand of helix P4 in orange, nucleotides of opened stem P5 in purple. The residues of rS2 forming the binding site for rS1 at the 30S ribosome are highlighted in beige[23,30]. **b**, **c** Base pairing interaction of the Shine-Dalgarno sequence of the ASW with the anti-Shine-Dalgarno sequence of the 16S rRNA.

to complete destabilization of the apoB conformation, complete opening of helices P4 and P5 and release of nucleotides 82–115 to adopt single strand conformation.

We built a molecular model of the initiating ribosome in complex with the 5′-end of the ASW mRNA, restrained by the base pairing structure derived from the NMR experiments (Fig. 4). The effect of rS1 on opening of the translation initiation site (P4, P5 and the single stranded 3′ nucleotides) was accompanied by a stabilization of base-pairs stemming from P1, P2 and the connecting loops and junctions of ASW$^{14–140}$, therefore the ligand bound aptamer was modeled in its compact tertiary structure. Base pairs from position 82 to 115 (P4 and P5) were molten by the 30 S complex and thus the nucleotides were kept in single stranded conformation. Consequently, the RBS including the SD sequence and neighboring secondary structure elements impeding productive mRNA binding becomes accessible for interaction with the 16S rRNA (Fig. 4c). The model showed that only if helix P5 was single stranded steric clash between r-proteins and mRNA could be avoided at the entrance of the tunnel. At this point, melting of P5 is a unique feature that only occurs through ribosome-bound rS1.

Here, rS1 exerts its RNA chaperoning function[20] directly at the ribosome, fulfilling its role in mRNA delivery by keeping the ribosome binding site on the mRNA in a single-stranded conformation. As it is shown that rS1 promotes partial mRNA unfolding also in other riboswitches[28], facilitating single stranded conformations within the effector domains of translational riboswitches might be a general function of rS1. However, for the ASW, the observed unfolding of this part of the riboswitch is prerequisite for productive incorporation of the mRNA into the 30S ribosome.

In summary, we found that for translational riboswitches binding of the inducer ligand stabilizes the aptamer domain and destabilizes the sequestering effect of the cis-acting complementary strand precluding SD opening. However, the presence of ligand and riboswitch RNA alone was insufficient to completely release the ribosome binding site. To achieve complete opening, the interaction with the translation machinery and in particular with the rS1 protein was essential. A translational ON-state in the complex between the riboswitch and the 30S ribosome became effectively populated only when both, the inducer ligand and the modulator rS1 were present. Only then, the SD sequence is completely released and could base pair with the aSD sequence of the 16S rRNA leading to correct positioning of the start codon. Thus, our model delineates all required components for translation initiation by riboswitches. We note that our data for the adenine-sensing riboswitch are consistent with the standby model of translation initiation[4,5,29]. Further, it challenges the “RNA-only” model of riboswitch function, as translational regulation only occurs upon concerted yet dynamic interaction of mRNA, 30S ribosomes and associated modulator proteins.

Methods

DNA template. The DNA template of 127 nt adenine-sensing riboswitch (ASW) from *Vibrio vulnificus* is flanked by two restriction sites and two ribozymes at both ends in order to maintain 3′- and 5′-end homogeneity:

GAATTC (EcoRI) – TAATACGACTCACTATAG (T7 Pol) – GGTATGAAGC CTGATGAGTCCGTGAGGACG AAAGACCGTCTTCGGACGGTCTC (Hammerhead ribozyme) – GCTTCATATAATCCTAATGATATGGTTT GGGA GTTTCTACCAAGAGCCTTAAACTCTTGATTATGAAGTCTGTCGCTTTATC CGAAATTTTATAAAGAGAAGACTCATGAATTACTTTGACCTGCCG (127 nucleotide ASW) – GGGTCGGCATGGCATCTCCACCTC CTCGCGGTCCGA CCTGGGCTACTTCGGTAGGCTAAGGGAGAAG (HDV ribozyme) – CCCG GG (SmaI).

Riboswitch preparation. ^{13}C–^{15}N-labeled 127-nucleotide *add* ASW was synthesized by in vitro transcription with T7 RNA polymerase from linearized plasmid DNA[31]. The ^{13}C–^{15}N-labeled (rATP, rCTP, rGTP, rUTP) nucleotides for transcription were purchased from Silantes (Munich, Germany). The construct was purified by preparative polyacrylamide gel electrophoresis according to standard protocols[32]. The construct was folded in water for 5 min at 95 °C and immediately diluted 10-fold with ice-cold water. The RNA was buffer-exchanged into NMR buffer (25 mM potassium phosphate, 150 mM KCl, 5 mM $MgCl_2$, pH 7.2). For protein-RNA studies, 5 mM DTT was added.

Purification of rS1, rS1ΔD6 and rS1ΔD56. The rS1 sequence was taken from *Vibrio vulnificus* (strain CMCP6). rS1, rS1ΔD6 and rS1ΔD56 were expressed in BL21(DE3) cells carrying a N-terminal polyhistidine- and thioredoxin-tag. Cell cultures were grown in LB-medium at 37 °C. At OD_{600} 0.6–0.8, protein expression was induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and incubated over night at 20 °C. The culture was harvested and resuspended in 50 mM Tris-HCl (pH 8), 300 mM NaCl, 10 mM imidazole, 10 mM β-mercapto-ethanol. The protein was purified from the soluble fraction with a 5 mL HisTrap column (GE Healthcare) using 50 mM Tris-HCl (pH 8), 300 mM NaCl, 500 mM imidazole, 10 mM β-mercapto-ethanol for elution. Tags were cleaved overnight and at 4 °C using TEV protease and removed via a second HisTrap column. The protein was further purified by size exclusion chromatography on a HiLoad 26/60 S75 column (GE Healthcare) using NMR buffer (25 mM potassium phosphate, 150 mM KCl, 5 mM DTT, pH 7.2).

Purification of 30S ribosomes and rS1-depleted 30S ribosomes. 30S ribosomes were purified based on His-tagged proteins L7/L12 following the single step protocol from Ederth et al.[33] Briefly, *E. coli* strain JE28 was grown in LB medium supplemented with 50 μg/mL kanamycin at 37 °C. At OD_{600} 0.6, the culture was harvested by centrifugation at 4000 g for 5 min. The cell pellet was resuspended in lysis buffer (20 mM Tris–HCl pH 7.6, 10 mM $MgCl_2$, 150 mM KCl, 30 mM NH_4Cl) supplemented with 0.5 mg/mL lysozyme, 50 μL DNase 1 (2000 U/mL) and cOmplete™ Protease Inhibitor Cocktail tablet (Roche) and further lysed using a Microfluidizer M-110P (Microfluidics, USA) in a final volume of 50 mL. The lysate was clarified by centrifuging twice at 38420 g at 4 °C, 20 min each. For affinity purification, a HisTrap column (Ni^{2+} sepharose pre-packed, 5 mL, GE Healthcare) was connected to an ÄKTA prime chromatography system (GE Healthcare) pre-equilibrated in lysis buffer. After loading the lysate, the column was washed with 10 column volumes of 7 mM imidazole in lysis buffer. The purification of 30SΔS1 ribosomes was performed as described before. To elute the ribosomal S1 protein, the column was washed with 1.5 M NH_4Cl in lysis buffer. To elute 30S ribosomes, a decreasing magnesium concentration gradient was applied from 10 mM to 1 mM magnesium. Fractions were concentrated to 1 mL using centrifugal filter units (Vivaspin 20, MWCO 30000) and loaded onto a sucrose gradient of 20% to 50% sucrose in sucrose gradient buffer (20 mM Tris-HCl, pH 7.6, 300 mM NH_4Cl, 5 mM $Mg(OAc)_2$, 0.5 mM EDTA and 7 mM β-mercapto-ethanol). Gradients were centrifuged for 16 h at 100000 g and 4 °C, fractionated with a Piston Gradient Fractionator (BioComp Instruments, Canada) and analyzed by SDS gel electrophoresis (Supplementary Fig. 8). 30S ribosomes were stored in sucrose buffer at 4 °C. Concentration of 30S ribosomes were determined at A_{260} where 1 A_{260} correspond to 72 pmol/mL[34]. To avoid disturbing effects of sucrose, sample concentration was determined in 1:100 dilutions after buffer exchange.

30S ribosome stability and integrity of NMR samples. High resolution NMR experiments require that the sample of the macromolecular complex has to be pure from contaminating proteins and nucleic acids, it must be monodispersed and due to the experimental restraints of the technique highly concentrated. Purity of the final 30S ribosomes utilized in all subsequent experiments was judged by gel-based analysis in comparison to similarly purified 50S and 70S particles (Supplementary Fig. 8).

Ribosomes were further analyzed by Western Blot using a 6x-His Tag Antibody (Thermo Fischer) that allows selective detection of the modified ribosomal protein L12. The Western Blot with this specific antibody clearly shows that 30S particles are free of impurities of 50S protein components.

Integrity of ribosomes were further analyzed by native composite gel electrophoresis and quantitative mass spectrometry (Supplementary Fig. 9). After acquiring NMR experiments, the NMR samples were analyzed for RNA integrity and 30S ribosome stability. RNA integrity was checked by denaturing polyacrylamide gel electrophoresis (Supplementary Fig. 8). Theoretically, RNA could possibly be degraded by co-purification of RNases during ribosome preparation. Ribosome stability was further analyzed by loading the measured samples on another sucrose gradient (Supplementary Fig. 10).

Electrophoretic mobility shift assay (EMSA). The EMSAs to investigate RNA-S1 binding were performed at constant RNA concentration (4 μM) and increasing amounts of protein (0–10 equivalents) in the absence and presence of adenine. All samples were prepared in EMSA buffer (25 mM potassium phosphate, 75 mM KCl, 5 mM DTT, pH 7.2) and were incubated for 30 min on ice. Bands were separated on continuous gels (8% PAGE) at low powers (<500 mW, 3.5 V/cm) in TAM buffer (40 mM Tris, 20 mM acetic acid, 5 mM $MgCl_2$, pH 8.3). The RNA was visualized upon staining with GelRed^{TM} (Biotium) and subsequent excitation of the bands at 254 nm.

Native composite gel. For the analysis of ribosome stability and electrophoretic mobility shift assays, native composite gels were performed which allow the migration of native ribosomes into the gel[35]. Optimized composite gels were composed of 1.76% acrylamide and 0.68% agarose. Therefore, 13.5 mL of 0.92% agarose solution was cooled to 65 °C. Subsequently, 3 mL of prewarmed buffer (0.2 M Tris, 0.125 M acetic acid, 35% (w/v) sucrose) were added to the agarose solution. Then, 1 mL of acrylamide (29:1 acrylamide:bisacrylamide), 8.5 μL of Triton-X100 and $MgCl_2$ (5 mM) was added and heated to 65 °C. Polymerization was induced by adding 100 μL of APS and 20 μL of TEMED and poured immediately into a prewarmed gel chamber. The gel was polymerized for 1 h at 4 °C. A pre-run was conducted at 200 V for 30 min under water cooling. The running buffer (40 mM Tris, 25 mM acetic acid, 5 mM $MgCl_2$) was exchanged.

The 127 nt *add* riboswitch was prepared by the ligation of a 97 nt fragment (G14–G110) with the Cy5-labeled 30 nt RNA (A111-G140) (Dharmacon) as described before[36]. The RNA was pre-incubated with 1.4 equivalents of adenine. Ribosome concentration was increased from 1 to 10 equivalents at a constant RNA concentration of 1.28 μM. All samples were prepared in NMR buffer (25 mM potassium phosphate, 150 mM KCl, 5 mM $MgCl_2$, pH 7.2) and a final glycerol concentration of 10% as loading buffer. Samples were loaded and the gel was run for 1 hour and 20 min at 200 V under water cooling. Bands were visualized by UV shadowing, fluorescence scan and further stained with Coomassie.

Ribosomal RNA extraction. rRNA of 30S ribosomes were analyzed by agarose gel electrophoresis. Therefore, rRNA was extracted with 1 vol. eq. PCI. Aqueous phase was mixed with gel loading dye and analyzed by 1% agarose gel electrophoresis. The RNA was visualized upon staining with GelRed^{TM} (Biotium).

NMR Spectroscopy. All experiments were performed at 25 °C in NMR buffer (25 mM potassium phosphate, pH 7.2, 150 mM KCl, 5 mM $MgCl_2$) and 5% to 10% D_2O. For protein-RNA studies, 5 mM DTT was added. For ligand bound RNA, 1.4 equivalents of ^{13}C-^{15}N-labeled adenine was added. ^{13}C-^{15}N-labeled adenine was synthesized as described[37].

All spectra were referenced to DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid). Nitrogen-15 and Carbon-13 chemical shifts were indirectly referenced using the ratio of the gyromagnetic ratios of proton to ^{15}N (0.101329118) and ^{13}C (0.251449530), respectively[38].

For titration experiments of the ^{15}N-labeled 127 nucleotide *add* adenine-sensing riboswitch with the ribosomal S1 protein, the RNA concentration was set to 40 μM. The unlabeled protein rS1 was stepwise added to the RNA in ratios ranging from 0 to 5.

For the investigation of mRNA-ribosome complexes with the 127 nucleotide ASW, samples were prepared with a concentration of 35 μM ribosomes and RNA, since higher concentrations lead to aggregation of the ribosomes.

NMR experiments were performed on an 800 MHz, 900 MHz or 950 MHz NMR spectrometer equipped with a 5 mm, z-axis gradient ^{1}H {^{13}C, ^{15}N} TCI cryogenic probe.

BEST-TROSY (Band-selective Excitation Short-Transient - Transverse relaxation-optimized spectroscopy) experiments were recorded to observe ^{1}H, ^{15}N correlations of ^{15}N isotope labeled RNA[39]. Spectra were recorded using a modified pulse program[40].

NMR Experiments were analyzed using Bruker Biospin software TopSpin 3.5. Assignments were performed using the software Sparky 3.114[41].

Microscale thermophoresis. The 127 nt add riboswitch was prepared by the ligation of a 97 nt fragment (G14–G110) with the Cy5-labeled 30 nt RNA (A111–G140) (Dharmacon) as described before[36]. 30SΔS1 ribosomes were labeled with Monolith NT Protein Labeling kit red-maleimide according to the protocol (NanoTemper Technologies). The Cy5-labeled 127 nt riboswitch (30 nM) was incubated with a 1:1 dilution series of 30S ribosomes (36 μM) and 30SΔS1 ribosomes (31 μM) in the absence and presence of adenine (500 μM). Labeled 30SΔS1 (30 nM) was incubated with a 1:1 dilution series of rS1 (3 μM). The complexes were incubated for 30 min at 25 °C in NMR buffer (25 mM potassium phosphate, pH 7.2, 150 mM KCl, 5 mM $MgCl_2$, 5 mM DTT) supplemented with 0.4 mg/mL BSA (New England Biolabs), 0.05% Tween20 and RNasin (Promega, Germany). The reactions were transferred to hydrophobic capillaries and measured with a microscale thermophoresis (MST) device (NanoTemper Monolith NT.115, Germany). After a primary capillary scan, the thermophoresis experiments were carried out at 25 °C with an excitation and MST power of 40% each. Four scans were recorded for each sample. The analysis was performed using MO.Affinity Analysis and the evaluation strategy T-Jump.

In vivo assay of adenine-sensing riboswitch regulated gene expression. BL21 (DE3) *E. coli* cells were transformed with pUC19 plasmid containing T7 regulatory elements and shifted green fluorescent protein (sGFP) gene under the translational control of the adenine-sensing riboswitch ASW14–122. For a pre-culture, a single colony was picked and added to 5 mL M9 minimal medium containing 100 mg/mL ampicillin and IPTG and incubated in a thermoshaker (37 °C, 120 rpm) overnight. The pre-culture was used to inoculate 50 mL M9/ampicillin main cultures with varying adenine and IPTG concentrations in biological triplicates with a start OD_{600} of 0.1, which was grown in a thermoshaker (37 °C, 120 r.p.m.). The production of sGFP was monitored via fluorescence emission (λ_{ex} = 484 nm and λ_{em} = 510 nm) with a Tecan Infinite M200 pro microplate reader (Tecan, Männedorf, Switzerland). For error estimation, three biological independent samples were measured in triplicates. Results were normalized to the growth condition of 0 mM adenine and 0 mM IPTG. As additional control, untransformed BL21(DE3) *E. coli* cells were measured.

Coupled in vitro transcription-translation assay. The coupled in vitro transcription-translation assays and the cell extract preparation were performed as previously described[42] to monitor protein expression levels of shifted green fluorescent protein (sGFP). The cell-free protein expression reactions were performed with in-house made S100 cell extracts from *E. coli* strain BL21 Star^{TM}(DE3) and rS1-depleted ribosomes (final concentration 0.46 μM). The S100 extract preparation was adapted from the S30 extract protocol by Schwarz et al.[43] Our preparation differs at the dialysis step. Here, the S100 extract was dialyzed at 4 °C against dialysis buffer (10 mM Tris/acetate (pH 8.2), 14 mM $Mg(OAc)_2$, 60 mM KOAc, 0.5 mM DTT) using a dialysis membrane (MWCO 12-14 kDa, ZelluTrans, Carl Roth, Germany). The dialysis was performed for 3 h and the buffer was exchanged every hour. After dialysis the S100 extract was centrifuged at 40000 rpm and 4 °C for 5 h under vacuum (Beckman Coulter Optima^{TM} L-90 K). The protein is synthesized from a pUC19 plasmid containing T7 regulatory elements. The sGFP gene is under the translational control of the adenine-sensing riboswitch ASW14–122 that contains a weak ribosome-binding site and the start codon AUG. The 30 μL reactions were performed as triplicates in μClear® 384-well flat bottom microplates (Greiner Bio-One, Frickenhausen, Germany) at 25 °C for 2 h. The production of sGFP was monitored via sGFP fluorescence (λ_{ex} = 470 nm and λ_{em} = 515 nm) with a Tecan Spark® microplate reader (Tecan, Männedorf, Switzerland). For error estimation, one biological sample was measured in triplicates. Native 70S ribosomes were used as positive control.

Structural modeling of mRNA-ribosome complexes. The riboswitch structure was modeled into the mRNA decoding channel of the 30S ribosome.

For the structural modeling, the 30S ribosomal pre-initiation complex structure with the PDB ID 5lmn was chosen. The proteins IF1, IF3 and Thx were removed from the structure.

The adenine-sensing riboswitch structure was built as follows: the holo aptamer structure (C13–G83) with the PDB ID 1y26 was used as starting point. Here, the nucleotides C13 and G83 were removed from the structure.

To build the single stranded region of the mRNA, the online tool RNAthread (ROSIE) was used[44,45]. Here, the single stranded mRNA from the 5lmn structure with 20 nucleotides was used as input structure. The nucleotides were replaced in steps of 10–20 nucleotides with the desired ASW sequence. The resulting structure was then renumbered accordingly to the ASW sequence using the renumber_PDB tool.

Sequences were aligned in bins of two, each containing two overlapping nucleotides. The alignment was performed in Pymol. This procedure ensured preservation of the correct backbone geometry. The redundant nucleotides from the alignment were removed manually in text editor. The mRNA in the 5lmn structure was replaced by the modeled ASW structure, maintaining the same position of the GAG triplet (G110–G112) from the Shine–Dalgarno sequence, that base pairs with the rRNA (C1536–C1538).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data needed to evaluate the conclusions in the paper are present in the paper and/or the supplementary information, and are available from the corresponding author upon reasonable request. The structural model that was generated in this study based on the publically available structures 1Y26 and 5MLN is provided in Supplementary Data 1. Source data are provided with this paper.

Received: 30 March 2021; Accepted: 17 July 2021;
Published online: 05 August 2021

References

1. Shine, J. & Dalgarno, L. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl Acad. Sci. USA* **71**, 1342–1346 (1974).
2. Steitz, J. A. & Jakes, K. How ribosomes select initiator regions in mRNA: base pair formation between the 3′ terminus of 16S rRNA and the mRNA during initiation of protein synthesis in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **72**, 4734–4738 (1975).
3. Marzi, S. et al. Structured mRNAs regulate translation initiation by binding to the platform of the ribosome. *Cell* **130**, 1019–1031 (2007).
4. De Smit, M. H. & Van Duin, J. Translational standby sites: how ribosomes may deal with the rapid folding kinetics of mRNA. *J. Mol. Biol.* **331**, 737–743 (2003).
5. Romilly, C., Deindl, S. & Wagner, E. G. H. The ribosomal protein S1-dependent standby site in tisB mRNA consists of a single-stranded region and a 5′ structure element. *Proc. Natl Acad. Sci. USA* **116**, 15901–15906 (2019).
6. Studer, S. M. & Joseph, S. Unfolding of mRNA secondary structure by the bacterial translation initiation complex. *Mol. Cell* **22**, 105–115 (2006).
7. Darfeuille, F., Unoson, C., Vogel, J. & Wagner, E. G. H. An antisense RNA inhibits translation by competing with standby ribosomes. *Mol. Cell* **26**, 381–392 (2007).
8. Simonetti, A. et al. A structural view of translation initiation in bacteria. *Cell. Mol. Life Sci.* **66**, 423–436 (2009).
9. Breaker, R. R. Riboswitches and translation control. *Cold Spring Harb. Perspect. Biol.* **10**, 1–14 (2018).
10. Rinaldi, A. J., Lund, P. E., Blanco, M. R. & Walter, N. G. The Shine-Dalgarno sequence of riboswitch-regulated single mRNAs shows ligand-dependent accessibility bursts. *Nat. Commun.* **7**, 1–10 (2016).
11. Bédard, A. S. V., Hien, E. D. M. & Lafontaine, D. A. Riboswitch regulation mechanisms: RNA, metabolites and regulatory proteins. *Biochim. Biophys. Acta* **1863**, 194501 (2020).
12. Reining, A. et al. Three-state mechanism couples ligand and temperature sensing in riboswitches. *Nature* **499**, 355–360 (2013).
13. Warhaut, S. et al. Ligand-modulated folding of the full-length adenine riboswitch probed by NMR and single-molecule FRET spectroscopy. *Nucleic Acids Res.* **45**, 5512–5522 (2017).
14. Lemay, J. F. et al. Comparative study between transcriptionally- and translationally-acting adenine riboswitches reveals key differences in riboswitch regulatory mechanisms. *PLoS Genet.* **7**, e1001278 (2011).
15. Fürtig, B. et al. Refolding through a linear transition state enables fast temperature adaptation of a translational riboswitch. *Biochemistry* **59**, 1081–1086 (2020).
16. Yusupova, G. Z., Yusupov, M. M., Cate, J. H. D. & Noller, H. F. The path of messenger RNA through the ribosome. *Cell* **106**, 233–241 (2001).
17. Coetzee, T., Herschlag, D. & Belfort, M. *Escherichia coli* proteins, including ribosomal protein S12, facilitate in vitro splicing of phage T4 introns by acting as RNA chaperones. *Genes Dev.* **8**, 1575–1588 (1994).

18. Semrad, K., Green, R. & Schroeder, R. RNA chaperone activity of large ribosomal subunit proteins from *Escherichia coli*. *RNA* **10**, 1855–1860 (2004).
19. Aseev, L. V. & Boni, I. V. Extraribosomal functions of bacterial ribosomal proteins. *Mol. Biol.* **45**, 739–750 (2011).
20. Duval, M. et al. *Escherichia coli* ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation. *PLoS Biol.* **11**, 12–14 (2013).
21. Hajnsdorf, E. & Boni, I. V. Multiple activities of RNA-binding proteins S1 and Hfq. *Biochimie* **94**, 1544–1553 (2012).
22. Sengupta, J., Agrawal, R. K. & Frank, J. Visualization of protein S1 within the 30S ribosomal subunit and its interaction with messenger RNA. *Proc. Natl Acad. Sci. USA* **98**, 11991–11996 (2001).
23. Byrgazov, K. et al. Structural basis for the interaction of protein S1 with the *Escherichia coli* ribosome. *Nucleic Acids Res.* **43**, 661–673 (2014).
24. Wilson, D. N. & Nierhaus, K. H. Ribosomal proteins in the spotlight. *Crit. Rev. Biochem. Mol. Biol.* **40**, 243–267 (2005).
25. Qureshi, N. S., Bains, J. K., Sreeramulu, S., Schwalbe, H. & Fürtig, B. Conformational switch in the ribosomal protein S1 guides unfolding of structured RNAs for translation initiation. *Nucleic Acids Res.* **46**, 10917–10929 (2018).
26. Saguy, M. et al. Ribosomal protein S1 influences trans -translation in vitro and in vivo. *Nucleic Acid Res.* **35**, 2368–2376 (2007).
27. Kitakawa, M. & Isono, K. An amber mutation in the gene rpsA for ribosomal protein S1 in *Escherichia coli. MGG Mol. Gen. Genet.* **185**, 445–447 (1982).
28. Lund, P. E., Chatterjee, S., Daher, M. & Walter, N. G. Protein unties the pseudoknot: S1-mediated unfolding of RNA higher order structure. *Nucleic Acid Res.* **48**, 2107–2125 (2019).
29. Sterk, M., Romilly, C. & Wagner, E. G. H. Unstructured 5′-tails act through ribosome standby to override inhibitory structure at ribosome binding sites. *Nucleic Acids Res.* **46**, 4188-4199 (2018).
30. Beckert, B. et al. Structure of a hibernating 100S ribosome reveals an inactive conformation of the ribosomal protein S1. *Nat. Microbiol.* **3**, 1115–1121 (2018).
31. Fürtig, B., Richter, C., Wöhnert, J. & Schwalbe, H. NMR spectroscopy of RNA. *ChemBioChem* **4**, 936–962 (2003).
32. Bains, J. K. et al. Combined smFRET and NMR analysis of riboswitch structural dynamics. *Methods* **153**, 22–34 (2019).
33. Ederth, J., Mandava, C. S., Dasgupta, S. & Sanyal, S. A single-step method for purification of active His-tagged ribosomes from a genetically engineered Escherichia coli. *Nucleic Acids Res.* **37**, e15 (2009).
34. Christodoulou, J. et al. Heteronuclear NMR investigations of dynamic regions of intact *Escherichia coli* ribosomes. *Proc. Natl Acad. Sci. USA* **101**, 10949–10954 (2004).
35. Nadano, D., Aoki, C., Yoshinaka, T., Irie, S. & Sato, T. A. Electrophoretic characterization of ribosomal subunits and proteins in apoptosis: specific downregulation of S11 in staurosporine-treated human breast carcinoma cells. *Biochemistry* **40**, 15184–15193 (2001).
36. Duss, O., Maris, C., Von Schroetter, C. & Allain, F. H. T. A fast, efficient and sequence-independent method for flexible multiple segmental isotope labeling of RNA using ribozyme and RNase H cleavage. *Nucleic Acids Res.* **38**, e188 (2010).
37. Noeske, J. et al. An intermolecular base triple as the basis of ligand specificity and affinity in the guanine- and adenine-sensing riboswitch RNAs. *Proc. Natl Acad. Sci. USA* **102**, 1372–1377 (2005).
38. Wishart, D. S. et al. 1H, 13C and 15N chemical shift referencing in biomolecular NMR. *J. Biomol. NMR* **6**, 135–140 (1995).
39. Solyom, Z. et al. BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. *J. Biomol. NMR* **55**, 311–321 (2013).
40. Favier, A. & Brutscher, B. Recovering lost magnetization: polarization enhancement in biomolecular NMR. *J. Biomol. NMR* **49**, 9–15 (2011).
41. Lee, W., Tonelli, M. & Markley, J. L. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **31**, 1325–1327 (2015).
42. Kaldenhoff, R., Bernhard, F., Kai, L. & Do, V. Artificial environments for the co-translational stabilization of cell-free expressed proteins. *PLoS ONE* **8**, e56637 (2013).
43. Schwarz, D. et al. Preparative scale expression of membrane proteins in Escherichia coli-based continuous exchange cell-free systems. *Nat. Protoc.* **2**, 2945–2957 (2007).
44. Watkins, A. M., Rangan, R. & Das, R. FARFAR2: improved de novo rosetta prediction of complex global RNA folds. *Structure* **28**, 963–976.e6 (2020).
45. Lyskov, S. et al. Serverification of molecular modeling applications: the rosetta online server that includes everyone (ROSIE). *PLoS ONE* **8**, e63906 (2013).

Acknowledgements

We are grateful to Suparna Sanyal for providing *E. Coli* JE28 cells. We thank Julian Langer for quantitative mass spectrometry analysis. We thank Frank Löhr for help with HNN-COSY experiments. Further, we thank Volker Dötsch for kindly providing the pIVEX-sGFP plasmid. VDJ was supported by Boehringer Ingelheim Fonds. The work was supported by funds from DFG in collaborative research center 902: "Molecular principles of RNA-based regulation" and in Graduate College GRK1986 (CLIC). Work at BMRZ is supported by the state of Hesse (HMWK).

Author contributions

V.D.J., N.S.Q., S.W. and J.K.B. prepared the samples. V.D.J., N.S.Q. and B.F. performed the NMR experiments. V.D.J. performed analytical gel electrophoresis, EMSAs, composite gels, western blot and sucrose gradients. V.D.J., N.S.Q. and S.W. performed the MST experiments. V.D.J., M.S.D. and M.H. performed binding studies. V.D.J. performed the in vivo assay. J.K.B. performed the coupled transcription-translation assay. B.F. performed rRNA analysis. B.F. and H.S. directed the research. V.D.J., B.F. and H.S. wrote the manuscript. All authors provided critical comments.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-25024-5.

Correspondence and requests for materials should be addressed to B.Für.

Peer review information *Nature Communications*thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

5.5 **Research article:** Wavelength-Selective Uncaging of Two Different Photoresponsive Groups on One Effector Molecule for Light-Controlled Activation and Deactivation

Isam Elamri*, Chahinez Abdellaoui*, Jasleen Kaur Bains, Katharina Felicitas Hohmann, Santosh Lakshmi Gande, Elke Stirnal, Josef Wachtveitl and Harald Schwalbe*; Journal of the American Chemical Society* **143**, 10596-10603 (2021)

* These authors contributed equally to this work.

In this work, an effector molecule with two photocleavable protecting groups was synthesized by orthogonal photolysis of the protecting groups. In a first step, the effector molecule is activated, followed by light-induced deactivation. The translation inhibitor puromycin was synthetically modified to an inactive puromycin using an *ortho*-nitrophenylalanine protecting group. In addition, the modified puromycin was protected by a second protecting group, thio-coumarylmethyl. The photolysis processes upon wavelength-selective irradiation were characterized by UV/Vis-, IR and time-resolved NMR spectroscopy. Furthermore, the translation inhibition effects of the modified puromycin upon illumination were investigated in a cell-free transcription-translation assay using sGFP as a read-out. Translation was inhibited 3-fold by the released *o*-nitro derivative of puromycin after irradiation of thio-DEACM-nitro-puromycin at 488 nm.

I. Elamri synthesized the compound, performed the time-resolved NMR experiments, and wrote the manuscript. C. Abdellaoui performed the UV/Vis- and IR spectroscopy and prepared the manuscript. The author of this thesis established and performed the light-controlled cell-free transcription-translation assay with the puromycin derivatives thio-DEACM-nitro-puromycin and *o*-nitro-puromycin to test its inhibitory effect on the protein expression of sGFP. In addition, the effects of the illumination wavelength on sGFP expression were tested by native polyacrylamide gel electrophoresis (PAGE). The author prepared the figures and wrote the corresponding paragraph in the manuscript.

https://doi.org/10.1021/jacs.1c02817

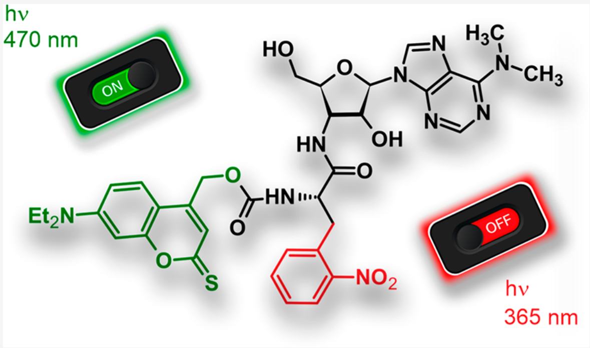
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS Article

Wavelength-Selective Uncaging of Two Different Photoresponsive Groups on One Effector Molecule for Light-Controlled Activation and Deactivation

Isam Elamri,[§] Chahinez Abdellaoui,[§] Jasleen Kaur Bains, Katharina Felicitas Hohmann, Santosh Lakshmi Gande, Elke Stirnal, Josef Wachtveitl,* and Harald Schwalbe*

Cite This: J. Am. Chem. Soc. 2021, 143, 10596−10603 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Photocleavable protecting groups (PPGs) play a pivotal role in numerous studies. They enable controlled release of small effector molecules to induce biochemical function. The number of PPGs attached to a variety of effector molecules has grown rapidly in recent years satisfying the high demand for new applications. However, until now molecules carrying PPGs have been designed to activate function only in a single direction, namely the release of the effector molecule. Herein, we present the new approach Two-PPGs-One-Molecule (TPOM) that exploits the orthogonal photolysis of two photoprotecting groups to first release the effector molecule and then to modify it to suppress its induced effect. The moiety resembling the tyrosyl side chain of the translation inhibitor puromycin was synthetically modified to the photosensitive *ortho*-nitrophenylalanine that cyclizes upon near UV-irradiation to an inactive puromycin cinnoline derivative. Additionally, the modified puromycin analog was protected by the *thio*-coumarylmethyl group as the second PPG. This TPOM strategy allows an initial wavelength-selective activation followed by a second light-induced deactivation. Both photolysis processes were spectroscopically studied in the UV/vis- and IR-region. In combination with quantum-chemical calculations and time-resolved NMR spectroscopy, the photoproducts of both activation and deactivation steps upon illumination were characterized. We further probed the translation inhibition effect of the new synthesized puromycin analog upon light activation/deactivation in a cell-free GFP translation assay. TPOM as a new method for precise triggering activation/deactivation of effector molecules represents a valuable addition for the control of biological processes with light.

■ INTRODUCTION

In recent years, photocleavable protecting groups (PPGs) have become a powerful tool for light-induced release of effector molecules in Chemical Biology.[1−8] The list of novel "caged" effector molecules is constantly growing. Besides their biological function, they satisfy a large number of different photophysical requirements[9] including high quantum yield φ,[10] absorption at $\lambda > 350$ nm,[11−14] high absorption coefficient ε,[15] rapid uncaging,[16,17] applicability for two-photon uncaging,[18,19] and wavelength-selective photolysis[20] as well as biocompatibility,[21,22] and solubility under physiological conditions,[23] cell permeability,[24−26] to name a few important design criteria of PPGs.

Here, we describe a new strategy that enables the light-controlled activation and subsequent deactivation of effector molecules. Figure 1 schematically demonstrates our approach: in addition to the conventional photocleavable protecting group (marked in orange), whose photochemical release activates an effector molecule, the compound contains a second photoresponsive group (marked in blue), ideally

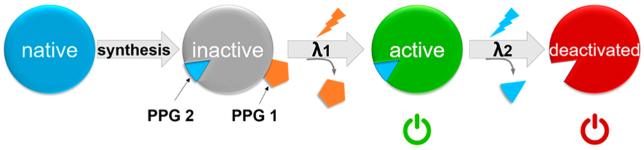
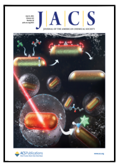


Figure 1. Schematic representation of a doubly caged effector molecule. PPG 1 (orange) is attached to the functional group of the molecule. A second protecting group is incorporated as a part of the active structure (blue triangle). Both can be removed selectively upon orthogonal irradiation (with λ_1 and λ_2), which activates or deactivates the effector molecule.

Received: March 15, 2021
Published: July 8, 2021



https://doi.org/10.1021/jacs.1c02817

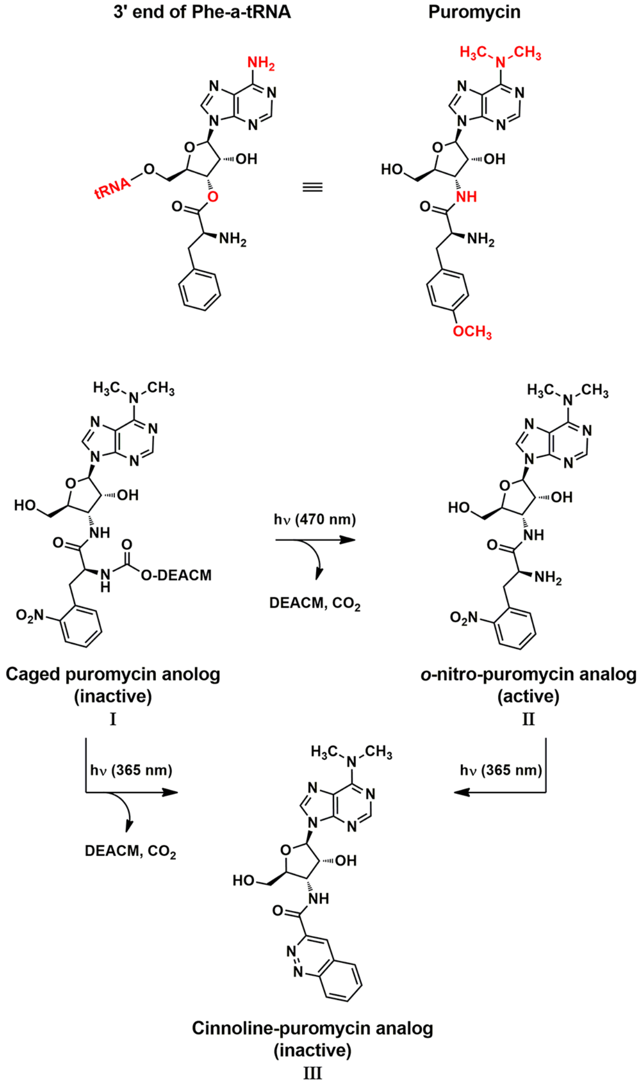
without affecting the biological activity induced by the effector. The "uncaged" effector molecule after the first irradiation (λ_1) can subsequently be deactivated by irradiation at a different wavelength (λ_2) as a response of the photoreaction induced by absorption of the second photocleavable protecting group. Achieving practical applicability of this process requires a number of aspects: 1. The first protecting group is able to suppress the biological activity of the effector molecule. 2. Both involved PPGs need to have absorption spectra with little spectral overlap enabling independent, wavelength-selective photoprocesses. 3. The second PPG attached to the effector induces a modification that should not induce a significant loss of effector activity. After removing the second PPG, the effector molecule has to be inactive; thus, the effector molecule has to be chemically modified in the second step.

To develop the new method TPOM, we selected puromycin as a naturally occurring antibiotic with widespread application in biochemical studies. The aminonucleoside puromycin produced by *Streptomyces alboniger* is a universal translation inhibitor of the peptidyl transferase reaction. Furthermore, it is frequently utilized to visualize newly synthesized proteins.[27] Puromycin mimics the phenylalanyl-tRNA (Phe-a-tRNA) (Scheme 1). It enters the ribosome and incorporates itself into the growing peptide chain. As amide, the puromycin rest cannot be released during translation. Antibodies raised against these conjugates carrying puromycin on the C-terminal are used to detect nascent chains.[28] Puromycin acts codon independently, does not rely on soluble translation factors, and does not induce EF-Tu-GTPase activity.[29] Interestingly, at low concentration (e.g., 0.04 μM) puromycin does not act as an inhibitor, but binds specifically to newly synthesized full-length proteins at their C-terminus.[30] Puromycin acts non-selectively and exhibits a high toxicity in humans. This high toxicity has prevented its application as an antibiotic. Nonetheless, the use of puromycin continues to be of vital importance for mechanistic studies of gene expression regulation under physiological and pathological conditions.[31]

Previously,[16,17,32,33] we developed two photolabile puromycin derivatives protected by a single PPG: either by 6-nitroveratryl oxycarbonyl (NVOC-puromycin) or by 7-*N*,*N*-(diethylaminocumarin-4-yl)-methoxycarbonyl (DEACM-puromycin). Both photosensitive derivatives showed promising potential for photocontrolled release of puromycin.

Our new puromycin derivative developed here extends the use of PPGs to two fundamental processes: the release of puromycin as an effector by light and the subsequent deactivation of this effector, again by light. In designing such a molecule, we came across previous related approaches of sequential photocleavage of different effectors from multi-protected substrates. Klan et al. reported a monochromophoric linker to orthogonally release two different effectors.[34] Similarly, Singh et al. developed a novel approach based on a synthesized *o*-hydroxycinnamate derivative that enables the controlled release of alcohol and carboxylic acid.[35] Sterner et al. developed light-sensitive diarylethene-based competitive inhibitors that upon irradiation with UV light were able to undergo an electro-cyclization affecting their affinity for the active site of an allosteric bienzyme complex.[36] These previous approaches could not be transferred to the controlled activation and deactivation of puromycin, because the introduced photolabile groups are incompatible for release of a functional puromycin.

Scheme 1. Schematic Representation of the Strategy of a Puromycin Analog with Two Photoactivatable Groups[a]



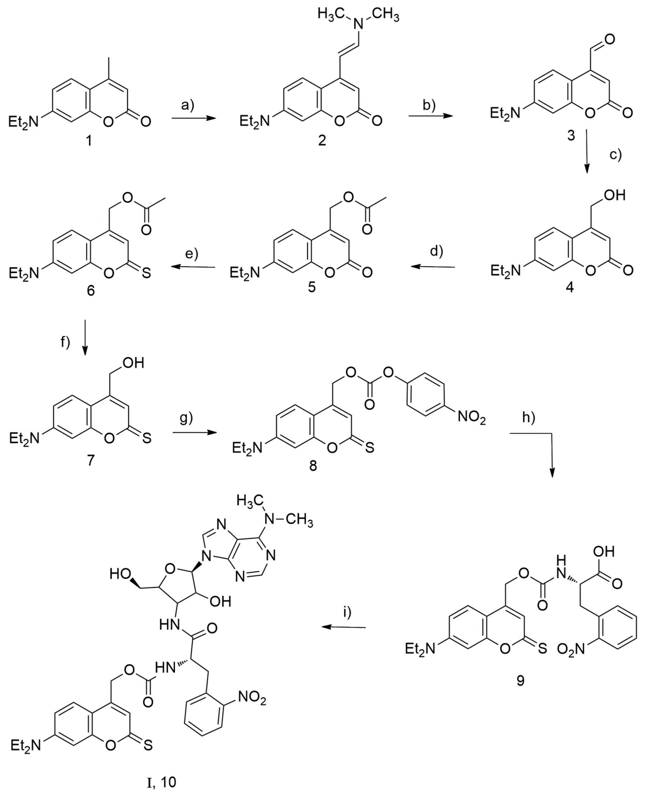
[a]Top: Structural similarity between 3′-end of a phenylalanyl-tRNA and the aminonucleoside antibiotic puromycin. Below: (Step 1) Uncaging of *thio*-coumarin group and release of active *o*-nitro-puromycin upon irradiation at 470 nm. (Step 2) Light-induced photocyclization at 355−365 nm to the inactive cinnoline derivative.

Our approach is based on work by Schultz et al., who demonstrated a technique in which a protein containing photosensitive 2-nitrophenylalanine (2-NPA) was expressed and site-specifically photocleaved.[37] The photochemistry of 2-NPA involves a cinnoline-forming photocyclization after UV-light illumination. The reported tolerance of puromycin to different structural modifications[38−42] inspired us to modify the amino acid part *O*-methyl-L-tyrosine of conventional puromycin to the photoactivable *o*-nitrophenylalanine. In such a derivative, the new photosensitive puromycin analogue (*o*-nitro-puromycin **II**) would structurally resemble the 3′end of $tRNA^{Phe}$ and contain a photosensitive group (Scheme 1). Additionally, as is the case with the conventional puromycin, the photoprotection of the α-amino group of modified *o*-nitro-puromycin (**II**) with a coumarylmethyl derivative as the carbamate (**I**) allows control of the release of biologically active *o*-nitro-puromycin (**II**). Simultaneously, the introduced *o*-nitrophenylalanine part enables a different photochemistry initiating the photocyclization to the nonactive cinnoline derivative **III**, due to the lack of the α-amino group.

■ RESULTS AND DISCUSSION

We synthesized *thio*-DEACM-nitro-puromycin (**I**) in nine steps (Scheme 2). Starting from 7-amino-4-methyl coumarin, we prepared the alcohol DEACM-OH **4** by condensation of coumarin **1** to the enamine **2** followed by treatment with sodium periodate to the aldehyde **3** which was subjected to the reduction using sodium borohydride to yield DEACM-alcohol **4**.[17] Using DEACM as one of the two photolabile groups in the TPOM strategy posed, however, a challenging obstacle exists for the intended photocyclization due to the high absorption of **4** that is released as a byproduct together with the effector molecule *o*-nitro-puromycin (**II**) (Figures S24 and S25). The attempt to generate the selectivity by applying single- and two-photon-induced uncaging of the nitrophenyl part and DEACM, respectively, remained unsuccessful, despite the large difference in the two-photon cross section. In the presence of DEACM an energy transfer to the nitrophenyl part takes place preventing a selective uncaging of DEACM during the activation step. Additionally, we observed by light-coupled NMR experiments that released DEACM-OH (with high ε = 20 800 M^{-1} cm^{-1}) absorbs the incident light intended for the photocyclization and undergoes a photo side-reaction to an unknown byproduct (Figure S15). To circumvent this disadvantage, we tested other PPG derivatives as the photolysis efficiency depends strongly on the type of the substituents of the coumarin scaffold (Figure 2).[3,21,43−45] As is apparent from the UV spectra, diethylamino-thiocoumarylmethyl (*thio*-DEACM) combines the advantages of the different derivatives as it features a diethylamino group on position 7 and a strong electron-withdrawing group at position 2. Incorporation of *thio*-DEACM would thus enable an excitation of the first PPG to yield an activation of the *thio*-DEACM-nitro-puromycin (**I**) with visible light. Of particular importance for the on/off functionality is its absorption gap around 370 nm (Figure 2, red area), which is advantageous in view of the deactivation step.

Scheme 2. Synthesis of Caged *thio*-DEACM-*ortho*-nitro-puromycin I[a]



[a](a) 7-Diethylamino-4-methylcoumarin **1** (1 equiv), DMF−DMA (2 equiv), DMF, 160°C, 7 h; (b) **2** (1 equiv), $NaIO_4$ (3 equiv), THF/H_2O 1:1, rt, 5 h; (c) **3** (1 equiv), $NaBH_4$ (1.5 equiv), EtOH, rt, 4 h, 98%; (d) **4** acetic acid, DCC, DMAP, CH_2Cl_2, reflux, 4 h; (e) **5** Lawesson's reagent, toluene, reflux, 24 h, (f) **6** 1.25 M HCl in ethanol, reflux, 15 h; (g) **7** (1 equiv), DMAP (2 equiv), 4-NO_2Ph-chloroformate (2 equiv), CH_2Cl_2, rt, 7 h; (h) **8** (1 equiv) 2-nitrophenylalanine.HCl (1 equiv), DMAP (2 equiv), DIPEA (20 equiv), CH_2Cl_2, rt, 24 h; (i) **9** (1 equiv), EDCI (1 equiv), NHS (1 equiv), Puromycin aminonucleoside (1 equiv), DMF, 0°−rt, 24 h. Synthesis procedure and characterization of *thio*-DEACM-nitro-puromycin (**I**,**10**) are presented in the Supporting Information (pp S2−S20).

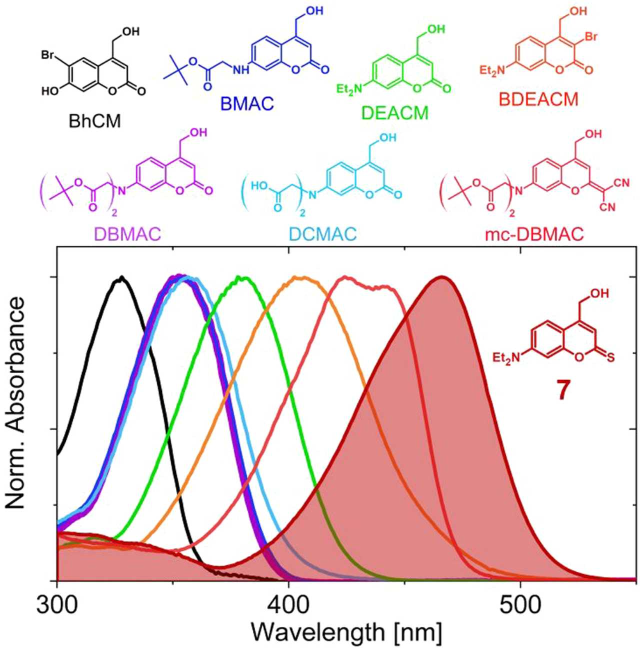


Figure 2. Main absorption bands of different coumarin derivatives in 1:1 acetonitrile/water mixtures. From left to right: BhCM (black, λ_{max} = 327 nm), DBMAC and BMAC (purple and blue, λ_{max} = 353 nm), DCMAC (cyan, λ_{max} = 357 nm), DEACM (green, λ_{max} = 380 nm), BDEACM (orange, λ_{max} = 406 nm), mc-DBMAC (red, λ_{max} = 425 nm/441 nm), *thio*-DEACM (dark red, λ_{max} = 470 nm). Corresponding synthesis procedures and characterizations of 4-hydroxymethyl-coumarin derivatives are presented in the Supporting Information (pp S10−S11).

Therefore, *thio*-coumarin alcohol **7** was synthesized as follows (Scheme 2): Alcohol **4** was converted to ester **5** with acetic acid catalyzed by DCC and subsequently thionylated by Lawesson's reagent.[46] **6** was hydrolyzed providing the desired alcohol **7**. Alcohol **7** was further activated by conversion to the nitro-phenyl carbonate **8** as an intermediate, which then reacts without further purification with the commercially available 2-nitrophenylalanine to yield carbamate **9**. In the last step, the obtained carbamate **9** was treated with puromycin-aminonucleoside in the presence of the coupling reagents EDCI and

NHS to yield doubly caged *thio*-DEACM-nitro-puromycin (**I**) in 43% yield after reversed-phase chromatography.

Since the modified puromycin analogue (**II**) varies from different reported nitrobenzyl scaffolds,[47,48] we decided to first verify the photosensitivity of the *o*-nitro-puromycin (**II**) by examining the proposed cyclization reaction by real-time laser-NMR.[49] We synthesized *o*-nitro-puromycin (**II**) as described in the Supporting Information (pp S7−S9). The laser-coupled ^{1}H NMR experiment showed that irradiation at 488 nm does not affect *o*-nitro-puromycin (**II**), which is favorable for the planned wavelength-selective uncaging of *thio*-DEACM (Figure S12) that should occur before the cyclization step. In the case of 355 nm laser irradiation, most signals stemming from the nitrophenyl ring and the amino nucleoside part of the illuminated compound (**II**) decayed exponentially (Figure 3a).

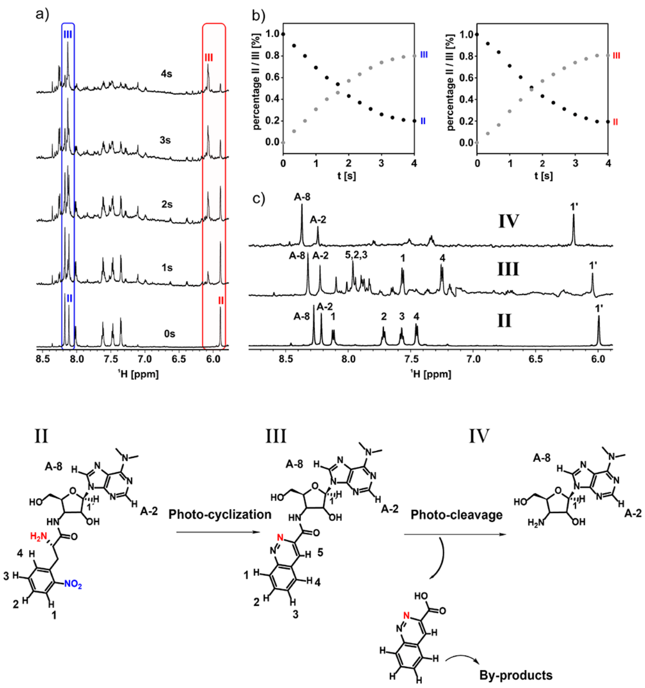


Figure 3. Laser-coupled NMR experiment for tracking the photolysis reaction of *o*-nitro-puromycin (**II**). Estimated Light power: 1 W. NMR-sample: 500 μM in HEPES/D_2O (10%), 355 nm. ^{1}H-spectrum measured after every 0.5 s irradiation. (a and b) Time-resolved ^{1}H-spectra and normalized peak integrals of H1′ and H2 corresponding to the two compounds showing the decay of *o*-nitro-puromycin (**II**) and the accumulation of the photoproducts **III** and **IV**. (c) Spectra after HPLC purification of irradiated sample of *o*-nitro-puromycin **II** showing: (d) photocyclized puromycin-cinnoline **III** and the release of aminonucleoside puromycin **IV** as final photolysis step in comparison with the initial structure before irradiation of *o*-nitro-puromycin (**II**).

This decay was accompanied by a corresponding rise of new photoproduct signals. As an example, Figure 3b shows the exponential signal decay at 5.82 and 8.12 ppm and a signal increase at 6.10 and 8.30 ppm, respectively. The new signals could be assigned to the expected cinnoline derivative **III**. Interestingly, an HPLC-purification of an irradiated sample containing all photoproducts showed a subsequent photo-induced cleavage of the cinnoline derivative **III** to the nucleoside and amino acid moieties as a consecutive reaction of derivative **III** (Figure 3c and 3d). Noteworthy, controlled irradiation of the amino acid part 2-nitrophenylalanine at 355 nm, as the only photolabile fragment of *o*-nitro-puromycin (**II**), led also to a cyclized cinnoline without a carboxylic acid group. Contrary, the same laser irradiation of conventional phenylalanine without the nitro group resulted in no structural changes confirming the photoreaction of the 2-nitrophenylalanine part, excluding any photodegradation of the amino acid itself (Figure S12).

The photophysical properties of the *thio*-coumarylmethyl and 2-nitrophenylalanine groups of the respective photocages were examined. In search of a deactivation wavelength that is within the absorption gap of *thio*-DEACM, several wavelengths between 280 and 450 nm were tested for the photocyclization of *o*-nitro-puromycin (**II**) in separate illumination experiments, although *o*-nitro-puromycin (**II**) seems to barely absorb beyond 300 nm (Figure 4, green line). DFT-calculated UV/vis absorption spectra of *o*-nitro-puromycin (**II**) and its cinnoline photoproduct **III** show that successful uncaging is achievable with wavelengths up to 420 nm (Figure S20a). An activation wavelength λ_a = 470 nm was applied to remove the *thio*-DEACM, while deactivation was performed at λ_d = 365 nm. Both the activation and deactivation wavelengths are within the spectral window suitable for biological applications. The cyclization of *o*-nitro-puromycin (**II**) without *thio*-DEACM attached is shown in Figure 4. Based on DFT calculations the emerging absorption band at 240 nm can be assigned as the main absorption band of the cinnoline photoproduct (Figure S20a). The deactivation step of *o*-nitro-puromycin (**II**) successfully proceeds via an intermediate to a cinnoline photoproduct. Considering the NMR results, further decomposition reactions can be explained by the higher illumination power compared to the UV/vis experiments. During the activation step, the absorption bands of the *thio*-DEACM **7** part of **I** at 474 and 250 nm disappear (Figure 5a, blue spectrum compared to green spectrum) (Figure S16). A residual band at 271 nm (Figure 5a, green solid line) can be assigned to the desired uncaging product *o*-nitro-puromycin (**II**) (Figure 5a, green area), indicating a successful activation of the compound. Upon *thio*-DEACM cleavage, the alcohol derivative *thio*-DEACM-OH is formed, in agreement with a previous report.[50] This particular *thio*-coumarin chromophore undergoes further photolysis upon 470 nm illumination, which yields DEACM-OH, but additional photoproducts are possible.[46] Irradiating pure *thio*-DEACM-OH at the activation wavelength in a separate experiment (Figures S17 and S18a) confirms the presence of DEACM-OH and another photolysis product absorbing around 370 nm. The latter was identified by

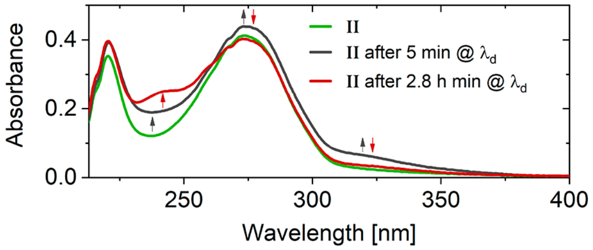


Figure 4. Absorbance spectra of *o*-nitro-puromycin (**II**) during continuous illumination at λ_d = 340 nm in 1:1 acetonitrile/water mixture before (green), after 5 min (gray line), and after 2.8 h (red) of illumination resulting in cinnoline product **III**. The arrows underline the direction of absorption change during the first 5 min of illumination in gray and during the following 2.7 h in red.

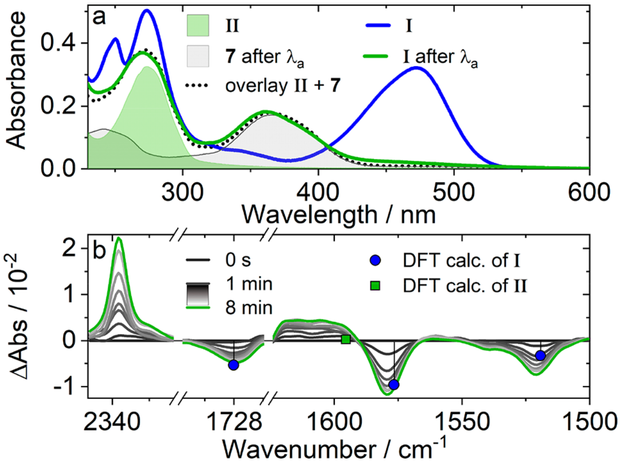


Figure 5. (a) Absorption spectra of activation in 1:1 acetonitrile/water mixture. *Thio*-DEACM-nitro-puromycin (**I**) before (blue line) and after 14 min illumination at 470 nm (green line), which is identical to the (2x(**7**):1x(**II**)) weighted overlay (black dotted line) of the synthesized *o*-nitro-puromycin (**II**) (green area) and the illuminated *thio*-DEACM (**7**) (gray area). (b) Difference IR-spectra measured every 60 s in DMSO during continuous illumination of (**I**) at the activation wavelength (470 nm) evolving from the black (zero line: no illumination of the starting compound (**I**)) to the green line (photoproduct of the activation step: *o*-nitro-puromycin (**II**)). DFT-calculated vibrational modes as sticks of *thio*-DEACM-nitro-puromycin (**I**) (circles) and *o*-nitro-puromycin (**II**) (square). (c) Determination of quantum yield by light induced absorption changes monitoring the photoinduced release of CO_2 upon irradiation of (**I**) at 2337 cm^{-1} with a fit through the first 8 data points (500 s of illumination) (p S30).

mass spectrometry (Figure S14) and NMR spectroscopy. The results indicate the formation of a coumarylmethyl hydrogen sulfate derivative DEACM-OSO_3^- as the major photoproduct of photolyzed *thio*-DEACM-OH.

Since the absorption spectrum of the doubly caged compound (**I**) is in principle an additive superposition of the individual components, with slightly red-shifted absorption bands (Figure S16), the uncaging products are hardly distinguishable in the UV/vis region. IR-experiments after illumination reveal, however, absorption changes that clearly confirm uncaging (Figure 5b). The newly appearing positive absorption band at 2337 cm^{-1} in combination with the vanishing band at 1728 cm^{-1} demonstrate the release of CO_2 from the carbamate linker during illumination taking place in the last step of uncaging.[37] Two other prominent absorption changes can be observed for IR bands at 1579 and 1522 cm^{-1} characteristic for ring vibrations of the coumarin moiety, confirming its photolysis. Compared to the reported DEACM-puromycin[17] the uncaging wavelength could successfully be red-shifted by about 100 nm without a significant loss of quantum yield (1.7%, Figure S18; 69% conversion within 2 min of light exposure: 2 × 10^{16} photons/s). In combination with the high extinction coefficient $\varepsilon_{474\ nm}$ = 17464 M^{-1} cm^{-1} this results in a reasonable uncaging cross section $\varphi{\cdot}\varepsilon$ = 297 M^{-1} cm^{-1}.

Figure 6 shows the deactivation step of a sample previously activated with 470 nm light. It contains the side products of the first uncaging step. Upon exposure to 365 nm the absorption band in the UV-region undergoes a bathochromic shift, which is indicative for the formation of a cinnoline product as known from DFT-calculations (Figure S20a). Regardless of the fact that the *thio*-DEACM side products absorb at the deactivation wavelength (λ_d), the UV/vis experiments unambiguously show that the deactivation step is not inhibited. Comparing the two uncaging quantum yields ($\varphi_{o\text{-nitro-puro/cinnoline}}$ = 2.5%, $\varphi_{\text{DEACM-OSO3}^-/\text{DEACM-OH}}$ = 0.1%) reveals that the deactivation of nitro-puromycin is favored over the photolysis of DEACM-OSO_3^- (Figures S18b and S20b, approximately 33% deactivated within 3 min of illumination: 5 × 10^{15} photons/s). Further changes of the absorption around 370 nm confirm a subsequent side reaction of the *thio*-coumarin photoproduct. The remaining band is assigned to DEACM-OH (Scheme 4).

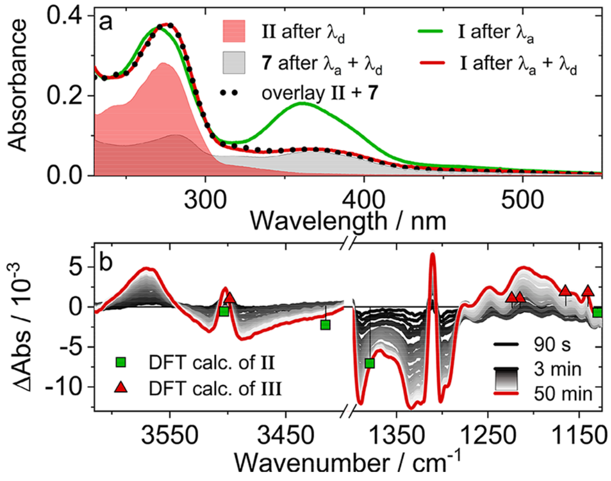
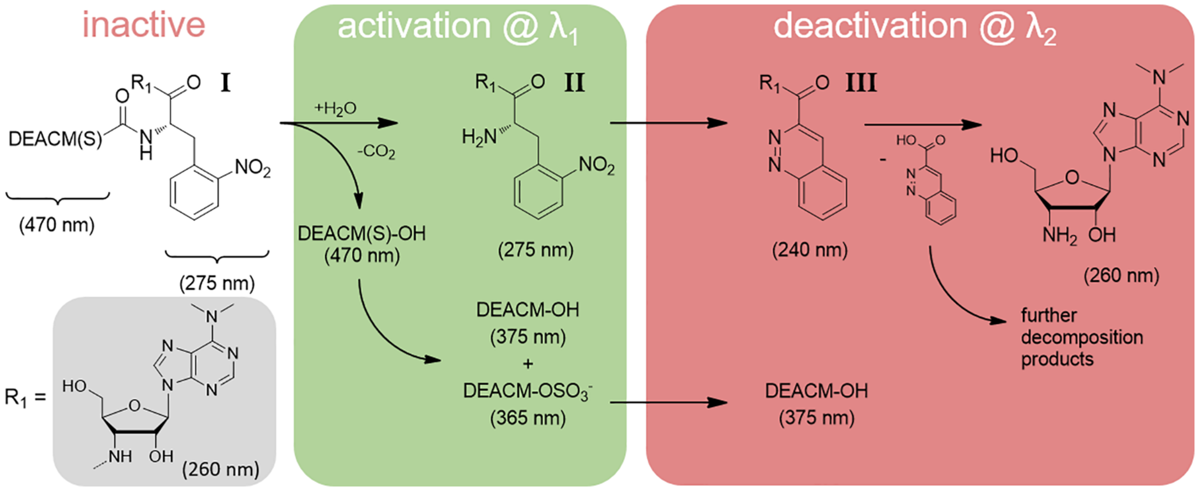


Figure 6. (a) UV/vis spectroscopic study of deactivation in 1:1 acetonitrile/water mixture. Starting material is the activated sample including side products (green line). After illumination at 365 nm for 1 h a spectrum (red line) arises, which can be reconstructed by the overlay (black dotted line) of illuminated *o*-nitro-puromycin (**II**) (red area) and the illuminated *thio*-DEACM with 365 nm for 1 h (gray area). (b) Difference IR-spectra measured every 60 s in DMSO of *o*-nitro-puromycin (**II**) including side product with illumination at 365 nm evolving from the black (zero line: no illumination of the sample) to the red line (photoproducts of the deactivation step: including cinnoline product **III**). DFT-calculated vibrational modes as sticks of *o*-nitro-puromycin (**II**) (squares) and cinnoline product **III** (triangles).

Additionally, the loss of the nitro-functionality of the *o*NB moiety at 1390 cm^{-1} and the symmetric stretch mode of the amino group at 3516 cm^{-1} confirm the deactivation of the central functional group of puromycin (Figure 6b). The formation of the cyclization product can be verified by the positive bands at 1139 and 1165 cm^{-1} that can be assigned to vibrational modes of the cinnoline moiety including the N=N double bond in accordance to DFT calculated modes (Figure 6b). For the presumed deactivation mechanism, water has to be eliminated in the last step, which is supported by the positive absorption change at 3569 cm^{-1}. A double uncaging can be carried out directly with 365 nm (Figures S22 and S23). Analysis of the absorption changes during the first 8 min reveals that both the activation and deactivation step occur simultaneously to a certain extent.

We used the newly developed TPOM puromycin to control protein synthesis in a cell-free coupled transcription−translation system (Figure 7). The expression inhibition induced by the released *o*-nitro-derivative of puromycin was monitored using shifted green fluorescent protein (sGFP) in a microplate reader. We monitored the sGFP expression as a function of increased concentrations of puromycin (Figure S27a) and *o*-

Scheme 4. Reaction Sequence Summary[a]



[a]Doubly caged Puromycin (*thio*-DEACM-nitro-puromycin **I**) is inactivated by thio-DEACM. After illumination at 470 nm the coumarin cage is removed, while the biologically active *o*-nitro-puromycin (**II**) is regenerated (activation step: green area). In a parallel pathway, two other photoproducts are produced as byproducts after photolysis of *thio*-DEACM: DEACM-OH and DEACM-OSO_3^-. In the deactivation pathway, *o*-nitro-puromycin (**II**) undergoes a photocyclization, resulting in a cinnoline product **III**, which is inactive due to the missing amino group (deactivation step: red area). Furthermore, the cinnoline moiety splits from the remaining puromycin scaffold. In this reaction route, the byproduct DEACM-OSO_3^- from the first step vanishes, leaving only the DEACM-OH behind.

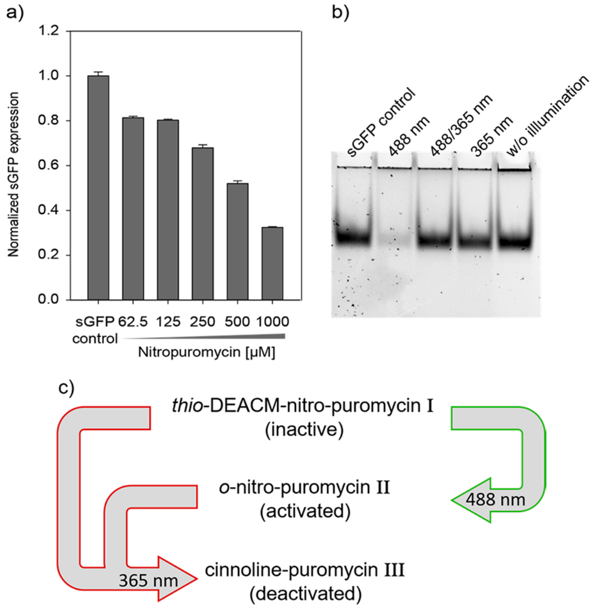


Figure 7. (a) Inhibition of sGFP expression with increasing o-nitro-puromycin (**II**) concentration. 1 mM *o*-nitro-puromycin (**II**) results in a 3-fold inhibition of sGFP expression. (b and c) *thio*-DEACM-nitro-puromycin (**I**) was illuminated, respectively with LEDs at 365 nm, 365/488 nm, and 488 nm, and the effects of the illumination wavelength were tested by native PAGE analysis of sGFP expression and compared with non-illuminated *thio*-DEACM-nitro-puromycin (**I**) (w/o illumination) and sGFP control sample without addition of inhibitor (sGFP control).

nitro-puromycin analog (**II**) concentration, resulting in a 5-fold and 3-fold inhibition, respectively (Figure 7a). Apparently, the structural modification of conventional puromycin by introducing the NO_2-group and by removing the *para*-methoxy group diminishes the inhibition activity of (sGFP) translation. However, the deactivation by protecting the primary α-amino group with *thio*-DEACM to the carbamate (**I**) was successful, since sGFP expression levels were comparable to those without any inhibitor addition. This and the effect of illumination at 488 and 365 nm were examined by using native PAGE analysis (Figure 7b), since *thio*-DEACM and sGFP absorption curves overlap justifying the fluorescence decay of this sample (Figure S27b).

In the next step, upon laser irradiation at 488 nm *thio*-DEACM was cleaved releasing *o*-nitro-puromycin (**II**) that inhibits almost 80% of the initial sGFP expression. In contrast, *thio*-DEACM-nitro-puromycin (**I**) and *o*-nitro-puromycin (**II**) which were deactivated by a 365 nm and 488/365 nm irradiation, respectively, did not show inhibition of the sGFP expression (Figures 7b and S27b). These results are in accordance with our above-mentioned spectroscopic and analytical investigations. Thus, *thio*-DEACM-nitro-puromycin (**I**) shows no inhibition effect and can be activated through 488 nm irradiation releasing the active *o*-nitro-puromycin (**II**). This compound can be again deactivated by a subsequent irradiation at 365 nm releasing the photocyclized cinnoline puromycin derivative. The same observation was made, when *thio*-DEACM-nitro-puromycin **I** was illuminated at 365 nm directly (Figure 7b and c). An obvious direct interpretation is that at 365 nm the cleavage of *thio*-DEACM and the cyclization to the cinnoline ring are simultaneously occurring events.

■ CONCLUSION

Here, we designed and probed TPOM as a new method to release and then destroy the function of a low molecular weight effector molecule in a biological system. A part of the active structure of the target molecule is synthetically modified to a second photoresponsive scaffold that undergoes a structural change to one or more inactive molecules upon wavelength selective irradiation. To demonstrate this dual function, we synthesized *thio*-DEACM-nitro-puromycin (**I**), a double photoresponsive analog of the frequently used aminonucleoside antibiotic puromycin. To achieve the required chromatic orthogonality, several 4-hydroxycoumarin derivatives were synthesized and screened. To report the successfully photocyclization of *o*-nitro-puromycin **II**, a laser-coupled ^{1}H NMR experiment was performed showing a light-induced formation of the cinnoline derivative followed by the release of puromycin aminonucleoside (Scheme 4). The successful activation/deactivation upon blue light uncaging of *thio*-DEACM and subsequent near-UV-induced photocyclization of *o*-nitro-puromycin analog **II** were confirmed by IR-measurements and DFT calculations. Furthermore, steady-state spectroscopy revealed a side path of coumarin byproducts. Finally, with *in vitro* transcription−translation experiments we examined the controlling ability of the translation inhibition effect of *thio*-DEACM-nitro-puromycin **I**. The expression inhibition was monitored using shifted green fluorescent protein (sGFP). The cleavage of *thio*-DEACM released the active *o*-nitro-puromycin **II**, which can be in turn deactivated by irradiating with λ_d. Moreover, direct deactivation of *thio*-DEACM-nitro-puromycin **I** was also observed. The introduction of the Two-PPGs-One-Molecule method for restricting the activity of target biomolecules is a promising strategy and a significant addition to the continuously growing toolbox of spatiotemporally photocontrolled elements for biological applications.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.1c02817.

> Synthesis procedures, characterization spectra, photolysis control experiments and details of additional spectroscopic investigation of uncaging processes (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Josef Wachtveitl − *Institute of Physical and Theoretical Chemistry, Goethe-University Frankfurt, Frankfurt am Main 60438, Germany;* orcid.org/0000-0002-8496-8240; Email: wveitl@theochem.uni-frankfurt.de

Harald Schwalbe − *Institute of Organic Chemistry and Chemical Biology, Goethe-University Frankfurt, Frankfurt am Main 60438, Germany;* orcid.org/0000-0001-5693-7909; Email: schwalbe@nmr.uni-frankfurt.de

Authors

Isam Elamri − *Institute of Organic Chemistry and Chemical Biology, Goethe-University Frankfurt, Frankfurt am Main 60438, Germany*

Chahinez Abdellaoui − *Institute of Physical and Theoretical Chemistry, Goethe-University Frankfurt, Frankfurt am Main 60438, Germany*

Jasleen Kaur Bains − *Institute of Organic Chemistry and Chemical Biology, Goethe-University Frankfurt, Frankfurt am Main 60438, Germany*

Katharina Felicitas Hohmann − *Institute of Organic Chemistry and Chemical Biology, Goethe-University Frankfurt, Frankfurt am Main 60438, Germany*

Santosh Lakshmi Gande − *Institute of Organic Chemistry and Chemical Biology, Goethe-University Frankfurt, Frankfurt am Main 60438, Germany*

Elke Stirnal − *Institute of Organic Chemistry and Chemical Biology, Goethe-University Frankfurt, Frankfurt am Main 60438, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.1c02817

Author Contributions

[§]I.E. and C.A. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Anna Wacker, Dr. Jan Ferner and Christoph Kaiser for helpful discussions. This work has been supported by GRK 1986 “CLiC” and by the DFG large scale equipment programme for the purchase of lasers coupled to NMR spectrometers (project number: 280251937). Work at BMRZ is supported by the state of Hesse.

■ REFERENCES

(1) Klán, P.; Šolomek, T.; Bochet, C. G.; Blanc, A.; Givens, R.; Rubina, M.; Popik, V.; Kostikov, A.; Wirz, J. Photoremovable Protecting Groups in Chemistry and Biology: Reaction Mechanisms and Efficacy. *Chem. Rev.* **2013**, *113* (1), 119−191.

(2) Brieke, C.; Rohrbach, F.; Gottschalk, A.; Mayer, G.; Heckel, A. Light-Controlled Tools. *Angew. Chem., Int. Ed.* **2012**, *51* (34), 8446−8476.

(3) Hansen, M. J.; Velema, W. A.; Lerch, M. M.; Szymanski, W.; Feringa, B. L. Wavelength-Selective Cleavage of Photoprotecting Groups: Strategies and Applications in Dynamic Systems. *Chem. Soc. Rev.* **2015**, *44* (11), 3358−3377.

(4) Bochet, C. G. Photolabile Protecting Groups and Linkers. *J. Chem. Soc. Perkin 1* **2002**, *2* (2), 125−142.

(5) Corrie, J. E. T.; Furuta, T.; Givens, R. S.; Yousef, A. L.; Goeldner, M. *Photoremovable Protecting Groups Used for the Caging of Biomolecules*; WILEY-VCH: Weinheim, 2005.

(6) Grün, J. T.; Hennecker, C.; Klötzner, D. P.; Harkness, R. W.; Bessi, I.; Heckel, A.; Mittermaier, A. K.; Schwalbe, H. Conformational Dynamics of Strand Register Shifts in DNA G-Quadruplexes. *J. Am. Chem. Soc.* **2020**, *142* (1), 264−273.

(7) Wenter, P.; Fürtig, B.; Hainard, A.; Schwalbe, H.; Pitsch, S. A Caged Uridine for the Selective Preparation of an RNA Fold and Determination of Its Refolding Kinetics by Real-Time NMR. *ChemBioChem* **2006**, *7* (3), 417−420.

(8) Fürtig, B.; Richter, C.; Schell, P.; Wenter, P.; Pitsch, S.; Schwalbe, H. NMR-Spectroscopic Characterisation of Phosphodiester Bond Cleavage Catalysed by the Minimal Hammerhead Ribozyme. *RNA Biol.* **2008**, *5* (1), 41−48.

(9) Bardhan, A.; Deiters, A. Development of Photolabile Protecting Groups and Their Application to the Optochemical Control of Cell Signaling. *Curr. Opin. Struct. Biol.* **2019**, *57*, 164−175.

(10) Slanina, T.; Shrestha, P.; Palao, E.; Kand, D.; Peterson, J. A.; Dutton, A. S.; Rubinstein, N.; Weinstain, R.; Winter, A. H.; Klán, P. In Search of the Perfect Photocage: Structure-Reactivity Relationships in Meso-Methyl BODIPY Photoremovable Protecting Groups. *J. Am. Chem. Soc.* **2017**, *139* (42), 15168−15175.

(11) Agarwal, H. K.; Janicek, R.; Chi, S. H.; Perry, J. W.; Niggli, E.; Ellis-Davies, G. C. R. Calcium Uncaging with Visible Light. *J. Am. Chem. Soc.* **2016**, *138* (11), 3687−3693.

(12) Umeda, N.; Takahashi, H.; Kamiya, M.; Ueno, T.; Komatsu, T.; Terai, T.; Hanaoka, K.; Nagano, T.; Urano, Y. Boron Dipyrromethene as a Fluorescent Caging Group for Single-Photon Uncaging with Long-Wavelength Visible Light. *ACS Chem. Biol.* **2014**, *9* (10), 2242−2246.

(13) Griepenburg, J. C.; Rapp, T. L.; Carroll, P. J.; Eberwine, J.; Dmochowski, I. J. Ruthenium-Caged Antisense Morpholinos for

Regulating Gene Expression in Zebrafish Embryos. *Chem. Sci.* **2015**, *6* (4), 2342−2346.

(14) Rapp, T. L.; Dmochowski, I. J. Ruthenium-Cross-Linked Hydrogels for Rapid, Visible-Light Protein Release. *Methods Enzymol.* **2019**, *624*, 151−166.

(15) Sitkowska, K.; Hoes, M. F.; Lerch, M. M.; Lameijer, L. N.; Van Der Meer, P.; Szymański, W.; Feringa, B. L. Red-Light-Sensitive BODIPY Photoprotecting Groups for Amines and Their Biological Application in Controlling Heart Rhythm. *Chem. Commun.* **2020**, *56* (41), 5480−5483.

(16) Herzig, L. M.; Elamri, I.; Schwalbe, H.; Wachtveitl, J. Light-Induced Antibiotic Release from a Coumarin-Caged Compound on the Ultrafast Timescale. *Phys. Chem. Chem. Phys.* **2017**, *19* (22), 14835−14844.

(17) Elamri, I.; Heumüller, M.; Herzig, L. M.; Stirnal, E.; Wachtveitl, J.; Schuman, E. M.; Schwalbe, H. A New Photocaged Puromycin for an Efficient Labeling of Newly Translated Proteins in Living Neurons. *ChemBioChem* **2018**, *19* (23), 2458−2464.

(18) Piant, S.; Bolze, F.; Specht, A. Two-Photon Uncaging, from Neuroscience to Materials. *Opt. Mater. Express* **2016**, *6* (5), 1679.

(19) (a) Becker, Y.; Unger, E.; Fichte, M. A. H.; Gacek, D. A.; Dreuw, A.; Wachtveitl, J.; Walla, P. J.; Heckel, A. A Red-Shifted Two-Photon-Only Caging Group for Three-Dimensional Photorelease. *Chem. Sci.* **2018**, *9* (10), 2797−2802. (b) Sikder, A.; Banerjee, M.; Singha, T.; Mondal, S.; Datta, P. K.; Anoop, A.; Singh, N. D. P. A Natural Alkaloid, *β*-Carboline, as a One- And Two-Photon Responsive Fluorescent Photoremovable Protecting Group: Sequential Release of the Same or Different Carboxylic Acids. *Org. Lett.* **2020**, *22* (17), 6998−7002.

(20) Rodrigues-Correia, A.; Weyel, X. M. M.; Heckel, A. Four Levels of Wavelength-Selective Uncaging for Oligonucleotides. *Org. Lett.* **2013**, *15* (21), 5500−5503.

(21) Fournier, L.; Gauron, C.; Xu, L.; Aujard, I.; Le Saux, T.; Gagey-Eilstein, N.; Maurin, S.; Dubruille, S.; Baudin, J. B.; Bensimon, D.; Volovitch, M.; Vriz, S.; Jullien, L. A Blue-Absorbing Photolabile Protecting Group for in Vivo Chromatically Orthogonal Photoactivation. *ACS Chem. Biol.* **2013**, *8* (7), 1528−1536.

(22) Neveu, P.; Aujard, I.; Benbrahim, C.; Le Saux, T.; Allemand, J. F.; Vriz, S.; Bensimon, D.; Jullien, L. A Caged Retinoic Acid for One- and Two-Photon Excitation in Zebrafish Embryos. *Angew. Chem., Int. Ed.* **2008**, *47* (20), 3744−3746.

(23) Noguchi, M.; Skwarczynski, M.; Prakash, H.; Hirota, S.; Kimura, T.; Hayashi, Y.; Kiso, Y. Development of Novel Water-Soluble Photocleavable Protective Group and Its Application for Design of Photoresponsive Paclitaxel Prodrugs. *Bioorg. Med. Chem.* **2008**, *16* (10), 5389−5397.

(24) Basa, P. N.; Barr, C. A.; Oakley, K. M.; Liang, X.; Burdette, S. C. Zinc Photocages with Improved Photophysical Properties and Cell Permeability Imparted by Ternary Complex Formation. *J. Am. Chem. Soc.* **2019**, *141* (30), 12100−12108.

(25) Deiters, A.; Groff, D.; Ryu, Y.; Xie, J.; Schultz, P. G. A Genetically Encoded Photocaged Tyrosine. *Angew. Chem., Int. Ed.* **2006**, *45* (17), 2728−2731.

(26) Lemke, E. A.; Summerer, D.; Geierstanger, B. H.; Brittain, S. M.; Schultz, P. G. Control of Protein Phosphorylation with a Genetically Encoded Photocaged Amino Acid. *Nat. Chem. Biol.* **2007**, *3* (12), 769−772.

(27) Nathans, D. Puromyicn Inhibition of Protein Synthesis: Incorporation of Puromycin. *Proc. Natl. Acad. Sci. U. S. A.* **1964**, *51*, 585−592.

(28) Yarmolinsky, M. B.; Haba, G. L. D. L. Inhibition by Puromycin of Amino Acid Incorporation Into Protein. *Proc. Natl. Acad. Sci. U. S. A.* **1959**, *45* (12), 1721−1729.

(29) Campuzano, S.; Modolell, J. Hydrolysis of GTP on Elongation Factor Tu·ribosome Complexes Promoted by 2′(3′)-O-L-Phenylalanyladenosine. *Proc. Natl. Acad. Sci. U. S. A.* **1980**, *77*, 905−909.

(30) Miyamoto-Sato, E.; Nemoto, N.; Kobayashi, K.; Yanagawa, H. Specific Bonding of Puromycin to Full-Length Protein at the C-Terminus. *Nucleic Acids Res.* **2000**, *28* (5), 1176−1182.

(31) Glock, C.; Heumüller, M.; Schuman, E. M. MRNA Transport & Local Translation in Neurons. *Curr. Opin. Neurobiol.* **2017**, *45*, 169−177.

(32) Kohl-Landgraf, J.; Buhr, F.; Lefrancois, D.; Mewes, J.-M.; Schwalbe, H.; Dreuw, A.; Wachtveitl, J. Mechanism of the Photoinduced Uncaging Reaction of Puromycin Protected by a 6-Nitroveratryloxycarbonyl Group. *J. Am. Chem. Soc.* **2014**, *136* (9), 3430−3438.

(33) Buhr, F.; Kohl-Landgraf, J.; Tomdieck, S.; Hanus, C.; Chatterjee, D.; Hegelein, A.; Schuman, E. M.; Wachtveitl, J.; Schwalbe, H. Design of Photocaged Puromycin for Nascent Polypeptide Release and Spatiotemporal Monitoring of Translation. *Angew. Chem., Int. Ed.* **2015**, *54* (12), 3717−3721.

(34) Kammari, L.; Šolomek, T.; Ngoy, B. P.; Heger, D.; Klán, P. Orthogonal Photocleavage of a Monochromophoric Linker. *J. Am. Chem. Soc.* **2010**, *132* (33), 11431−11433.

(35) Paul, A.; Bera, M.; Gupta, P.; Singh, N. D. P. O -Hydroxycinnamate for Sequential Photouncaging of Two Different Functional Groups and Its Application in Releasing Cosmeceuticals. *Org. Biomol. Chem.* **2019**, *17* (33), 7689−7693.

(36) Kneuttinger, A. C.; Winter, M.; Simeth, N. A.; Heyn, K.; Merkl, R.; König, B.; Sterner, R. Artificial Light Regulation of an Allosteric Bienzyme Complex by a Photosensitive Ligand. *ChemBioChem* **2018**, *19* (16), 1750−1757.

(37) Peters, F. B.; Brock, A.; Wang, J.; Schultz, P. G. Photocleavage of the Polypeptide Backbone by 2-Nitrophenylalanine. *Chem. Biol.* **2009**, *16* (2), 148−152.

(38) Kirsch, N.; Kusunose, N; Aikawa, J.-i. Synthesis of N-Acetylglucosaminyl Asparagine-Substituted Puromycin Analogues. *Bioorg. Med. Chem.* **1995**, *3* (12), 1631−1636.

(39) Du, S.; Wang, D.; Lee, J. S.; Peng, B.; Ge, J.; Yao, S. Q. Cell Type-Selective Imaging and Profiling of Newly Synthesized Proteomes by Using Puromycin Analogues. *Chem. Commun.* **2017**, *53* (60), 8443−8446.

(40) Ge, J.; Zhang, C. W.; Ng, X. W.; Peng, B.; Pan, S.; Du, S.; Wang, D.; Li, L.; Lim, K. L.; Wohland, T.; Yao, S. Q. Puromycin Analogues Capable of Multiplexed Imaging and Profiling of Protein Synthesis and Dynamics in Live Cells and Neurons. *Angew. Chem., Int. Ed.* **2016**, *55* (16), 4933−4937.

(41) Gilbert, C. L. K.; Lisek, C. R.; White, R. L.; Gumina, G. Synthesis of L,L-Puromycin. *Tetrahedron* **2005**, *61* (35), 8339−8344.

(42) Mizusawa, K.; Abe, K.; Sando, S.; Aoyama, Y. Synthesis of Puromycin Derivatives with Backbone-Elongated Substrates and Associated Translation Inhibitory Activities. *Bioorg. Med. Chem.* **2009**, *17* (6), 2381−2387.

(43) Blanc, A.; Bochet, C. G. Wavelength-Controlled Orthogonal Photolysis of Protecting Groups. *J. Org. Chem.* **2002**, *67* (16), 5567−5577.

(44) Bochet, C. G. Wavelength-Selective Cleavage of Photolabile Protecting Groups. *Tetrahedron Lett.* **2000**, *41* (33), 6341−6346.

(45) Bochet, C. G. Orthogonal Photolysis of Protecting Groups. *Angew. Chem., Int. Ed.* **2001**, *40* (11), 2071.

(46) Fournier, L.; Aujard, I.; Le Saux, T.; Maurin, S.; Beaupierre, S.; Baudin, J. B.; Jullien, L. Coumarinylmethyl Caging Groups with Redshifted Absorption. *Chem. - Eur. J.* **2013**, *19* (51), 17494−17507.

(47) Vorob'ev, A. Y.; Dranova, T. Y.; Moskalensky, A. E. Photolysis of Dimethoxynitrobenzyl-"Caged" Acids Yields Fluorescent Products. *Sci. Rep.* **2019**, *9* (1), 1−7.

(48) Beaudoin, D.; Wuest, J. D. Dimerization of Aromatic C-Nitroso Compounds. *Chem. Rev.* **2016**, *116* (1), 258−286.

(49) Kühn, T.; Schwalbe, H. Monitoring the Kinetics of Ion-Dependent Protein Folding by Time-Resolved NMR Spectroscopy at Atomic Resolution. *J. Am. Chem. Soc.* **2000**, *122* (26), 6169−6174.

(50) Schade, B.; Hagen, V.; Schmidt, R.; Herbrich, R.; Krause, E.; Eckardt, T.; Bendig, J. Deactivation Behavior and Excited-State Properties of (Coumarin-4- Yl)Methyl Derivatives. 1. Photocleavage of (7-Methoxycoumarin-4-Yl)Methyl- Caged Acids with Fluorescence Enhancement. *J. Org. Chem.* **1999**, *64* (25), 9109−9117.

5.6 **Research article:** ^{1}H, ^{13}C and ^{15}N chemical shift assignment of the stem-loops 5b + c from the 5′-UTR of SARS-CoV-2

Klara R. Mertinkus*, J. Tassilo Grün*, Nadide Altincekic, Jasleen Kaur Bains, Betül Ceylan, Jan-Peter Ferner, Lucio Frydman, Boris Fürtig, Martin Hengesbach, Katharina F. Hohmann, Daniel Hymon, Jihyun Kim, Božana Knezic, Mihajlo Novakovic, Andreas Oxenfarth, Stephen A. Peter, Nusrat S. Qureshi, Christian Richter, Tali Scherf, Andreas Schlundt, Robbin Schnieders, Harald Schwalbe, Elke Stirnal, Alexey Sudakov, Jennifer Vögele, Anna Wacker, Julia E. Weigand, Julia Wirmer-Bartoschek, Maria A. Wirtz Martin and Jens Wöhnert; *Biomolecular NMR Assignments* **16**, 17-25 (2022)

* These authors contributed equally to this work.

This article presents NMR studies of the stem-loops 5 (SL5) located in the 5'-UTR of the SARS-CoV-2 genome. The ^{1}H, ^{13}C, and ^{15}N resonance assignments of the stem-loops SL5b and 5SLc are presented as a basis for further in-depth structural and ligand screening studies.

The project was designed and coordinated by the Covid-19 NMR consortium. K. R. Mertinkus and J. T. Grün assigned the NMR spectra and wrote the manuscript. The author of this thesis was part of the Covid-19 NMR consortium and the RNA production team that prepared and purified various isotope-labeled RNA samples for NMR spectroscopy.

https://doi.org/10.1007/s12104-021-10053-4

Biomolecular NMR Assignments (2022) 16:17–25
https://doi.org/10.1007/s12104-021-10053-4

ARTICLE

^{1}H, ^{13}C and ^{15}N chemical shift assignment of the stem-loops 5b + c from the 5′-UTR of SARS-CoV-2

Klara R. Mertinkus[1,3] · J. Tassilo Grün[4] · Nadide Altincekic[1,3] · Jasleen Kaur Bains[1,3] · Betül Ceylan[1,3] · Jan-Peter Ferner[1,3] · Lucio Frydman[4] · Boris Fürtig[1,3] · Martin Hengesbach[1,3] · Katharina F. Hohmann[1,3] · Daniel Hymon[1,3] · Jihyun Kim[4] · Božana Knezic[1,3] · Mihajlo Novakovic[4,7] · Andreas Oxenfarth[1,3] · Stephen A. Peter[8] · Nusrat S. Qureshi[6] · Christian Richter[1,3] · Tali Scherf[5] · Andreas Schlundt[2,3] · Robbin Schnieders[1,3,9] · Harald Schwalbe[1,3] · Elke Stirnal[2,3] · Alexey Sudakov[1,3] · Jennifer Vögele[2,3] · Anna Wacker[1,3] · Julia E. Weigand[8] · Julia Wirmer-Bartoschek[1,3] · Maria A. Wirtz Martin[1,3] · Jens Wöhnert[2,3]

Received: 4 August 2021 / Accepted: 16 October 2021 / Published online: 18 February 2022

Abstract

The ongoing pandemic of the respiratory disease COVID-19 is caused by the SARS-CoV-2 (SCoV2) virus. SCoV2 is a member of the Betacoronavirus genus. The 30 kb positive sense, single stranded RNA genome of SCoV2 features 5′- and 3′-genomic ends that are highly conserved among Betacoronaviruses. These genomic ends contain structured *cis*-acting RNA elements, which are involved in the regulation of viral replication and translation. Structural information about these potential antiviral drug targets supports the development of novel classes of therapeutics against COVID-19. The highly conserved branched stem-loop 5 (SL5) found within the 5′-untranslated region (5′-UTR) consists of a basal stem and three stem-loops, namely SL5a, SL5b and SL5c. Both, SL5a and SL5b feature a 5′-UUUCGU-3′ hexaloop that is also found among Alphacoronaviruses. Here, we report the extensive ^{1}H, ^{13}C and ^{15}N resonance assignment of the 37 nucleotides (nts) long sequence spanning SL5b and SL5c (SL5b + c), as basis for further in-depth structural studies by solution NMR spectroscopy.

Keywords SARS-CoV-2 · 5′-UTR · SL5b + c · SL5b · SL5c · Solution NMR spectroscopy · COVID19-NMR

Biological context

The ongoing global pandemic associated with the coronavirus disease (COVID-19) is caused by the human Betacoronavirus SARS-CoV-2 (SCoV2), a close relative of the

Klara R. Mertinkus and J. Tassilo Grün have contributed equally to this work.

✉ Harald Schwalbe
covid19-nmr@dlist.server.uni-frankfurt.de;
schwalbe@nmr.uni-frankfurt.de

[1] Institute for Organic Chemistry and Chemical Biology, Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue-Str. 7, 60438 Frankfurt, Germany

[2] Institute for Molecular Biosciences, Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue-Str. 7, 60438 Frankfurt, Germany

[3] Center for Biomolecular Magnetic Resonance (BMRZ), Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue-Str. 7, 60438 Frankfurt, Germany

[4] Department of Chemical and Biological Physics, Weizmann Institute of Science, Herzl St. 234, 760001 Rehovot, Israel

[5] Department of Chemical Research Support, Weizmann Institute of Science, Herzl St. 234, 760001 Rehovot, Israel

[6] Present Address: EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany

[7] Present Address: Institute for Biochemistry, ETH Zürich, Hönggerbergring 64, 8093 Zürich, Switzerland

[8] Department of Biology, Technical University of Darmstadt, Schnittspahnstr. 10, 64287 Darmstadt, Germany

[9] Present Address: Deutero GmbH, Am Ring 29, 56288 Kastellaun, Germany

severe acute respiratory syndrome (SARS) causing agent SARS-CoV. Betacoronaviruses have large positive sense, single-stranded RNA genomes, with highly conserved 5′- and 3′-untranslated regions (UTRs) that do not code for viral proteins. These structured UTRs are highly conserved among Betacoronaviruses and are important for the replication, balanced transcription of subgenomic mRNAs and translation of viral proteins. (Yang and Leibowitz 2015) So far, most efforts for the development of new antiviral drugs target the proteins of SARS-CoV-2. The structured regulatory elements of the approx. 30,000 nucleotides (nts) long RNA genome remain unexploited as potential target sites for antiviral drugs. Between different Coronaviruses, the sequence of the individual elements varies, but their secondary structures reveal remarkably high conservation, suggesting a critical importance for viral viability and pathogenesis. (Madhugiri et al. 2016) Until now, a large number of sequence-based computational predictions and different chemical probing approaches have been reported to map the architecture of these viral RNA elements. (Zhao et al. 2020; Huston et al. 2021; Lan et al. 2021; Manfredonia and Incarnato 2021; Rangan et al. 2021) However, to establish the viral RNA as an antiviral drug target, high-resolution structural data are important that can also visualize structural dynamics and tertiary structure interactions. In response to the pandemic situation, the international COVID19-NMR initiative (https://covid19-NMR.de) has set the goal to provide this information by solution NMR, in order to initiate and guide structure-based drug screening, design and synthesis. The structured parts of the SARS-CoV-2 genome have been divided into fragments in a ‘divide and conquer’ approach, allowing us to determine the secondary structures of these RNA elements. (Wacker et al. 2020) Further, fragment screening campaigns demonstrated that the RNA structural elements can be targeted differentially, revealing low micromolar binding affinities specific to molecules of low molecular weight. (Sreeramulu et al. 2021).

An intriguing example of an RNA regulatory element from SARS-CoV-2 is the comparably large structural element of SL5 spanning nts 149–265. The entire SL5 element consists of four helices, joining three sub-elements with stem loop motifs to the SL5 basal stem by a four-way junction. These sub-elements are termed SLs 5a, 5b and 5c. Interestingly, SL5 is forming junction-connected elements in the genomes of both Alpha- and Betacoronaviruses. (Madhugiri et al. 2014, 2016) The regulatory function of SL5 has been linked to maintaining efficient viral replication. (Chen and Olsthoorn 2010; Guan et al. 2011) In SL5b, an apical hexaloop sequence is found that is identical to the loop in SL5a (5'-UUUCGU-3′). Similar loop sequences with 5′-UUYCGU-3′ motifs can also be found in members of the Alphacoronavirus genus, suggesting a conserved function e.g. in viral packaging. (Masters 2019) Interestingly, currently available sequencing data for new SCoV-2 variants emerging since March 2020 show that the 5′-UUUCGU-3′ loop in SL5a remains conserved compared to the original virus strain, while a C241U mutation resulting in a 5′-UUU UGU-3′ loop appeared in SL5b.

In SCoV2, SL5 contains the first 29 nts of the open reading frame ORF 1a/b that codes for nsp1, the first of the non-structural proteins (Fig. 1) including the start codon A266 to G268, suggesting that the complex structural arrangement in SL5 is important for translation initiation. The SL5b stem-loop contains nucleotides 228 to 252 (25 nts), while the downstream located SL5c consists of 10 nts, 253 to 262. We report here the NMR chemical shift assignments

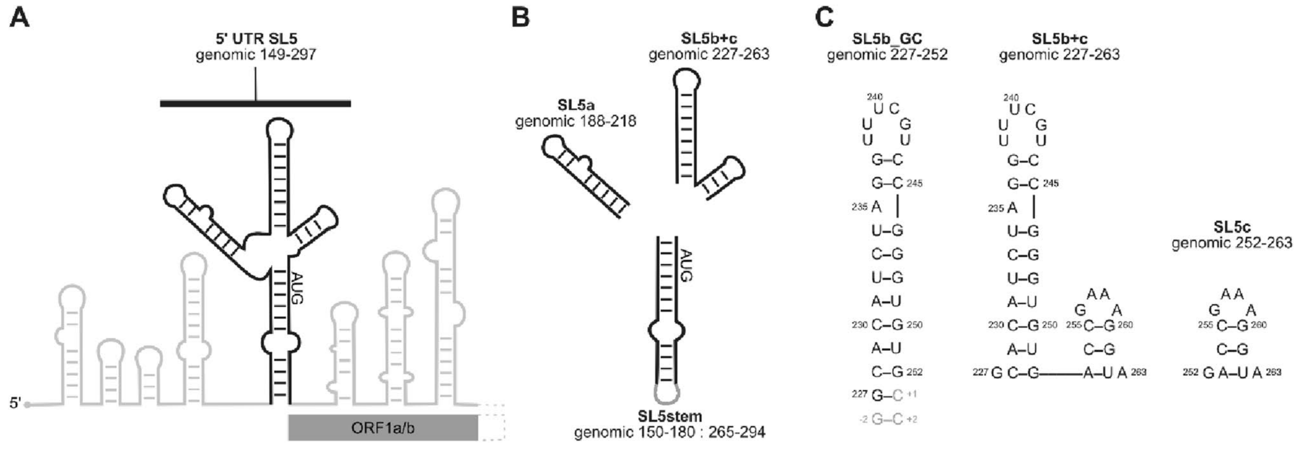


Fig. 1 **A** Schematic overview of 5′-UTR RNA elements of the SCoV2 genome. Black: SL5 element; AUG start codon and the 5′-terminal structural elements of the open reading frame ORF1a/b are highlighted in grey. **B** Elements used for the NMR-based divide-and-conquer approach. **C** Predicted secondary structures of RNA (sub-) elements used for the NMR chemical shift assignment of SL5b+c reported here. Genomic region, numbering and sample titles are given. **B/C** Black regions according to genomic sequence, grey regions contain stabilizing nucleotides. The actual investigated RNAs are represented by the sequences including the grey regions

for SL5b + c (nts 227–263) containing both stems, which was aided by assigning the isolated SL elements based on initial ^{1}H and ^{15}N assignments of all sub-elements of SL5 (a–c) and the basal stem. (Wacker et al. 2020) More recently, we reported the chemical shift assignments including ^{13}C chemical shifts for SL5a. (Schnieders et al. 2021).

Methods and experiments

Sample preparation

RNA synthesis for NMR experiments: For DNA template production, the sequences of SL5b + c (genomic nucleotides 227 to 263) and SL5b_GC, (5′-G-(genomic 227 to 252)-CC-3′) (Fig. 1C), together with the T7 promoter were generated by hybridization of complementary oligonucleotides and introduced into the *EcoR*I and *Nco*I sites of a plasmid, based on the pSP64 vector (Promega) encoding an HDV ribozyme (Schürer et al. 2002). RNAs were transcribed as HDV ribozyme fusions to obtain homogeneous 3′-ends. The recombinant vectors pHDV-5_SL5b + c and pHDV-5_SL5b_GC were transformed and amplified in the *Escherichia coli* strain DH5α. Plasmid-DNA was purified with a large scale DNA isolation kit (Gigaprep; Qiagen) according to the manufacturer's instructions and linearized with *Hind*III prior to *in-vitro* transcription by T7 RNA polymerase [P266L mutant, prepared as described in (Guillerez et al. 2005; Schnieders et al. 2020)]. 15 ml transcription reactions [20 mM DTT, 2 mM spermidine, 200 ng/µl template, 200 mM Tris/glutamate (pH 8.1), 40 mM $Mg(OAc)_2$, 12 mM NTPs, 32 µg/ml T7 RNA Polymerase, 20% DMSO (b + c) or 0% DMSO (b_GC)] were performed to obtain sufficient amount of RNA. Preparative transcription reactions (6 h at 37 °C and 70 rpm) were terminated by addition of 150 mM EDTA. SL5b + c and 5SL5b_GC RNAs were purified as follows: RNAs were precipitated with one volume of 2-propanol at − 20 °C overnight. RNA fragments were separated on 12–15% denaturing polyacrylamide (PAA) gels and visualized by UV shadowing at 254 nm. SL5b + c and SL5b_GC containing RNA bands were excised from the gel and then incubated for 30 min at − 80 °C, followed by 15 min at 65 °C. The elution was done overnight by passive diffusion into 0.3 M NaOAc, precipitated with EtOH and desalted via PD10 columns (GE Healthcare). Residual PAA was removed by reversed-phase HPLC using a Kromasil RP 18 column and a gradient of 0–40% 0.1 M triethylammonium acetate in acetonitrile. After freeze-drying of RNA-containing fractions and cation exchange by $LiClO_4$ precipitation (2% in acetone), the RNA was folded in water by heating to 80 °C followed by rapid cooling on ice. Buffer exchange into NMR buffer (95% H_2O/5% D_2O, 25 mM potassium phosphate buffer, pH 6.2, 50 mM potassium chloride) was performed using Vivaspin centrifugal concentrators (2 kDa molecular weight cut-off). The purity of SL5b + c and SL5b_GC was verified by denaturing PAA gel electrophoresis and homogeneous folding was monitored by native PAA gel electrophoresis, loading the same RNA concentration as used in NMR experiments. 100% D_2O samples were prepared by lyophilisation and redissolving in identical volumes of pure D_2O to keep the buffer salt concentration constant.

Using this protocol, three NMR samples of SL5b + c were prepared: a 350 µM uniformly ^{15}N- and two uniformly ^{13}C, ^{15}N-labelled samples (750 µM in buffer with 95% H_2O/5% D_2O and 430 µM in buffer with 100% D_2O). In addition, two uniformly ^{13}C, ^{15}N-labelled samples of SL5b_GC were prepared: an H_2O sample at a concentration of 510 µM in buffer with 95% H_2O/5% D_2O and a D_2O sample at a concentration of 300 µM in buffer with 100% D_2O. For the divide-and-conquer approach, the unlabelled single stem-loop 5c (Fig. 1C) was purchased from Horizon Discovery LTD (Cambridge, UK). The sample was processed by reversed-phase HPLC identical to the in-vitro transcribed samples. $LiClO_4$ precipitation, buffer exchange, folding check and sample preparation was performed as mentioned above (1.1 mM in NMR buffer with 95% H_2O/5% D_2O, 1.0 mM in NMR buffer with 100% D_2O).

NMR experiments

NMR experiments were carried out at the Weizmann Institute (WIS) using a Bruker AVIII 600 MHz NMR spectrometer equipped with a 5 mm, z-axis gradient ^{1}H [^{13}C, ^{15}N]-TCI prodigy probe and a Bruker AVANCE Neo 1 GHz spectrometer equipped with a 5 mm, z-axis gradient ^{1}H [^{13}C, ^{15}N]-TCI cryogenic probe and at the Center for Biomolecular Magnetic Resonance (BMRZ) at the Goethe University Frankfurt using Bruker NMR spectrometers from 600 to 800 MHz, which are equipped with AVANCE Neo, AVIIIHD, AVIII and AVI consoles and the following cryogenic probes: 5 mm, z-axis gradient ^{1}H [^{13}C, ^{31}P]-TCI cryogenic probe (600 MHz), 5 mm, z-axis gradient ^{1}H/^{19}F [^{13}C, ^{15}N]-TCI prodigy probe (600 MHz), 5 mm, z-axis gradient ^{1}H [^{13}C, ^{15}N]-TCI cryogenic probe (600 MHz), 5 mm, z-axis gradient ^{1}H [^{13}C, ^{15}N,^{31}P]-QCI cryogenic probe (700 MHz) and ^{13}C-optimized 5 mm, z-axis gradient ^{13}C [^{15}N, ^{1}H]-TXO cryogenic probe (800 MHz).

Experiments were performed in a temperature range spanning 274 to 298 K. NMR spectra were processed and analysed using Topspin (versions 3.6.2 to 4.1.1), and chemical shift assignment was conducted using Sparky. (Lee et al. 2015) NMR data were managed and archived using the platform LOGS (2020, version 2.1.54, Signals GmbH & Co KG, www.logs.repository.com). ^{1}H chemical shifts were referenced externally to DSS and ^{13}C and ^{15}N chemical shifts were indirectly referenced from the ^{1}H chemical shift as previously described. (Wishart et al. 1995).

Assignment and data deposition

The imino, aromatic and ribose resonances of the SL5b + c were assigned using a ^{13}C, ^{15}N-labelled SL5b_GC sample and an unlabelled SL5c model RNA (SL5c) (SI Fig. 1).

The assignment of SL5b_GC is described in the following using the experiments summarized in SI Table SL5b_GC. From the imino proton chemical shift assignment by ^{1}H, ^{15}N-TROSY (SI Fig. 2A), ^{1}H,^{1}H-NOESY (SI Fig. 2B) and HNN-COSY (SI Fig. 2C) spectra, the U-C2 and -C4 as well as G-C2 and -C6 could be assigned in the ^{1}H, ^{13}C-HNCO, and as well for aromatic carbon resonances U-H3/C6 and G-H1/C8 in the ^{1}H, ^{13}C-HCCNH (Fig. 2C). Using an ^{1}H,^{1}H-TOCSY (Fig. 2E) to selectively assign cytidine and uridine H5-H6 resonances and ^{1}H, ^{13}C-HSQC's (Fig. 2A, B, and F) for aromatic carbon resonances, the pyrimidine base C5-H5 and C6-H6 were obtained. Assignment of purine C8-H8 was aided by an ^{1}H,^{1}H-x-filter-NOESY (Fig. 2D). The latter experiment was also used to confirm assigned shifts for aromatic H6/H8 and H1′ found in a 3D-NOESY-HSQC and 3D-HCN. Additional carbons C4, C5 and C6 for adenosine were assigned using a 3D TROSY-HCCH-COSY. Nucleobase intra and sequential aromatic-to-ribose H1′ correlations were successfully detected for the predicted helical stem part from G-2 to U238 and from U243 to C + 2. The H1′ assignments were confirmed by a 3D-NOESY-HSQC leading to almost complete H1′ assignments with the help of a ^{1}H,^{13}C-HSQC for the H1′–C1′ region (Fig. 2A). From here, a ^{1}H,^{13}C-ct-HSQC and 3D-HCCH-TOCSY's with different mixing times gave further insight to the CH ribose resonance shifts. A canonical shift analysis for the sugar puckers of the almost completely assigned C1′ to C5′ resonances showed a C3′-endo conformation except for the bulge and loop nucleotides (Fig. 3). Additionally, chemical shift assignments for the nitrogens N1 or N9 could be assigned in the ^{1}H,^{15}N-HCN experiment.

The assignments obtained for the SL5b_GC sample were transferred to ^{1}H,^{15}N-TROSY, ^{1}H,^{1}H-NOESY, ^{1}H,^{13}C-HNCO and ^{1}H,^{13}C-HSQC spectra of SL5b + c (Table 1I to IV), showing a fit of the shifts of SL5b_GC from nucleotides

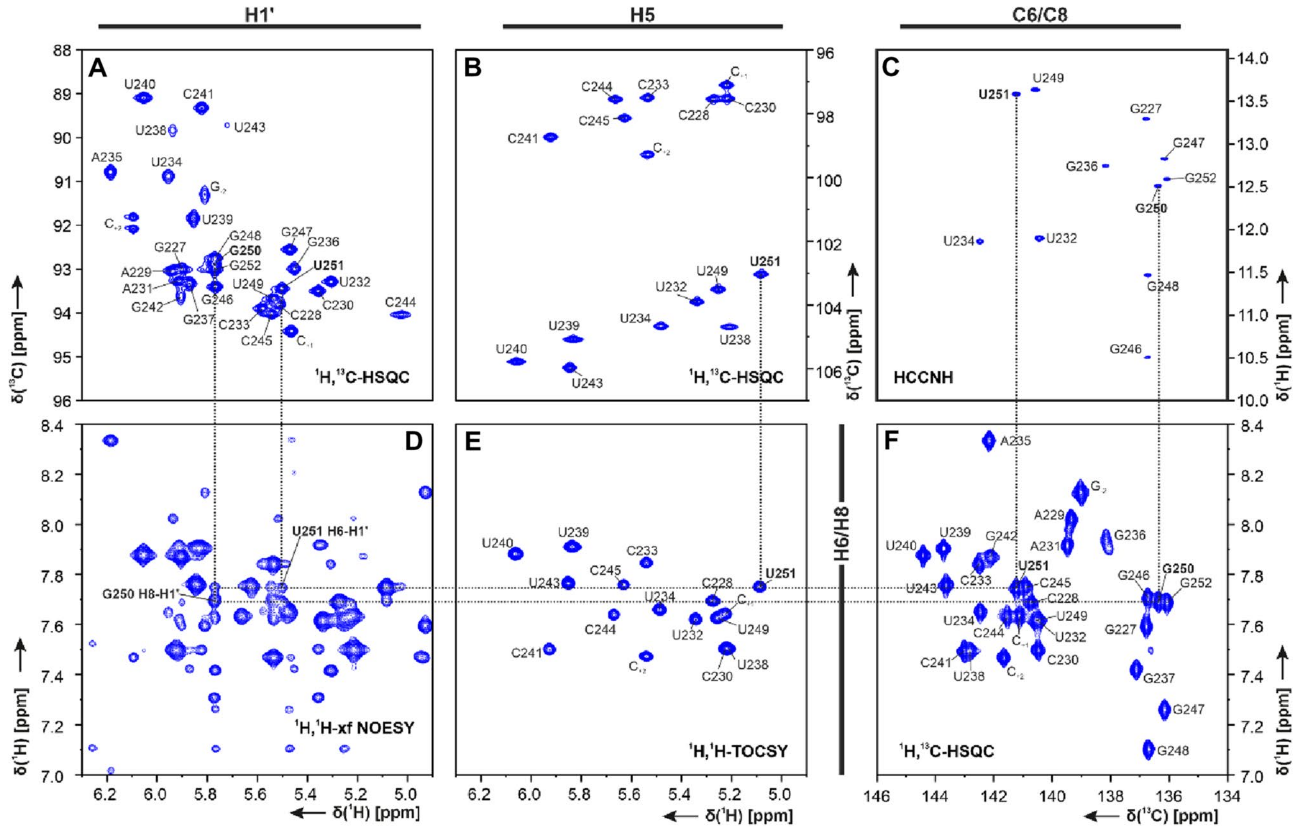


Fig. 2 Spectra of SL5b_GC, in NMR buffer in 95% H_2O/5% D_2O, 298 K: **A** ^{1}H,^{13}C-HSQC (C1′–H1′ region), **B** ^{1}H,^{13}C-HSQC (C5–H5 region), **C** HCCNH, **D** ^{1}H,^{1}H-xfilter NOESY, **E** ^{1}H,^{1}H-TOCSY and **F** ^{1}H,^{13}C-HSQC (C6–H6/C8–H8 region). Annotation of nucleobase assignment uses genomic numbering. Additional closing base pairs are annotated with '±x'. Dashed lines showing examples of ribose-to-aromatic atom relations for bases G250 and U251 of the helical region. (For experimental details see SI Table 2)

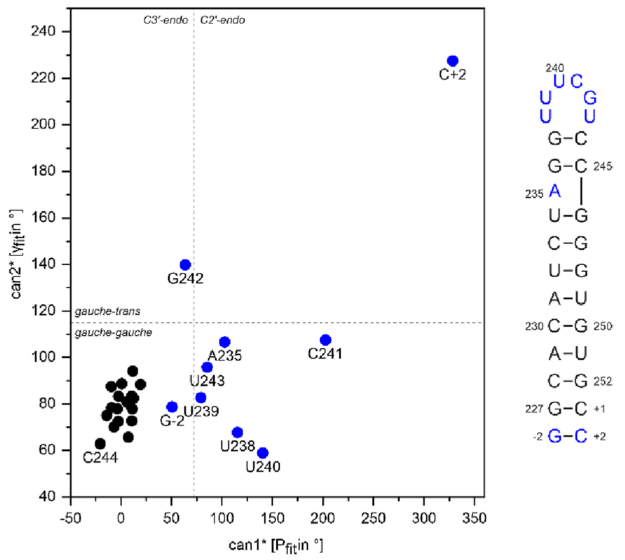


Fig. 3 Graph of canonical coordinates can1*[Pfit in °] and can2*[γfit in °] for SL5b_GC, calculated as in (Cherepanov et al. 2010). Data points are annotated by base numbering as used in the RNA secondary structure scheme on the right. Blue (both in the graph and the secondary structure) highlights residues with non-C3′-endo conformation or deviations in exocyclic torsion angle γ

ranging from C230 to G250. Small chemical shift differences are in line with the differences in primary chemical structure.

The assigned imino resonances of the SL5c sample provided a starting point for the aromatic proton resonances assignment of the ^{1}H,^{1}H-NOESY experiment (for experimental data see SI Table 2). The sequential walk was only interrupted by missing cross peaks between G256-H8 to A257-H8 in the loop region in both the 95% H_2O and the D_2O (Fig. 4A) samples in NMR buffer. Using ^{1}H,^{1}H-TOCSY and ^{1}H,^{13}C-HSQC (Fig. 4B and C) the aromatic resonances of protons H2, H6, H8 and H5 as well as their corresponding carbons except for U262-C5 were obtained. Using a ^{1}H,^{1}H-NOESY recorded for a sample diluted in D_2O (Fig. 4A), the H1′,C1′ ribose resonances were assigned by the analysis of intra-nucleotide and sequential NOEs. Similar to this, the H2′–C2′ assignments were obtained by identification of H1′–H2′ intra- nucleotide and sequential NOEs.

The 5′-UUUCGU-3′ hexaloop in SL5b

The entire SL5 motif (Fig. 1A) spans three stem-loop elements. Interestingly, SL5a and SL5b both possess the same 5′-UUUCGU-3′ hexaloop consisting of nucleotides 238–243 (SL5a: 200–205). While in SL5a, the loop closing stem is formed by at least three base pairs, the hexaloop of SL5b is closed by a stem consisting of only two GC base pairs, preceded by a bulge at residue A235. This bulged A235 shows downfield shifted signals for the aromatic H2 and H8 in the aromatic ^{1}H,^{13}C-HSQC, as typically observed for non-stacked purines. (Aeschbacher et al. 2013) Starting from H8 of residue G237, assignment of the ribose H1′ and aromatic H6/H8 was obtained by a sequential walk obtained in an amino nitrogen filtered NOESY (x-filter-NOESY). Loop residues U240 and C241 show characteristic C1′–H1′ shifts in the ^{1}H,^{13}C-HSQC, which reveal a fingerprint characteristic for the hexaloop. The overall resonance assignment of the hexaloop is in excellent agreement with the observations in SL5a. This loop arrangement is similar to a UUCG-tetraloop, for which detailed structural restraints are available. (Fürtig et al. 2004; Nozinovic et al. 2010).

Remarkably, available sequencing data for observed mutations in SARS-CoV2 (Hadfield et al. 2018; Cao et al. 2021), show significantly different vulnerability for mutations for the two hexaloop sequences. While the SL5a loop sequence remained mostly conserved until recently, in SL5b mutation C241U appeared in variants emerging since March 2020. A first study on mutational frequency indicates, among others, high C-to-U mutation rates in the SCoV2 genome. (Mourier et al. 2021) With the most recent mutation at SL5a C203U, both hexaloop sequences change to 5′-UUUUGU-3′. With the chemical shifts provided here, the delineation of structural differences for mutant versions of SCoV2 from changes in chemical shifts can be monitored by NMR spectroscopy.

The GAAA-tetraloop of SL5c

The GAAA-tetraloop of the SL5c stem consists of G256 to A259, closed by two GC and one AU base pairs. All three G N1–H1 resonances observed for the shorter construct were superimposable in the ^{1}H,^{15}N TROSY spectrum of SL5b + c. The chemical shifts observed in SL5b + c are in agreement with chemical shifts for a GAAA tetraloop (Jucker et al. 1996). Particularly, the ribose H1′ shift of ~ 3.5 ppm of the guanosine residue in the loop closing base pairing is characteristic. ^{31}P 1D data support the formation of a typical GAAA-tetraloop (SI Fig. 3, (Legault and Pardi 1994)). Legault and Pardi detected shows a stabilization of the G imino ^{1}H (corresponding to G256 in our construct) by interaction with a phosphate oxygen in the backbone for a GAAA tetraloop. While we could not assign crosspeaks between the loop guanosine (G256) and the loop adenosine (A259) that would have confirmed the reported loop geometry in SL5c, we found an additional imino crosspeak assigned to G256. In addition to the imino and amino proton assignment, which is consistent with the published chemical shifts for SL5b + c (Wacker et al. 2020), complete assignments of the aromatic C–H resonances as well as the ribose resonances C1′–H1′ and C2′–H2′ were obtained for element SL5c. While in SL5c imino signals of base pairs were observable at 298 K, a vanishing of these signals as well as appearance of additional

Table 1 List of NMR experiments for SL5b + c conducted at WIS and BMRZ at temperatures a: 275 K, b: 283 K and c: 298 K. Spectra were recorded in NMR buffer with A: 95% H_2O/5% D_2O or B: 100% D_2O. Experimental parameters and experiment-specific parameters are given

#	NMR experiment	Experimental parameters	Characteristic parameters
I	^{1}H,^{15}N-TROSY^{BMRZ}	**A** a,b: 700 MHz, ns: 16, sw(f2): 22.04 ppm, sw(f1): 25.62 ppm, aq(f2): 70.4 ms, aq(f1): 64.9 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 101 ppm, o3(^{15}N): 153 ppm, rel. delay: 0.3 s, time: 30 min **A** c: 700 MHz, ns: 16, sw(f2): 21.0 ppm, sw(f1): 25.3 ppm, aq(f2): 71.3 ms, aq(f1): 69.6 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 101 ppm, o3(^{15}N): 153 ppm, rel. delay: 0.3 s, time: 30 min	
II	^{1}H,^{1}H-NOESY^{WIS} jump-return water suppression	**A** a imino: 600 MHz, ns: 32, sw(f2): 22.03 ppm, sw(f1): 12.00 ppm, aq(f2): 77.4 ms, aq(f1): 53.3 ms, o1(^{1}H): 4.692 ppm, o3(^{15}N): 155 ppm, rel. delay: 2.0 s, time: 11 h 30 min **A** b imino: 600 MHz, ns: 40, sw(f2): 20.83 ppm, sw(f1): 11.49 ppm, aq(f2): 81.9 ms, aq(f1): 37.1 ms, o1(^{1}H): 4.7 ppm, o3(^{15}N): 153 ppm, rel. delay: 1.5 s, time: 10 h	**A**: NOE mixing time 150 ms, JR-delay 200 µs
III	^{1}H,^{15}N-cmpg-NOESY^{WIS} imino and amino cross−correlations	**A** b 600 MHz, ns: 96, sw(f2): 20.83 ppm, sw(f1): 102.72 ppm, aq(f2): 81.9 ms, aq(f1): 20.4 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 101 ppm, o3(^{15}N): 117 ppm, rel. delay: 2.0 s, time: 15 h 30 min	
IV	^{1}H,^{13}C-HSQC^{WIS} aromatic region (Bodenhausen and Ruben 1980)	**A** b: 600 MHz, ns: 16, sw(f2): 9.80 ppm, sw(f1): 26.00 ppm, aq(f2): 87.04 ms, aq(f1): 16.3 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 143 ppm, o3(^{15}N): 153 ppm, rel. delay: 1.0 s, time: 40 min **A** c: 600 MHz, ns: 16, sw(f2): 9.80 ppm, sw(f1): 26 ppm, aq(f2): 87.04 ms, aq(f1): 16.3098 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 143 ppm, o3(^{15}N): 153 ppm, rel. delay: 1.0 s, time: 40 min	INEPT transfer time 2.7 ms ($^{1}J_{CH}$ 185 Hz), off-resonant Q3 shaped pulse for C5 decoupling at 95 ppm with 25 ppm bandwidth
III	^{1}H,^{13}C-HSQC^{BMRZ} full	**A** b 800 MHz, ns: 8, sw(f2): 10.0 ppm, sw(f1): 110.4 ppm, aq(f2): 64.0 ms, aq(f1): 11.5 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 105 ppm, o3(^{15}N): 153 ppm, rel. delay: 1.0 s, time: 1 h 15 min	HSQC with gradient selection, INEPT transfer time 1.6 ms ($^{1}J_{CH}$ 160 Hz)
IV	^{1}H,^{13}C-ct-HSQC^{BMRZ} ribose region C1' to C5′	**B** c 900 MHz, ns: 4, sw(f2): 8.29 ppm, sw(f1): 38.0 ppm, aq(f2): 68.6 ms, aq(f1): 14.8 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 77 ppm, o3(^{15}N): 150 ppm, rel. delay: 1.0 s, time: 20 min	INEPT transfer time 1.6 ms ($^{1}J_{CH}$ 160 Hz), CT period 12.5 ms ($^{1}J_{CC}$ 77 Hz)
V	(H)C(CCN)H^{WIS} imino-to-aromatics (Piotto et al. 1992; Sklenář et al. 1996)	**A** b 600 MHz, ns: 256, sw(f2): 20.8 ppm, sw(f1): 9.94 ppm, aq(f2): 81.9 ms, aq(f1): 42.6 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 137 ppm, o3(^{15}N): 154 ppm, rel. delay: 1.5 s, time: 16 h **A** c 600 MHz, ns: 208, sw(f2): 20.8 ppm, sw(f1): 9.94 ppm, aq(f2): 90.2 ms, aq(f1): 42.6 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 137 ppm, o3(^{15}N): 154 ppm, rel. delay: 1.8 s, time: 15 h	CC-TOCSY mixing time 28 ms
VI	H(N)CO^{BRMZ,WIS} imino-to-carbon (Favier and Brutscher 2011; Solyom et al. 2013)	**A** a 700 MHz, ns: 128, sw(f2): 20.8 ppm, sw(f1): 31.0 ppm, aq(f2): 60.3 ms, aq(f1): 23.4 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 153 ppm, o3(^{15}N): 158 ppm, rel. delay: 0.3 s, time: 4 h 15 min **A** b 1000 MHz, ns: 64, sw(f2): 21.7 ppm, sw(f1): 22.0 ppm, aq(f2): 94.2 ms, aq(f1): 11.6 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 153 ppm, o3(^{15}N): 159 ppm, rel. delay: 0.5 s, time: 9 h	
VII	3D ^{1}H,^{13}C-NOESY-HSQC^{WIS} aromatics (Piotto et al. 1992; Sklenář et al. 1993)	**A** b,c 600 MHz, ns: 16, sw(f3,^{1}H): 9.8 ppm, sw(f2,^{13}C): 21.0 ppm, sw(f1,^{1}H): 14.4 ppm, aq(f3): 87.0 ms, aq(f2): 10.1 ms, aq(f1): 8.33 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 142 ppm, o3(15N): 154 ppm, rel. delay: 1.0 s, time: 2 d 7 h	NOE mixing time 200 ms

Table 1 (continued)

#	NMR experiment	Experimental parameters	Characteristic parameters
VIII	3D HCCH-TOCSY^{BMRZ} ribose C1′-to-C2′ (Kay et al. 1993; Richter et al. 2010)	**B** c 800 MHz, ns: 8, sw(f3, ^{1}H): 8.5 ppm, sw(f2, ^{13}C): 9.5 ppm, sw(f1, ^{13}C): 35.5 ppm, aq(f3): 74.7 ms, aq(f2): 21 ms, aq(f1): 11.2 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 76.5 ppm, o3(^{13}C): 76.5 ppm, rel. delay: 1.0 s, time: 1 d 10 h	CC-TOCSY mixing time 6 ms
IX	3D HCCH-TOCSY^{BMRZ} ribose C1′-to-C5′ (Kay et al. 1993; Richter et al. 2010)	**B** c 800 MHz, ns: 8, sw(f3, ^{1}H): 8.5 ppm, sw(f2, ^{13}C): 9.5 ppm, sw(f1, ^{13}C): 35.5 ppm, aq(f3): 74.7 ms, aq(f2): 21 ms, aq(f1): 11.2 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 76.5 ppm, o3(^{13}C): 76.5 ppm, rel. delay: 1.0 s, time: 1 day 10 h	CC-TOCSY mixing time 18 ms

ns number of scans, *sw* spectral width, *aq* acquisition time, *o1/2/3* carrier frequencies on channels *1/2/3*, *rel. delay* relaxation delay, *CT* constant time, *JR* jump-return

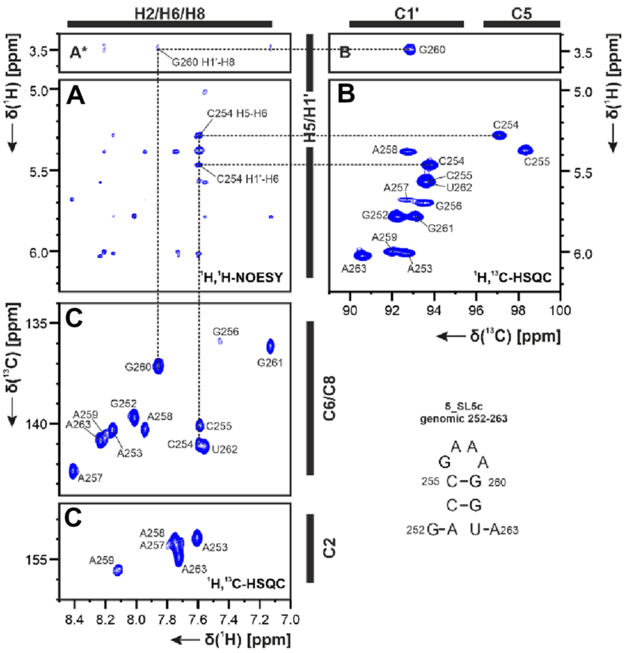


Fig. 4 **A** ^{1}H,^{1}H-NOESY, **B** ^{1}H,^{13}C-HSQC (C1' region) and **C** ^{1}H,^{13}C-HSQC (aromatic region) spectra for aromatic and ribose resonances of SL5c at 283 K in NMR buffer with 100% D_2O. *Lower contour level setting. Exemplary correlations are annotated by dashed lines and using the genomic numbering (for experimental details see SI Table 2)

signals and shifting of signals in the aromatic region of SL5c was noticed within the larger SL5b + c context. Thus, the SL5c stem is more stable and opens only at higher temperatures in the full-length construct (SI Fig. 4).

Summary

We herein present the ^{1}H, ^{13}C, ^{15}N chemical shifts of SL5b + c, using two fragments, SL5b_GC and SL5c, that subsequently allowed assignment of the SL5b + c element. The assignments of the sub-constructs were used as starting points for advancing the SL5b + c assignment. For the combined construct, an overall assignment was conducted at temperatures of 274 to 298 K. 95% of the aromatic H6–C6 and H8–C8 resonances were assigned for SL5b + c as well as the 8 adenosine H2–C2, and 17 uridine and cytidine H5–C5. Carbonyl and other quaternary carbon atoms of the nucleobases were partly assigned: in purines (C2: 75%, C4: 15% and C6: 50%) and pyrimidines (C2: 29% and C4: 24%). The assignment of the nitrogen atoms includes 70% N1 for purines and 65% N3 for pyrimidines, which represent mostly those involved in hydrogen bonding interactions. With the assignment transfer 90% of H1′–C1′ chemical shift assignment was obtained. Further ribose resonances for H2′ to H5″ and C2′ to C5′ are partially assigned in ranges of 10 to 30%.

In summary, an assignment of 65% of the ^{1}H, 53% of the ^{13}C and 63% of the ^{15}N atoms in the nucleobases of SL5b + c has been achieved.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12104-021-10053-4.

Acknowledgements Work at BMRZ is supported by the state of Hesse. Work in COVID19-NMR was supported by the Goethe Corona Funds, by the IWB-EFRE-programme 20007375 of state of Hesse, and the DFG in CRC902: "Molecular Principles of RNA-based regulation." and infrastructure funds (Project Numbers: 277478796, 277479031, 392682309, 452632086, 70653611). Work at Weizmann was supported by the EU Horizon 2020 program (FET-OPEN Grant 828946, PATHOS), Israel Science Foundation Grants 965/18 and 3572/20, and the Perlman Family Foundation. J.K. and J.T.G. are supported by the Israel Academy of Sciences and Humanities & Council for Higher Education Excellence Fellowship Program for International Postdoctoral Researchers. H.S. and B.F. are supported by the DFG in graduate school CLIC (GRK 1986). L.F. holds the Bertha and Isadore Gudelsky Professorial Chair and heads the Clore Institute for High-Field Magnetic Resonance Imaging and Spectroscopy, whose support is acknowledged.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability We updated the BMRB deposition for SL5b + c with code 50339. Data of SL5b_GC are found in BMRB code 51138 and SL5c BRMB code 51137.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Aeschbacher T, Schmidt E, Blatter M et al (2013) Automated and assisted RNA resonance assignment using NMR chemical shift statistics. Nucleic Acids Res 41:e172–e172. https://doi.org/10.1093/NAR/GKT665

Bodenhausen G, Ruben DJ (1980) Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem Phys Lett 69:185–189. https://doi.org/10.1016/0009-2614(80)80041-8

Cao C, Cai Z, Xiao X et al (2021) (2021) The architecture of the SARS-CoV-2 RNA genome inside virion. Nat Commun 121(12):1–14. https://doi.org/10.1038/s41467-021-22785-x

Chen SC, Olsthoorn RCL (2010) Group-specific structural features of the 5′-proximal sequences of coronavirus genomic RNAs. Virology 401:29–41. https://doi.org/10.1016/J.VIROL.2010.02.007

Cherepanov AV, Glaubitz C, Schwalbe H (2010) High-resolution studies of uniformly 13C,15N-labeled RNA by solid-state NMR spectroscopy. Angew Chem Int Ed Engl 49:4747–4750. https://doi.org/10.1002/anie.200906885

Favier A, Brutscher B (2011) Recovering lost magnetization: polarization enhancement in biomolecular NMR. J Biomol NMR 49:9–15. https://doi.org/10.1007/s10858-010-9461-5

Fürtig B, Richter C, Bermel W (2004) Schwalbe H (2004) New NMR experiments for RNA nucleobase resonance assignment and chemical shift analysis of an RNA UUCG tetraloop. J Biomol NMR 281(28):69–79. https://doi.org/10.1023/B:JNMR.0000012863.63522.1F

Guan B-J, Wu H-Y, Brian DA (2011) An optimal cis -replication stem-loop IV in the 5′ untranslated region of the mouse coronavirus genome extends 16 nucleotides into open reading frame 1. J Virol 85:5593–5605. https://doi.org/10.1128/JVI.00263-11

Guillerez J, Lopez PJ, Proux F et al (2005) A mutation in T7 RNA polymerase that facilitates promoter clearance. Proc Natl Acad Sci 102:5958–5963. https://doi.org/10.1073/PNAS.0407141102

Hadfield J, Megill C, Bell SM et al (2018) Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123. https://doi.org/10.1093/BIOINFORMATICS/BTY407

Huston NC, Wan H, Strine MS et al (2021) Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. Mol Cell 81:584-598.e5. https://doi.org/10.1016/J.MOLCEL.2020.12.041

Ikura M, Bax A (2002) Isotope-filtered 2D NMR of a protein-peptide complex: study of a skeletal muscle myosin light chain kinase fragment bound to calmodulin. J Am Chem Soc 114:2433–2440. https://doi.org/10.1021/JA00033A019

Jucker FM, Heus HA, Yip PF et al (1996) A network of heterogeneous hydrogen bonds in GNRA tetraloops. J Mol Biol 264:968–980. https://doi.org/10.1006/JMBI.1996.0690

Kay LE, Xu GY, Singer AU et al (1993) A gradient-enhanced HCCH-TOCSY experiment for recording side-chain 1H and 13C correlations in H_2O samples of proteins. J Magn Reson Ser B 101:333–337. https://doi.org/10.1006/JMRB.1993.1053

Lan TCT, Allan MF, Malsick LE et al (2021) Insights into the secondary structural ensembles of the full SARS-CoV-2 RNA genome in infected cells. bioRxiv. https://doi.org/10.1101/2020.06.29.178343

Lee W, Tonelli M, Markley JL (2015) NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics 31:1325–1327. https://doi.org/10.1093/BIOINFORMATICS/BTU830

Legault P, Pardi A (1994) 31P chemical shift as a probe of structural motifs in RNA. J Magn Reson Ser B 103:82–86. https://doi.org/10.1006/JMRB.1994.1012

Madhugiri R, Fricke M, Marz M, Ziebuhr J (2014) RNA structure analysis of alphacoronavirus terminal genome regions. Virus Res 194:76–89. https://doi.org/10.1016/J.VIRUSRES.2014.10.001

Madhugiri R, Fricke M, Marz M, Ziebuhr J (2016) Coronavirus cis-acting RNA elements. Adv Virus Res 96:127–163. https://doi.org/10.1016/BS.AIVIR.2016.08.007

Manfredonia I, Incarnato D (2021) Structure and regulation of coronavirus genomes: state-of-the-art and novel insights from SARS-CoV-2 studies. Biochem Soc Trans 49:341–352. https://doi.org/10.1042/BST20200670

Marino JP, Schwalbe H, Anklin C et al (1995) Sequential correlation of anomeric ribose protons and intervening phosphorus in RNA oligonucleotides by a 1H,13C,31P triple resonance experiment:

HCP-CCH-TOCSY. J Biomol NMR 5:87–92. https://doi.org/10.1007/BF00227473

Masters PS (2019) Coronavirus genomic RNA packaging. Virology 537:198–207. https://doi.org/10.1016/J.VIROL.2019.08.031

Mourier T, Sadykov M, Carr MJ et al (2021) Host-directed editing of the SARS-CoV-2 genome. Biochem Biophys Res Commun 538:35–39. https://doi.org/10.1016/J.BBRC.2020.10.092

Nozinovic S, Fürtig B, Jonker HRA et al (2010) High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. Nucleic Acids Res 38:683–694. https://doi.org/10.1093/NAR/GKP956

Piotto M, Saudek V (1992) Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. J Biomol NMR 26(2):661–665. https://doi.org/10.1007/BF02192855

Rangan R, Watkins AM, Chacon J et al (2021) De novo 3D models of SARS-CoV-2 RNA elements from consensus experimental secondary structures. Nucleic Acids Res 49:3092–3108. https://doi.org/10.1093/NAR/GKAB119

Richter C, Kovacs H, Buck J et al (2010) 13C-direct detected NMR experiments for the sequential J-based resonance assignment of RNA oligonucleotides. J Biomol NMR 474(47):259–269. https://doi.org/10.1007/S10858-010-9429-5

Schnieders R, Knezic B, Zetzsche H et al (2020) NMR spectroscopy of large functional rnas: from sample preparation to low-gamma detection. Curr Protoc Nucleic Acid Chem 82:e116. https://doi.org/10.1002/cpnc.116

Schnieders R, Peter SA, Banijamali E et al (2021) ^{1}H, ^{13}C and ^{15}N chemical shift assignment of the stem-loop 5a from the 5′-UTR of SARS-CoV-2. Biomol NMR Assign 151(15):203–211. https://doi.org/10.1007/S12104-021-10007-W

Schürer H, Lang K, Schuster J, Mörl M (2002) A universal method to produce in vitro transcripts with homogeneous 3′ ends. Nucleic Acids Res 30:e56–e56. https://doi.org/10.1093/NAR/GNF055

Schwalbe H, Marino JP, Glaser SJ, Griesinger C (1995) Measurement of, -coupling constants associated with vi, Vi, and V3 in uniformly 13C-labeled RNA by HCC-TOCSY-CCH-E.COSY. J Am Chem Soc 117:7251–7252

Simon B, Zanier K (2001) A TROSY relayed HCCH-COSY experiment for correlating adenine H2/H8 resonances in uniformly 13C-labeled RNA molecules. J Biomol NMR 202(20):173–176. https://doi.org/10.1023/A:1011214914452

Sklenář V, Piotto M, Leppik R, Saudek V (1993) Gradient-tailored water suppression for 1H–15N HSQC experiments optimized to retain full sensitivity. J Magn Reson Ser A 102:241–245. https://doi.org/10.1006/JMRA.1993.1098

Sklenář V, Peterson RD, Rejante MR (1993) Two-and three-dimensional HCN experiments for correlating base and sugar resonances in 15N, 13C-labeled RNA oligonucleotides. J Biomol NMR 36(3):721–727. https://doi.org/10.1007/BF00198375

Sklenář V, Dieckmann T, Butcher SE (1996) Through-bond correlation of imino and aromatic resonances in 13C-,15N-labeled RNA via heteronuclear TOCSY. J Biomol NMR 71(7):83–87. https://doi.org/10.1007/BF00190460

Solyom Z, Schwarten M, Geist L et al (2013) BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. J Biomol NMR 554(55):311–321. https://doi.org/10.1007/S10858-013-9715-0

Sreeramulu S, Richter C, Berg H et al (2021) Exploring the druggability of conserved RNA regulatory elements in the SARS-CoV-2 genome. Angew Chemie Int Ed. https://doi.org/10.1002/ANIE.202103693

Vuister GW, Bax A (1992) Resolution enhancement and spectral editing of uniformly 13C-enriched proteins by homonuclear broadband 13C decoupling. J Magn Reson 98:428–435. https://doi.org/10.1016/0022-2364(92)90144-V

Wacker A, Weigand JE, Akabayov SR et al (2020) Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy. Nucleic Acids Res 48:12415–12435. https://doi.org/10.1093/nar/gkaa1013

Wishart DS, Bigam CG, Yao J et al (1995) 1H, 13C and 15N chemical shift referencing in biomolecular NMR. J Biomol NMR 62(6):135–140. https://doi.org/10.1007/BF00211777

Yang D, Leibowitz JL (2015) The structure and functions of coronavirus genomic 3′ and 5′ ends. Virus Res 206:120–133

Zhao J, Qiu J, Aryal S et al (2020) The RNA architecture of the SARS-CoV-2 3′-untranslated region. Viruses 12:1473. https://doi.org/10.3390/V12121473

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5.7 **Research article:** ^{1}H, ^{13}C and ^{15}N assignment of stem-loop SL1 from the 5'-UTR of SARS-CoV-2

Christian Richter*, Katharina F. Hohmann*, Sabrina Toews, Daniel Mathieu, Nadide Altincekic, Jasleen Kaur Bains, Oliver Binas, Betül Ceylan, Elke Duchardt-Ferner, Jan Ferner, Boris Fürtig, J. Tassilo Grün, Martin Hengesbach, Daniel Hymon, Hendrik R. A. Jonker, Bozana Knezic, Sophie M. Korn, Tom Landgraf, Frank Löhr, Stephen A. Peter, Dennis J. Pyper, Nusrat S. Qureshi, Andreas Schlundt, Robbin Schnieders, Elke Stirnal, Alexey Sudakov, Jennifer Vögele, Julia E. Weigand, Julia Wirmer-Bartoschek, Kerstin Witt, Jens Wöhnert, Harald Schwalbe and Anna Wacker; *Biomolecular NMR Assignments* **15**, 467-474 (2021)

* These authors contributed equally to this work.

In this article, the stem-loop 1 (SL1) located in the 5'-UTR of the SARS-CoV-2 genome was characterized by NMR spectroscopy. The ^{1}H, ^{13}C, and ^{15}N chemical shift assignments of SL1 were presented as a basis for further in-depth structural and ligand screening studies.

The project was designed and coordinated by the Covid-19 NMR consortium. C. Richter and K. F. Hohmann assigned the NMR spectra and wrote the manuscript. The author of this thesis was part of the Covid-19 NMR consortium and the RNA production team that prepared and purified various isotope-labeled RNA samples for NMR spectroscopy.

https://doi.org/10.1007/s12104-021-10047-2

Biomolecular NMR Assignments (2021) 15:467–474
https://doi.org/10.1007/s12104-021-10047-2

ARTICLE

^{1}H, ^{13}C and ^{15}N assignment of stem-loop SL1 from the 5'-UTR of SARS-CoV-2

Christian Richter[1,2] · Katharina F. Hohmann[1,2] · Sabrina Toews[1,2] · Daniel Mathieu[3] · Nadide Altincekic[1,2] · Jasleen Kaur Bains[1,2] · Oliver Binas[1,2,4] · Betül Ceylan[1,2] · Elke Duchardt-Ferner[2,5] · Jan Ferner[1,2] · Boris Fürtig[1,2] · J. Tassilo Grün[1,2,6] · Martin Hengesbach[1,2] · Daniel Hymon[1,2] · Hendrik R. A. Jonker[1,2] · Bozana Knezic[1,2] · Sophie M. Korn[2,5] · Tom Landgraf[1,2] · Frank Löhr[2,7] · Stephen A. Peter[8] · Dennis J. Pyper[1,2] · Nusrat S. Qureshi[1,2,9] · Andreas Schlundt[2,5] · Robbin Schnieders[1,2,10] · Elke Stirnal[1,2] · Alexey Sudakov[1,2] · Jennifer Vögele[2,5] · Julia E. Weigand[8] · Julia Wirmer-Bartoschek[1,2] · Kerstin Witt[1,2] · Jens Wöhnert[2,5] · Harald Schwalbe[1,2] · Anna Wacker[1,2]

Received: 29 June 2021 / Accepted: 19 August 2021 / Published online: 28 August 2021

Abstract
The stem-loop (SL1) is the 5'-terminal structural element within the single-stranded SARS-CoV-2 RNA genome. It is formed by nucleotides 7–33 and consists of two short helical segments interrupted by an asymmetric internal loop. This architecture is conserved among Betacoronaviruses. SL1 is present in genomic SARS-CoV-2 RNA as well as in all subgenomic mRNA species produced by the virus during replication, thus representing a ubiquitous *cis*-regulatory RNA with potential functions at all stages of the viral life cycle. We present here the ^{1}H, ^{13}C and ^{15}N chemical shift assignment of the 29 nucleotides-RNA construct 5_SL1, which denotes the native 27mer SL1 stabilized by an additional terminal G-C base-pair.

Keywords SARS-CoV-2 · 5'-UTR · SL1 · Solution NMR spectroscopy · COVID19-NMR

Biological context

The 5'-untranslated regions (5'-UTR) of Betacoronavirus RNA genomes contain several highly conserved, structured RNA elements that play essential roles in viral RNA synthesis. SL1, the first of these RNA stem-loops, has been structurally characterized by NMR spectroscopy in Mouse hepatitis virus (MHV), Bovine coronavirus (BCoV), and the human coronavirus HCoV-OC43 (Liu et al. 2007). Despite local differences in RNA sequences, the ~ 37 nucleotides (nt)

Christian Richter and Katharina F. Hohmann have contributed equally to this work.

✉ Harald Schwalbe
schwalbe@nmr.uni-frankfurt.de

✉ Anna Wacker
wacker@nmr.uni-frankfurt.de

[1] Institute for Organic Chemistry and Chemical Biology, Goethe-University Frankfurt, Max-von-Laue- Straße 7, 60438 Frankfurt, Germany

[2] Center for Biomolecular Magnetic Resonance (BMRZ), Goethe-University Frankfurt, Max-von-Laue- Straße 7, 60438 Frankfurt, Germany

[3] Bruker BioSpin, Silberstreifen 4, 76287 Rheinstetten, Germany

[4] Present Address: BioNTech SE, An der Goldgrube 12, 55131 Mainz, Germany

[5] Institute for Molecular Biosciences, Goethe-University Frankfurt, Max-von-Laue-Straße 9, 60438 Frankfurt, Germany

[6] Present Address: Faculty of Chemistry, Weizmann Institute of Science, 7610001 Rehovot, Israel

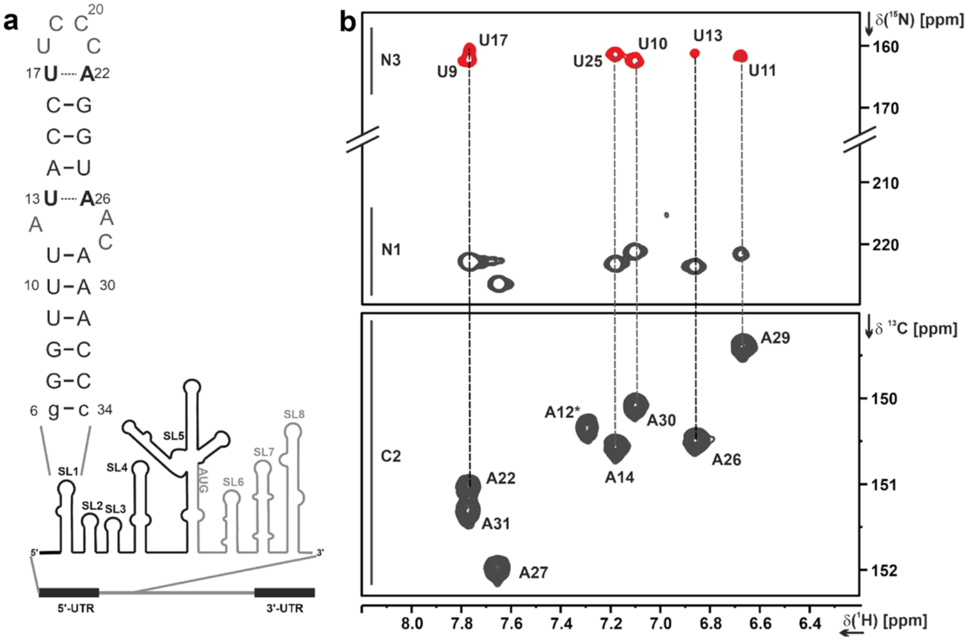
[7] Institute for Biophysical Chemistry, Goethe-University Frankfurt, Max-von-Laue-Straße 9, 60438 Frankfurt, Germany

[8] Department of Biology, Technical University of Darmstadt, Schnittspahnstraße 10, 64287 Darmstadt, Germany

[9] Present Address: EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany

[10] Present Address: Deutero GmbH, Am Ring 29, 56288 Kastellaun, Germany

Fig. 1 **a** Secondary structure of 5_SL1 and its genomic position within the 5'-UTR of the SARS-CoV-2 genome. **b** Detection of the W–C base-pairs U13-A26 and U17-A22 in the lrHNN-COSY experiment (Table 1, XIII.). Adenosine C2H2 resonances (lower spectrum, ^{1}H,^{13}C-HSQC) were used to assign the ^{2}J-N1H2 diagonal peaks and the corresponding uridine N3 cross peaks. Note that the A12 N1H2 resonance is broadened beyond detection. The U13-A22 and U17-A22 correlations are shown in black, the other base-pairs in grey in panel **a**



stem-loop adopts a very similar secondary structure in all three viruses, consisting of two helical parts interrupted by a stretch of nucleotides with mismatched bases and capped by a less conserved apical loop. Extensive mutational studies of MHV SL1 accompanied by NMR showed that virus viability depends on the sequence of the lower part of SL1 and on the stability of the upper part of SL1 (Li et al. 2008). For SL1 from SARS-CoV, it was shown that it can replace MHV SL1 and restore virus replication (Kang et al. 2006), suggesting a functionally equivalent role for SL1 in Betacoronaviruses in general. Subsequently, for the human pathogenic viruses MERS-CoV, SARS-CoV, and SARS-CoV-2, an additional function for SL1 was described. Here, SL1 is involved in viral escape from non-structural protein 1-mediated translational shutdown (Tanaka et al. 2012; Terada et al. 2017; Tidu et al. 2020). At present, the predicted secondary structure of stem-loop SL1 in SARS-CoV-2 (Fig. 1) has been experimentally verified (Miao et al. 2020; Wacker et al. 2020; Iserman et al. 2020; Manfredonia et al. 2020). SL1 is formed by nucleotides 7–33 of the 5'-UTR. The 5-base-pair (bp) lower helix is separated from the 3-bp upper helix by an asymmetric 5-nt internal loop flanked on both sides by A–U Watson–Crick (W–C) base-pairs. The UUCCCA apical loop has been mapped as an interaction site with the host protein LARP1 (Schmidt et al. 2020).

Methods and NMR experiments

RNAs were synthesized by *in vitro* run-off transcription from linearized DNA plasmids as previously described (Wacker et al. 2020; Schnieders et al. 2021; Vögele et al. 2021). For DNA template production, the sequence of SL1 (RNA sequence 5'gGGUUUAUACCUUCCCAGGUAACAAACCc-3') together with the T7 promoter was generated by hybridization of complementary oligonucleotides and introduced into the EcoRI and NcoI sites of an HDV ribozyme (Schürer et al. 2002) encoding plasmid, based on the pSP64 vector (Promega). RNAs were transcribed as a fusion construct with the 3'-HDV ribozyme to obtain homogeneous 3RNAs were transcribed as a fusion construct with the 3'-HDV ribozyme to obtain homogeneous 3'-ends. Transformation and amplification of the recombinant vector pHDV-5_SL1 was done in the *Escherichia coli* strain DH5α. Plasmid-DNA was purified using a large scale DNA isolation kit (Gigaprep; Qiagen) according to the manufacturer's instructions and linearized with HindIII prior to *in vitro* transcription with T7 RNA polymerase [P266L mutant, prepared as described in (Guilleres et al. 2005)]. RNA amounts sufficient for NMR experiments were produced in 15 ml preparative transcription reactions [20 mM dithiothreitol, 2 mM spermidine, 200 ng/µl plasmid template, 200 mM Tris/glutamate (pH 8.1), 30 mM Mg(OAc)$_2$, 12 mM rNTPs, 32 µg/ml (^{15}N,^{13}C-labelled RNAs)/150 µg/ml (uniformly ^{15}N labelled RNA) T7 RNA Polymerase]. After 1 h incubation time, yeast inorganic phosphatase [9.6 µg/mL (^{15}N,^{13}C-labelled RNAs)/4.8 µg/mL (uniformly ^{15}N labelled RNA) final concentration] was added. Transcription reactions (6 h at 37 °C and 70 rpm) were terminated by addition of EDTA (80 mM final concentration) and NaOAc (0.3 M final concentration). After transcription, RNAs were precipitated by adding 1 volume equivalent of ice-cold 2-propanol and incubation for 1 h at − 20 °C. For purification, RNA fragments were separated on 12 % denaturing polyacrylamide (PAA) gels and visualized by UV shadowing at 254 nm. SL1

RNAs were excised from the gel and incubated at − 80 °C for 30 min, followed by 15 min at 65 °C in 0.3 M NaOAc. Elution was achieved overnight by passive diffusion into 30 mL 0.3 M NaOAc solution. RNAs were precipitated by addition of 4 volume equivalents of ethanol at − 20 °C overnight. If the absorption ratio 220/260 nm of the RNA after dissolving in water was higher than 1.5, RNA was desalted via PD10 columns (GE Healthcare) for the following HPLC. Residual PAA was removed by reversed-phase HPLC using a Kromasil RP-18 column and a gradient of 0–40 % 0.1 M acetonitrile/triethylammonium acetate. After freeze-drying of RNA-containing fractions and cation exchange by $LiClO_4$ precipitation [2 % (w/v) in acetone], the RNA was folded in water by heating to 80 °C followed by rapid cooling on ice. Buffer exchange to NMR buffer (25 mM potassium phosphate buffer, pH 6.2, 50 mM potassium chloride) was performed using Vivaspin centrifugal concentrators (2 kDa molecular weight cut-off, Sarstedt). Purity of SL1 was verified by denaturing PAA gel electrophoresis and homogenous folding was monitored by native PAA gel electrophoresis, loading the same RNA concentration as used in NMR experiments (Fig. S1).

Using this protocol, four NMR samples of 5_SL1 were prepared and used for the assignment presented herein: A 0.64 mM uniformly ^{15}N labelled RNA sample and a 1.2 mM uniformly ^{15}N,^{13}C-labelled RNA sample, each in NMR buffer with 5 % (v/v) D_2O for a 5 mm Shigemi tube and 7 % (v/v) D_2O for a 1.7 mm NMR tube, a 1.33 mM uniformly ^{15}N,^{13}C-labelled RNA in 99.95 % (v/v) D_2O and an 0.87 mM selectively ^{15}N,^{13}C(A/C)-labelled RNA in NMR buffer (5 % (v/v) D_2O).

Assignment strategy and extent of assignment

Based on our previously reported assignment of the base-paired imino groups, the amino groups of base-paired cytidines and the adenosine H2 protons for 5_SL1 (Wacker et al. 2020), we have already confirmed the overall secondary structure of 5_SL1 consisting of two helical regions. For the stably base-paired adenosine and cytidine residues, we have previously also reported the assignments of the hydrogen bond-acceptor nitrogens in the HNN-COSY experiment.

Starting from these available assignments and following the classical NOE-based strategy, we first assigned all anomeric H1′ protons and all aromatic H6 (pyrimidine)/H8 (purine) protons via one single "sequential walk" in a 2D NOESY spectrum acquired in D_2O (Table 1, I.). For the nucleotides U9/U10, U18/C19, and C20/C21, the anomeric-aromatic walk was ambiguous in the H1′–H6/8-region due to severe signal overlap. However, these connectivities could be unambiguously established via the intra-nucleotide and sequential $H2'_i$–$H8/H6_{i,\ (i-1)}$ NOEs. Within the H1′–H6/H8 region of the NOESY, also the pyrimidine (intraresidual) H5–H6 and adenosine $H1'_i$–$H2_{(i+1)\ intra-strand,\ (i+1)\ cross-strand}$ NOE signals are typically observed. The 2D NOESY experiment, in combination with a 2D ^{1}H,^{1}H-TOCSY experiment showing only the pyrimidine H5–H6 cross peaks, thus allowed the unambiguous assignment of all pyrimidine H5 and adenosine H2 protons. All protonated nucleobase carbons as well as the C1' carbons were assigned in ^{1}H,^{13}C-HSQCs optimized for the respective CH-transfer (Table 1, II. and III.). Correlations from purine C8H8 and adenosine C2H2 resonances were used as starting points to assign all adenosine and guanosine N7/N9 resonances and adenosine N1/N3 resonances in the 2D ^{1}H,^{15}N-^{2J}HSQC as described in (Wacker et al. 2020), except for the A12 N1 resonance, which was not observable, most likely due to exchange broadening. For the adenosines, all base ^{13}C nuclei were assigned by correlating the C2H2 and C8H8 resonances with the quaternary base carbons C4, C5, and C6 in the 3D TROSY-(H)CCH-COSY experiment (Table 1, IV.). The same experiment also yielded assignments for guanosine C4 and C5 resonances. Uridine C2/C4 and guanosine C2/C6 resonances were assigned by correlating the respective imino protons to the carbonyl resonances in a 2D H(N)CO experiment (Table 1, V.). ^{15}N resonances of all exocyclic adenosine amino groups were identified in a ^{13}C-detected 2D ^{13}C,^{15}N-HSQC (Table 1, VI.). Ribose spin systems were connected to their respective nucleobases by simultaneously correlating C1' and C6 (for pyrimidine nucleobases) or C8 (for purine nucleobases) to the glycosidic (N1/N9) nitrogen atom in ^{1}H-detected 3D HCN and ^{13}C-detected 3D (H)CNC experiments (Table 1, VII. and VIII.), verifying the sequential NOE-based assignment of the H1′ protons. 3D (H)CCH-TOCSY experiments were used to identify the carbon resonances of the ribose spin systems. Discrimination of C2′ and C3′ was achieved by varying the CC-TOCSY mixing time to either correlate C1′and C2′ during a short TOCSY mixing time (6 ms) or to correlate C1′ to all ribose carbons via a long TOCSY mixing time of 18 ms (Table 1, IX). Due to severe resonance overlap of the respective C1′H1′ resonances, the carbon spin systems for G6, G7, and G24 were not unambiguously resolved. In summary, about 90 % of the ribose H2′–H5′/H5″ resonances were assigned via a 3D forward-directed HCCH-TOCSY experiment (Table 1, X.), a 3D ^{13}C-NOESY-HSQC (Table 1, XI.) and 2D ^{13}C-filtered/edited NOESY experiments (Table 1, X. and XI.) on a selectively ^{13}C,^{15}N (A/C)-labelled sample.

Internal loop

According to our previously reported secondary structure determination of 5_SL1, the internal loop consists of nucleotides A12-U13 and A26–A27-C28 (Wacker et al. 2020). A26 and A27 could both be potential interaction partners for U13, as observed for the homologous RNA element in MHV for A35 and A36 (Liu et al. 2007). However, formation of a W-C-type U13-A26 interaction was unambiguously observed in the lrHNN-COSY experiment (Table 1, XIII. and Fig. 1)), which in turn precluded a significantly populated U13-A27 interaction and eventually confined the internal loop to nucleotides A12, A27 and A28. The $^{2}J_{NN}$ coupling for U13N3-A26N1 was 4.5 Hz as derived from the intensity ratio of cross peak to diagonal peak according to $I_{cross}/I_{dia} = -\tan^{2}(\pi J_{NN}\tau)$ (Bax et al. 1994). For comparison, $^{2}J_{NN}$ couplings for U11N3-A29N1, U10N3-A30N1, and U25N3-A14N1 were around 6.4 Hz, 6.6 Hz, and 6.7 Hz, respectively. The intraresidual N1 resonance of A12 was the only missing signal in the H2-N1/N3 correlation experiment, hinting at severe exchange-induced line-broadening. Note that this experiment clearly rules out disappearance of signals due to solvent exchange.

Empirical determination of ribose conformation by means of the canonical coordinates yielded no significant deviation from A-form helical structure for A12 and C28 (Fig. 2), whereas A27 was found to adopt a C2′-endo conformation. Qualitative evaluation of glycosidic torsion angles via the intensity of the intra-base H1′–H6/H8 NOESY cross peak did not reveal a tendency for *syn* conformation for any of the internal loop nucleotides. Furthermore, global chemical shift analysis using CS-Annotate (Zhang et al. 2021) supported a largely stacked arrangement of all nucleobases of the internal loop, except for C28 (SI Fig. S2).

Pyrimidine loop

The apical loop of 5_SL1 is formed by nucleotides U17-A22. For U17-A22, formation of a labile W–C base-pair was observed in the lrHNN-COSY (Fig. 1). Overlap of the A22 and A27 N1H2 resonances did not allow us to derive the $^{2}J_{NN}$ coupling constant for A22N1-U17N3 in the same way as for the other A–U base-pairs as described above, but the U17N3 cross peak showed a reduced intensity compared to the canonical A–U base-pairs (Fig. 1). Ribose carbon chemical shifts of both nucleotides yielded canonical coordinates consistent with A-form conformation. Taken together, these results indicated that U17-A22 rather extends the upper helix by one base-pair, while the apical loop is a tetraloop formed by nucleotides U18 to C21. Linewidths in the TOCSY experiment were narrow for U18, C19, C20 and medium for C21, indicating conformational flexibility of this region (Fig. 3). The downfield chemical shifts of the U18 and C19 C6H6 groups were a further indication that these nucleotides are solvent-exposed and likely not participate in extensive stacking interactions. The Y-rich loop of 5_SL1 is currently discussed as a binding site for the Y-motif binding protein LARP1 (Schmidt et al. 2020). This protein-RNA interaction would severely impact the conformational flexibility of the involved nucleotides. Thus, the resonances of pyrimidines U18, C19, C20 and C21 may serve as valuable reporters for future structural investigations of RNA-protein interactions involving the apical loop of 5_SL1.

Conclusions

It is common in NMR spectroscopy of RNA to consider W–C base-pairs as "stable" if the H-bonding imino proton is significantly protected from solvent-exchange and gives rise to an observable imino proton signal. Relying on the presence of imino proton signals only, the upper helix of SARS-CoV-2 5_SL1 consists only of three stable base-pairs, as these signals for U13 and U17 are missing even at 275 K. Available secondary structure predictions (Tavares et al. 2020; Rangan et al. 2020; Andrews et al. 2021), however, base pairs U13-A26 and U17-A22 are consistently present. We show here that these base pairs are at least significantly populated via the lrHNN-COSY experiment. This demonstrates the unique ability of solution NMR spectroscopy to capture subtle differences in secondary structure stability under given conditions. In SARS-CoV-2, the lower helix appears to be the most stable part of 5_SL1, which is in contradiction to the putative function in genome cyclization and the observed lability of the lower SL1 helices in MHV, HCoV-OC43, and BCoV (Li et al. 2008). Interestingly, long-range RNA-RNA interactions have been recently mapped for SARS-CoV-2 involving the 5'-UTR downstream elements SL2 and SL3 as interaction sites with the 3'-UTR (Ziv et al. 2020). Thus, the function of genome cyclization might have been handed over to other conserved RNA structures in SARS-CoV-2 while acquiring distinct functions for SL1 not yet described for its counterparts in MHV or BCoV. These functions may include protecting viral mRNA from translation shutdown (Tidu et al. 2020). Our extensive assignment of ^{1}H, ^{13}C and ^{15}N chemical shifts for 5_SL1 provides experimental data as the basis for in-depth structural characterization of this stem-loop RNA and refines the currently available structure models in terms of structural dynamics, which is essential e.g., for the identification of potential drug binding sites.

Table 1 List of NMR experiments, "(Bruker)" indicates the NMR experiments that were carried out at Bruker BioSpin, Rheinstetten

NMR experiment	Experimental parameters
I. 2D ^{1}H,^{1}H NOESY **A**: (Bruker) aromatics, in 99.95 % D_2O **B**: Iminos and aromatics with excitation sculpting (Hwang and Shaka 1995; Sklenar 1995)	A: 800 MHz, 298 K, ns: 16, sw(f2): 12.0 ppm, sw(f1): 6.5 ppm, aq(f2): 319 ms, aq(f1): 162 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 118 ppm, o3(^{15}N): 190 ppm, rel. delay: 1.5 s, NOE mixing time: 150 and 300 ms, time: 14 h B: 900 MHz, 283 K, ns: 64, sw(f2): 22.2 ppm, sw(f1): 11.8 ppm, aq(f2): 102 ms, aq(f1): 45 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 110 ppm, o3(^{15}N): 153 ppm, rel. delay: 1.45 s, NOE mixing time: 80 , 160 and 240 ms, time: 29 h
II. 2D ^{1}H,^{13}C-HSQC **A**: Aromatics **B**: C1′-H1′ (Bodenhausen and Ruben 1980), optimized in-house	A: 700 MHz, 298 K, ns: 4, rel. delay: 1.0 s, sw(f2): 9.2 ppm, sw(f1): 10 ppm, aq(f2): 67 ms, aq(f1): 85 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 142.5 ppm, o3(^{15}N): 153 ppm, INEPT transfer time: 2.7 ms, off-resonant Q3 shaped pulse for C5 decoupling at 95 ppm with 25 ppm bandwidth, time: 35 min B: 600 MHz, 298 K, ns: 4, rel. delay: 1.0 s, sw(f2): 8.7 ppm, sw(f1): 22.7 ppm, aq(f2): 84 ms, aq(f1): 32 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 90.5 ppm, o3(^{15}N): 154 ppm, INEPT transfer time: 2.9 ms, off-resonant Q3 shaped pulse for C2′ decoupling at 72 ppm with 12 ppm bandwidth, time: 20 min
III. 2D ^{1}H,^{13}C-ct-HSQC All CH, optimized for ribose resonances (Vuister and Bax 1992)	700 MHz, 298 K, ns: 32, sw(f2): 8.3 ppm, sw(f1): 105 ppm, aq(f2): 102 ms, aq(f1): 16 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 105 ppm, rel. delay: 1.0 s, INEPT transfer time 2.9 ms, constant-time period: 25 ms, time: 5 h
IV. 3D TROSY-(H)CCH-COSY Adenine base sin systems (Simon et al. 2001)	950 MHz, 298 K, ns: 8, sw(f3, ^{1}H): 9.0 ppm, sw(f2, ^{13}C): 26.2 ppm, sw(f1, ^{13}C): 58.1 ppm, aq(f3): 119 ms, aq(f2): 5.1 ms, aq(f1): 4.6 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 142.5 ppm, o3(^{15}N): 150 ppm, rel. delay: 1.0 s, time: 21 h
V. 2D BEST-TROSY-H(N)CO (Favier and Brutscher 2011; Solyom et al. 2013)	600 MHz, 283 K, ns: 128, sw(f2): 21.0 ppm, sw(f1): 31 ppm, aq(f2): 63 ms, aq(f1): 13,6 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 157 ppm, o3(^{15}N): 153 ppm, rel. delay: 0.3 s, HN-INEPT transfer time: 5.2 ms, NC-INEPT transfer time 18 ms, time: 1.5 h
VI. 2D ^{13}C-detected ^{1} C,^{15}N-HSQC C2/4/6 to Amino-N2/4/6′ (Bermel et al. 2006; Fiala and Sklenár 2007)	800 MHz, 298 K, ns: 32, rel. delay: 2.5 s, sw(f2, ^{13}C): 50 ppm, sw(f1, ^{15}N): 43 ppm, aq(f2): 51 ms, aq(f1): 16 ms, o1(^{13}C): 160 ppm, o2(^{15}N): 86.5 ppm, INEPT CN transfer time: 18 ms, time: 2.5 h
VII. 3D HCN (Bruker) H6/8/H1′-to-N9/N1, in 99.95 % D_2O (Fiala et al. 1998)	800 MHz, 298 K, ns: 8, sw(f3, ^{1}H): 8.9 ppm, sw(f2, ^{13}C): 28 ppm, sw(f1, ^{15}N): 31 ppm, aq(f3): 143 ms, aq(f2): 8.5 ms, aq(f1): 32 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 113.5 ppm, o3(^{15}N): 157 ppm, rel. delay: 1.0 s, INEPT HC transfer time: 2.8 ms, INEPT CN transfer time: 30 ms, time: 1 d 15 h
VIII. 3D ^{13}C-detected (H)CNC C1′-to-C6/8 Modified from Fiala et al. (1998)	800 MHz, 298 K,, ns: 24, sw(f3, ^{13}C): 24 ppm, sw(f2, ^{15}N): 34 ppm, sw(f1, ^{13}C): 12 ppm, aq(f3): 67 ms, aq(f2): 23 ms, aq(f1): 25 ms, o1(^{13}C): 90 ppm, o2(^{1}H): 7.6 ppm, o3(^{15}N): 157 ppm, rel. delay: 0.5 s, C6/8-N1/9 transfer time 30 ms, C–H transfer time 2.9 ms (1′) and 2.6 ms (6/8), time: 2 d 10 h
IX. 3D (H)CCH TOCSY **A**: C1′ to C2′; **B**: C1′ to C5′ (Kay et al. 1993; Richter et al. 2010)	700 MHz, 298 K, ns: 16, sw(f3,^{1}H): 10.4 ppm, sw(f2,^{13}C): 10.0 ppm, sw(f1,^{13}C): 35.4 ppm, aq(f3): 82 ms, aq(f2): 26 ms, aq(f1): 12 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 39 ppm, o3(^{31}P): − 1 ppm, rel. delay: 1.0 s, CC-TOCSY mixing time (dipsi3 spin-lock): A: 6 ms, B: 18 ms, time: 2 d 2 h
X. 3D FW-directed H(C)CH-TOCSY (Schwalbe et al. 1995; Glaser et al. 1996)	700 MHz, 298 K, ns: 8, sw(f3,^{1}H): 8.3 ppm, sw(f2,^{13}C): 38.5 ppm, sw(f1,^{1}H): 4.1 ppm, aq(f3): 87 ms, aq(f2): 8 ms, aq(f1): 27 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 77 ppm, o3(^{15}N): 155 ppm, rel. delay: 1.0 s, constant-time period: 8.3ms; CC-TOCSY mixing time (dipsi3 spin-lock): 9.2 ms, time: 1 d 22 h
XI. ^{13}C-NOESY-HSQC **A**: 3D (Bruker) in 99.95 % D_2O; **B**: 2D, sel. ^{13} C,^{15} N(A,C)-labelled RNA (Sklenář et al. 1993; Piotto et al. 1992)	800 MHz, 298 K, A (constant time in t2): ns: 8, sw(f3,^{1}H): 12 ppm, sw(f2,^{13}C): 105 ppm, sw(f1,^{1}H): 5.9 ppm, aq(f3): 106 ms, aq(f2): 23 ms, aq(f1): 17 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 108.5 ppm, o3(^{15}N): 105 ppm, rel. delay: 1.0 s, HC-INEPT transfer time: 3 ms, constant-time period: 8.8 ms, NOE mixing time: 150 ms, time: 1 d 19 h B: ns: 64, sw(f2): 8.8 ppm, sw(f1,^{1}H): 6.2 ppm, aq(f2): 73 ms, aq(f1): 51 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 144 ppm, o3(^{15}N): 154 ppm, rel. delay: 0.9 s, HC-INEPT transfer time: 2.8 ms NOE mixing time: 200 ms, time: 11 h
XII. 2D ^{13}C,^{15}N(F2)-filtered NOESY All-to-G/U protons (Ogura et al. 1996; Zwahlen et al. 1997; Breeze 2000; Iwahara et al. 2001)	900 MHz, 298 K, ns: 48, sw(f2): 12 ppm, sw(f1,^{1}H): 9 ppm, aq(f2): 94 ms, aq(f1): 51 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 120 ppm, o3(^{15}N): 117 ppm, rel. delay: 1.5 s, NOE mixing time: 150 ms, time: 14 h
XIII. 2D ^{1}H,^{15}N-BEST-TROSY-lrHNN-COSY (Sklenár et al. 1994; Hennig and Williamson 2000; Farjon et al. 2009; Dingley and Grzesiek 1998; Dingley et al. 2008)	600 MHz, 298 K, ns: 512, sw(f2): 9.8 ppm, sw(f1): 88.9 ppm, aq(f2): 87 ms, aq(f1): 14.8 ms, o1(^{1}H): 7 ppm, o2(^{13}C): 150 ppm, o3(^{15}N): 192 ppm, rel. delay: 0.3 s, HN-INEPT transfer time: 19 ms, NN-transfer time 22.5 ms, time: 11 h
XIV. ^{1}H,^{1}H-TOCSY (Shaka et al. 1988; Hwang and Shaka 1995)	700 MHz, 283 K, ns: 16, sw(f2): 8.8 ppm, sw(f1): 6.2 ppm, aq(f2): 100 ms, aq(f1): 51 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 101 ppm, o3(^{15}N): 85 ppm, rel. delay: 1.0 s, TOCSY mixing time (dipsi3 spin-lock): 30 ms, time: 3 h

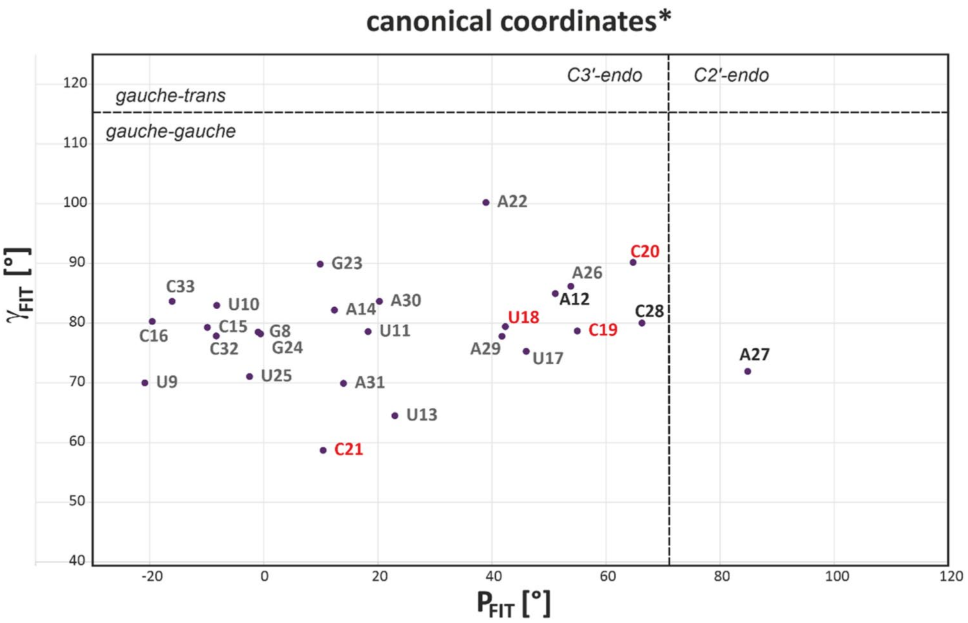


Fig. 2 Plot of γ_{FIT} against P_{FIT} as calculated from ribose ^{13}C chemical shifts according to (Cherepanov et al. 2010). Residues from the apical loop are marked in red, bulge residues in black. C34 is omitted due to its low-field C2′ chemical shift typical for the 3'-terminal nucleotide, resulting in exceptionally high values of the canonical coordinates

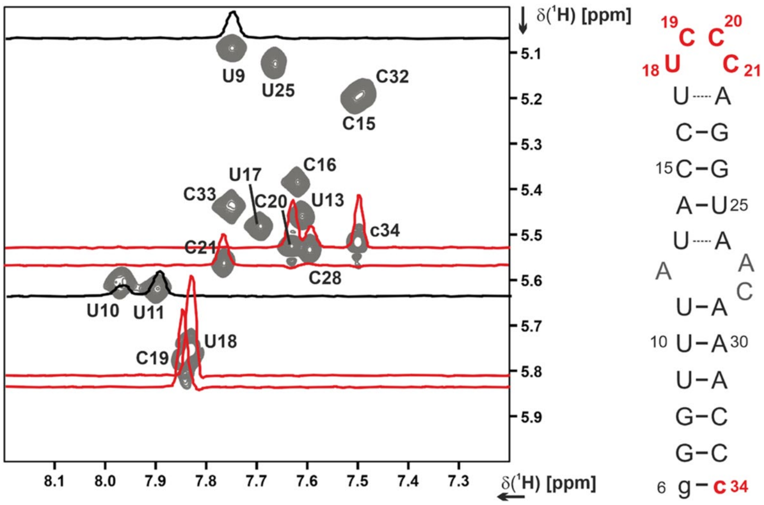


Fig. 3 Expanded region of the 2D ^{1}H,^{1}H TOCSY experiment (Table 1, XIV.) correlating pyrimidine H5–H6 proton chemical shifts via their ^{3}J coupling. Linewidths are approximately inversely proportional to the base order parameter, resulting in sharp signals for flexible residues that exhibit a lower than the global τ_c. 1D traces for selected residues are shown in the 2D. The flexible loop residues U18, C19, and C20 and the non-native 3'-terminal c34 are highlighted in red; helical residues U9 and U11 are shown in black

Data deposition

The BMRB deposition with the accession code 50349 was updated with the assignments reported herein.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12104-021-10047-2.

Acknowledgements We thank Bruker BioSpin for measurement time to record several NMR spectra at Rheinstetten, Germany (indicated in Table 1), and Dr. Wolfgang Bermel, Dr. Klaus-Peter Neidig, Dr. Kristof Grohe at Bruker BioSpin for excellent support. We also wish to thank Marie Hutchison for helpful discussions and the COVID19-NMR consortium for providing an inspiring research atmosphere. We are grateful to Dr. Martin Hähnke (Signals) for constant technical support.

Funding Open Access funding enabled and organized by Projekt DEAL. Work at the Center for Biomolecular Magnetic Resonance (BMRZ) at the Goethe-University Frankfurt is supported by the state of Hesse. Work in Covid19-nmr was supported by the Goethe Corona Funds, by the IWB-EFRE-programme 20007375 of the state of Hesse, and the DFG in CRC902: "Molecular Principles of RNA-based regulation." and infrastructure funds (Project Numbers: 277478796, 277479031, 392682309, 452632086, 70653611). K. F. H., B. F., B. K., H. S., and J. T. G. are supported by the DFG in graduate school CLIC (GRK 1986). D. J. P. and H. S. are supported by the DFG in SPP2002. A.W., B. F., O. B., S. K., A. S., M. H., Ju. Wei, Je. Woe. and H. S. are supported by the DFG in CRC902. T. L. and H. S. are supported by the DFG in SPP1879. J. K. B. and H. S. are supported by the DFG in TRR 267, H. R. A. Jonker and H.S. are supported by the DFG in FOR2509. A. S. is funded by the Deutsche Forschungsgemeinschaft through Grant number SCHL2062/2-1 and by the Johanna Quandt Young Academy at Goethe (Grant number 2019/AS01).

Declarations

Conflict of interest The authors declare the following competing financial interest(s): Daniel Mathieu is an employee of Bruker BioSpin.

References

Andrews RJ, O'Leary CA, Tompkins VS et al (2021) A map of the SARS-CoV-2 RNA structurome. NAR Genom Bioinform. https://doi.org/10.1093/nargab/lqab043

Bax A, Vuister GW, Grzesiek S et al (1994) Measurement of homo- and heteronuclear J couplings from quantitative J correlation. Nuclear magnetic resonance, Part C. Academic Press, Cambridge, pp 79–105

Bermel W, Bertini I, Felli IC et al (2006) 13 C-detected protonless NMR spectroscopy of proteins in solution. Prog Nucl Magn Reson Spectrosc 48:25–45

Bodenhausen G, Ruben DJ (1980) Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem Phys Lett 69:185–189. https://doi.org/10.1016/0009-2614(80)80041-8

Breeze AL (2000) Isotope-filtered NMR methods for the study of biomolecular structure and interactions. Prog Nucl Magn Reson Spectrosc 4:323–372. https://doi.org/10.1016/S0079-6565(00)00020-0

Cherepanov AV, Glaubitz C, Schwalbe H (2010) High-resolution studies of uniformly 13 C,15 N-labeled RNA by solid-state NMR spectroscopy. Angew Chem Int Ed Engl 49:4747–4750. https://doi.org/10.1002/anie.200906885

Dingley AJ, Grzesiek S (1998) Direct observation of hydrogen bonds in nucleic acid base pairs by internucleotide 2J(NN) couplings. J Am Chem Soc 120:8293–8297. https://doi.org/10.1021/ja981513x

Dingley AJ, Nisius L, Cordier F, Grzesiek S (2008) Direct detection of N – H···N hydrogen bonds in biomolecules by NMR spectroscopy. Nat Protoc 3:242–248. https://doi.org/10.1038/nprot.2007.497

Farjon J, Boisbouvier JR, Schanda P et al (2009) Longitudinal-relaxation-enhanced NMR experiments for the study of nucleic acids in solution. J Am Chem Soc 131:8571–8577. https://doi.org/10.1021/ja901633y

Favier A, Brutscher B (2011) Recovering lost magnetization: Polarization enhancement in biomolecular NMR. J Biomol NMR 49:9–15. https://doi.org/10.1007/s10858-010-9461-5

Fiala R, Jiang F, Sklenář V (1998) Sensitivity optimized HCN and HCNCH experiments for 13 C/15 N labeled oligonucleotides. J Biomol NMR 12:373–383. https://doi.org/10.1023/A:1008369515755

Fiala R, Sklenár V (2007) 13 C-detected NMR experiments for measuring chemical shifts and coupling constants in nucleic acid bases. J Biomol NMR 39:153–163. https://doi.org/10.1007/s10858-007-9184-4

Glaser SJ, Schwalbe H, Marino JP, Griesinger C (1996) Directed TOCSY, a method for selection of directed correlations by optimal combinations of isotropic and longitudinal mixing. J Magn Reson B 112:160–180. https://doi.org/10.1006/jmrb.1996.0126

Guilleres J, Lopez PJ, Proux F et al (2005) A mutation in T7 RNA polymerase that facilitates promoter clearance. Proc Natl Acad Sci U S A 102:5958–5963. https://doi.org/10.1073/pnas.0407141102

Hennig M, Williamson JR (2000) Detection of N–H ··· N hydrogen bonding in RNA via scalar couplings in the absence of observable imino proton resonances. Nucleic Acids Res 28:1585–1593. https://doi.org/10.1093/nar/28.7.1585

Hwang TL, Shaka AJ (1995) Water suppression that works. Excitation sculpting using arbitrary wave-forms and pulsed-field gradients. J Magn Reson Ser A 112:275–279. https://doi.org/10.1006/jmra.1995.1047

Iserman C, Roden CA, Boerneke MA et al (2020) Genomic RNA elements drive phase separation of the SARS-CoV-2 nucleocapsid. Mol Cell 80:1078-1091.e6. https://doi.org/10.1016/j.molcel.2020.11.041

Iwahara J, Wojciak JM, Clubb RT (2001) Improved NMR spectra of a protein-DNA complex through rational mutagenesis and the application of a sensitivity optimized isotope-filtered NOESY experiment. J Biomol NMR 19:231–241

Kang H, Feng M, Schroeder ME et al (2006) Putative cis-acting stem-loops in the 5' untranslated region of the severe acute respiratory syndrome coronavirus can substitute for their mouse hepatitis virus counterparts. J Virol 80:10600–10614. https://doi.org/10.1128/JVI.00455-06

Kay LE, Xu GY, Singer AU et al (1993) A gradient-enhanced HCCH-TOCSY experiment for recording side-chain 1H and 13 C correlations in H2O samples of proteins. J Magn Reson Ser B 101:333–337. https://doi.org/10.1006/jmrb.1993.1053

Li L, Kang H, Liu P et al (2008) Structural lability in stem-loop 1 drives a 5′ UTR-3′ UTR interaction in coronavirus replication. J Mol Biol 377:790–803. https://doi.org/10.1016/j.jmb.2008.01.068

Liu P, Li L, Millership JJ et al (2007) A U-turn motif-containing stem-loop in the coronavirus 5′ untranslated region plays a functional role in replication. RNA 13:763–780. https://doi.org/10.1261/rna.261807

Manfredonia I, Nithin C, Ponce-Salvatierra A et al (2020) Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. Nucleic Acids Res 48:12436–12452. https://doi.org/10.1093/nar/gkaa1053

Miao Z, Tidu A, Eriani G (2020) Martin F (2020) Secondary structure of the SARS-CoV-2 5′-UTR. RNA Biol 10(1080/15476286):1814556

Ogura K, Terasawa H, Inagaki F (1996) An improved double-tuned and isotope-filtered pulse scheme based on a pulsed field gradient and a wide-band inversion shaped pulse. J Biomol NMR 8:492–498. https://doi.org/10.1007/BF00228150

Piotto M, Saudek V, Sklenár V (1992) Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. J Biomol NMR 2:661–665

Rangan R, Zheludev IN, Das R (2020) RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. RNA 26:937–959. https://doi.org/10.1261/RNA.076141.120

Richter C, Kovacs H, Buck J et al (2010) 13 C-direct detected NMR experiments for the sequential J-based resonance assignment of RNA oligonucleotides. J Biomol NMR 47:259–269

Schmidt N, Lareau CA, Keshishian H et al (2020) The SARS-CoV-2 RNA–protein interactome in infected human cells. Nat Microbiol. https://doi.org/10.1038/s41564-020-00846-z

Schnieders R, Peter SA, Banijamali E et al (2021) ^{1}H, ^{13} C and ^{15} N chemical shift assignment of the stem-loop 5a from the 5′-UTR of SARS-CoV-2. Biomol NMR Assign. https://doi.org/10.1007/s12104-021-10007-w

Schürer H, Lang K, Schuster J, Mörl M (2002) A universal method to produce in vitro transcripts with homogeneous 3' ends. Nucleic Acids Res 30:56. https://doi.org/10.1093/nar/gnf055

Schwalbe H, Marino JP, Glaser SJ et al (1995) Measurement of H,H-coupling constants associated with ν1, ν2, and ν3 in uniformly 13 C-labeled RNA by HCC-TOCSY-CCH-E.COSY. J Am Chem Soc 117:7251–7252. https://doi.org/10.1021/ja00132a028

Shaka A, Lee C, Pines A (1988) Iterative schemes for bilinear operators; application to spin decoupling. J Magn Reson 77:274–293. https://doi.org/10.1016/0022-2364(88)90178-3

Simon B, Zanier K, Sattler M (2001) A TROSY relayed HCCH-COSY experiment for correlating adenine H2/H8 resonances in uniformly 13 C-labeled RNA molecules. J Biomol NMR 20:173–176. https://doi.org/10.1023/a:1011214914452

Sklenar V (1995) Suppression of radiation damping in multidimensional nmr experiments using magnetic field gradients. J Magn Reson Ser A 114:132–135. https://doi.org/10.1006/jmra.1995.1119

Sklenár V, Peterson RD, Rejante MR, Feigon J (1994) Correlation of nucleotide base and sugar protons in a 15 N-labeled HIV-1 RNA oligonucleotide by 1H-15 N HSQC experiments. J Biomol NMR 4:117–122. https://doi.org/10.1007/BF00178339

Sklenář V, Piotto M, Leppik R, Saudek V (1993) Gradient-tailored water suppression for 1H–15N HSQC experiments optimized to retain full sensitivity. J Magn Reson Ser A 102(2):241–245

Solyom Z, Schwarten M, Geist L et al (2013) BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. J Biomol NMR 55:311–321. https://doi.org/10.1007/s10858-013-9715-0

Tanaka T, Kamitani W, DeDiego ML et al (2012) Severe acute respiratory syndrome coronavirus nsp1 facilitates efficient propagation in cells through a specific translational shutoff of host mRNA. J Virol 86:11128–11137. https://doi.org/10.1128/JVI.01700-12

Tavares RDCA, Mahadeshwar G, Pyle AM (2020) The global and local distribution of RNA structure throughout the SARS-CoV-2 genome. J Virol. https://doi.org/10.1101/2020.07.06.190660

Terada Y, Kawachi K, Matsuura Y, Kamitani W (2017) MERS coronavirus nsp1 participates in an efficient propagation through a specific interaction with viral RNA. Virology 511:95–105. https://doi.org/10.1016/j.virol.2017.08.026

Tidu A, Janvier A, Schaeffer L et al (2020) The viral protein NSP1 acts as a ribosome gatekeeper for shutting down host translation and fostering SARS-CoV-2 translation. RNA. https://doi.org/10.1261/rna.078121.120

Vögele J, Ferner JP, Altincekic N et al (2021) 1H, 13 C, 15 N and 31P chemical shift assignment for stem-loop 4 from the 5′-UTR of SARS-CoV-2. Biomol NMR Assign 1:3. https://doi.org/10.1007/s12104-021-10026-7

Vuister GW, Bax A (1992) Resolution enhancement and spectral editing of uniformly 13 C-enriched proteins by homonuclear broadband 13 C decoupling. J Magn Reson 98:428–435. https://doi.org/10.1016/0022-2364(92)90144-V

Wacker A, Weigand JE, Akabayov SR et al (2020) Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy. Nucleic Acids Res. https://doi.org/10.1093/nar/gkaa1013

Zhang K, Abdallah K, Ajmera P et al (2021) CS-annotate: a tool for using NMR chemical shifts to annotate RNA structure. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.1c00006

Ziv O, Price J, Shalamova L et al (2020) The short- and long-range RNA-RNA interactome of SARS-CoV-2. Mol Cell. https://doi.org/10.1016/j.molcel.2020.11.004

Zwahlen C, Legault P, Vincent SJF et al (1997) Methods for measurement of intermolecular NOEs by multinuclear NMR spectroscopy: application to a bacteriophage λ N-peptide/boxB RNA complex. J Am Chem Soc 119:6711–6721. https://doi.org/10.1021/ja970224q

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5.8 Research article: ^{1}H, ^{13}C, ^{15}N and ^{31}P chemical shift assignment for stem-loop 4 from the 5′-UTR of SARS-CoV-2

Jennifer Vögele*, Jan-Peter Ferner*, Nadide Altincekic, Jasleen Kaur Bains, Betül Ceylan, Boris Fürtig, J. Tassilo Grün, Martin Hengesbach, Katharina F. Hohmann, Daniel Hymon, Bozana Knezic, Frank Löhr, Stephen A. Peter, Dennis Pyper, Nusrat S. Qureshi, Christian Richter, Andreas Schlundt, Harald Schwalbe, Elke Stirnal, Alexey Sudakov, Anna Wacker, Julia E. Weigand, Julia Wirmer-Bartoschek, Jens Wöhnert and Elke Duchardt-Ferner; *Biomolecular NMR Assignments* **15**, 335-340 (2021)

* These authors contributed equally to this work.

In this work, the stem-loop 4 (SL4) located in the 5'-UTR of the SARS-CoV-2 genome was characterized by NMR spectroscopy. ^{1}H, ^{13}C, ^{15}N, and ^{31}P chemical shift assignments of SL1 were presented as a basis for further in-depth structural and ligand screening studies.

The project was designed and coordinated by the Covid-19 NMR consortium. J. Vögele and J.-P. Ferner assigned the NMR spectra and wrote the manuscript. The author of this thesis was part of the Covid-19 NMR consortium and the RNA production team that prepared and purified various isotope-labeled RNA samples for NMR spectroscopy.

https://doi.org/10.1007/s12104-021-10026-7

Biomolecular NMR Assignments
https://doi.org/10.1007/s12104-021-10026-7

ARTICLE

^{1}H, ^{13}C, ^{15}N and ^{31}P chemical shift assignment for stem-loop 4 from the 5′-UTR of SARS-CoV-2

Jennifer Vögele[1,2] · Jan-Peter Ferner[2,3] · Nadide Altincekic[2,3] · Jasleen Kaur Bains[2,3] · Betül Ceylan[2,3] · Boris Fürtig[2,3] · J. Tassilo Grün[2,3] · Martin Hengesbach[2,3] · Katharina F. Hohmann[2,3] · Daniel Hymon[2,3] · Bozana Knezic[2,3] · Frank Löhr[2,4] · Stephen A. Peter[5] · Dennis Pyper[2,3] · Nusrat S. Qureshi[6] · Christian Richter[2,3] · Andreas Schlundt[1,2] · Harald Schwalbe[2,3] · Elke Stirnal[2,3] · Alexey Sudakov[2,3] · Anna Wacker[2,3] · Julia E. Weigand[5] · Julia Wirmer-Bartoschek[2,3] · Jens Wöhnert[1,2] · Elke Duchardt-Ferner[1,2]

Received: 3 March 2021 / Accepted: 16 April 2021

Abstract

The SARS-CoV-2 virus is the cause of the respiratory disease COVID-19. As of today, therapeutic interventions in severe COVID-19 cases are still not available as no effective therapeutics have been developed so far. Despite the ongoing development of a number of effective vaccines, therapeutics to fight the disease once it has been contracted will still be required. Promising targets for the development of antiviral agents against SARS-CoV-2 can be found in the viral RNA genome. The 5′- and 3′-genomic ends of the 30 kb SCoV-2 genome are highly conserved among Betacoronaviruses and contain structured RNA elements involved in the translation and replication of the viral genome. The 40 nucleotides (nt) long highly conserved stem-loop 4 (5_SL4) is located within the 5′-untranslated region (5′-UTR) important for viral replication. 5_SL4 features an extended stem structure disrupted by several pyrimidine mismatches and is capped by a pentaloop. Here, we report extensive ^{1}H, ^{13}C, ^{15}N and ^{31}P resonance assignments of 5_SL4 as the basis for in-depth structural and ligand screening studies by solution NMR spectroscopy.

Keywords SARS-CoV-2 · RNA genome · 5′-UTR · 5_SL4 · Solution NMR spectroscopy · COVID19-NMR

Biological context

SARS-CoV-2 is a human Betacoronavirus which causes the severe acute respiratory syndrome COVID-19. The virus contains a large single-stranded (+) RNA genome with a length of approximately 30,000 nt. Besides the coding regions for the viral proteins, the genome also includes extended, highly structured and conserved 5′- and 3′-untranslated regions (UTRs) with important functional roles in genome replication, transcription of subgenomic (sg) mRNAs and the balanced translation of viral proteins. So far, efforts aiming at the development of new antiviral agents against SARS-CoV-2 have been largely restricted to studies of the viral proteins, leaving the potentially vast reservoir of putative drug-targets to be found in the structured, conserved and functional genomic RNA elements essentially untapped. Triggered by the COVID-19 pandemic, the Covid19-NMR initiative (https://covid19-NMR.de) has united structural biologists and RNA biologists around the globe in a concerted initiative to make these viral RNA

✉ Elke Duchardt-Ferner
duchardt@bio.uni-frankfurt.de

[1] Institute for Molecular Biosciences, Goethe-University Frankfurt, Max-von-Laue-Str. 9, 60438 Frankfurt/M., Germany

[2] Center for Biomolecular Magnetic Resonance (BMRZ), Goethe-University Frankfurt, Max-von-Laue-Str. 7, 60438 Frankfurt/M., Germany

[3] Institute for Organic Chemistry and Chemical Biology, Goethe-University Frankfurt, Max-von-Laue-Str. 7, 60438 Frankfurt/M., Germany

[4] Institute for Biophysical Chemistry, Goethe-University Frankfurt, Max-von-Laue-Str. 9, 60438 Frankfurt/M., Germany

[5] Department of Biology, Technical University of Darmstadt, Schnittspahnstr. 10, 64287 Darmstadt, Germany

[6] Present Address: EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany

Published online: 29 April 2021

elements amenable as therapeutic targets as well as to pilot structure-guided drug screening efforts against these RNA targets. At the heart of this effort is the conviction that drug development can profit from and be efficiently guided by high resolution structural data. As a starting point of the initiative, the individual structured elements of the SARS-CoV-2 genome were therefore subjected to high resolution structure determination by NMR spectroscopy in a 'divide-and-conquer' approach.

The 5′- region (Fig. 1a) of the SARS-CoV-2 genome consists of eight stem-loop (SL) structures. Stem-loops 5_SL1 to 5_SL5 are located in the 5′-untranslated region (5′-UTR). While the sequences of the individual structural elements vary between different coronaviruses, their ubiquitous presence and highly conserved secondary structures suggest that these elements are critically important for viral viability and pathogenesis (Madhugiri et al. 2016). Stem-loop 4 of the 5′-UTR (5_SL4, nt 86–125), a 40 nt predicted hairpin capped by a pentaloop, is structurally conserved among the members of the Betacoronavirus family. Interestingly, 5_SL4 carries an upstream open reading frame (uORF) with its AUG start codon as integral part of the stem. This uORF is conserved within the Betacoronaviruses. Its function, however, is still under debate. On the one hand, genetic pressure to preserve the uORF has been observed. On the other hand, mutations manipulating the uORF yet retaining the 5_SL4 structure were still viable (Wu et al. 2014). We have recently established the secondary structure of 5_SL4 based on initial ^{1}H and ^{15}N NMR resonance assignments (Wacker et al. 2020). As a further step to guide structure-based studies of 5_SL4 amenable we provide here the almost complete ^{1}H, ^{13}C,^{15}N and ^{31}P NMR chemical shift assignment.

Methods and experiments

In order to adapt 5_SL4 for enzymatic synthesis, the 40 nt sequence (residues 86–125 of the SARS_COV2 genome) was extended by two guanine residues at the 5′- and two cytidine residues at the 3′-end yielding the 44 nt sequence 5′-GG**GUGUGGCUGUCACUCGGCUGCAUGCUU AGUGCACUCACGC**CC-3′ (5_SL4) with the wt-sequence shown in bold letters. In addition, a shorter construct comprising only the apical residues 96–116 again flanked by

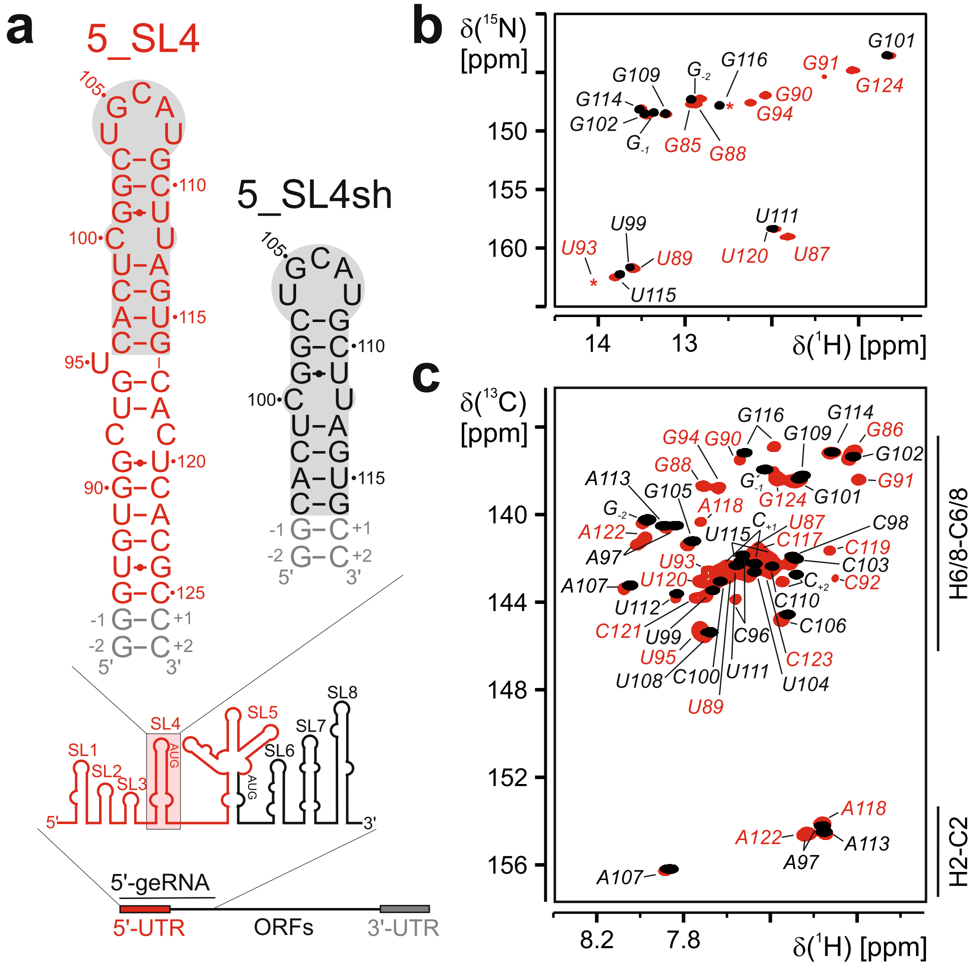


Fig. 1 **a** Sequence and secondary structure of the 44 nt 5_SL4 (top, left) and the smaller construct 5_SL4sh (top, right) and genomic context of 5_SL4. Scheme of the SARS-CoV2 genome (bottom) and overview of cis-acting RNA elements of the 5′ genomic end (middle). **b** Overlay of the imino ^{1}H, ^{15}N-BEST-TROSY spectra of 5_SL4 (red) and 5_SL4sh (black). Assignments are given. The asterisks mark signals that are visible at a lower contour threshold. **c** Overlay of the aromatic region of ^{1}H,^{13}C-HSQC spectra of 5_SL4 (red) and 5_SL4sh (black). Assignments are given. Those that only belong to the full-length construct are highlighted in red

two G-C base pairs was synthesized (5_SL4sh, 25 nt, sequence 5′-GG**CACUCGGCUGCAUGCUUAGUG**CC-3′). Of 5_SL4, a uniformly ^{15}N- and selectively A/C- and G/U-^{13}C/^{15}N-labeled samples and of 5_SL4sh a uniformly ^{13}C/^{15}N-labeled sample were prepared as described in detail previously (Wacker et al. 2020). The final RNA concentrations in all NMR samples varied between 0.35 and 0.79 mM in 25 mM potassium phosphate buffer, pH 6.2, with 50 mM potassium chloride and either 5% or 100% (v/v) D_2O.

NMR measurements were carried out at the Center for Biomolecular Magnetic Resonance (BMRZ) on 600, 800, 900 and 950 MHz Bruker Avance NMR-spectrometers equipped with 5-mm cryogenic triple resonance TCI-N probe heads, a 700 MHz spectrometer equipped with a quadruple resonance QCI-P probe and an 800 MHz spectrometer equipped with a ^{13}C-optimized TXO cryogenic probe (800 MHz^{TXO}). ^{1}H chemical shifts were referenced directly to an external DSS standard, and ^{13}C, ^{15}N, ^{31}P and chemical shifts were indirectly referenced from the ^{1}H chemical shift as described earlier (Maurer and Kalbitzer 1996; Wishart et al. 1995). All NMR experiments conducted for the resonance assignment of 5_SL4 and 5_SL4sh are summarized in Supplementary Table 1. If not indicated otherwise, experiments were performed at 25 °C. NMR data were processed using TOPSPIN 4.0.6 software (Bruker, BioSpin, Germany) and analyzed using CARA (Keller 2004).

Extent of assignments and data deposition

5_SL4 is a 44 nt long predicted stem-loop capped by an apical loop of five nucleotides (Fig. 1a). Given the rather large size of this RNA together with its expected rod-like extended shape and high content of canonical base-pairs, unfavorable relaxation behavior is expected to combine with limited resonance dispersion to interfere with a complete resonance assignment. We therefore also investigated a smaller construct containing only the apical stem loop and the predicted C100-U112 mismatch (5_SL4sh, Fig. 1a) to allow for an unambiguous sequential assignment of the apical loop as well as the acquisition of the chemical shifts of the mismatch residues as a prerequisite for the subsequent structure determination. For transcriptional reasons we added two terminal G-C base pairs at the end of the stem in both constructs.

The comparison of the imino ^{1}H,^{15}N-TROSY spectra of the full-length construct and 5_SL4sh indicates the complete preservation of the RNA structure in the truncated variant (Fig. 1b). This is also confirmed by comparing the aromatic region of ^{1}H,^{13}C-HSQC spectra recorded for both constructs (Fig. 1c). Only residues at the bottom of the 5_SL4sh stem display minor shift changes compared to the full length construct as expected.

We have previously reported an initial resonance assignment of 5_SL4 comprising the imino and amino groups and extending to the aromatic and H1′ protons (Wacker et al. 2020). On that basis, we followed essentially the classical assignment pathway using the NMR experiments listed in Supplementary Table S1. For 5_SL4sh, the assignment of the imino- and amino-group resonances could be readily transferred from the full length 5_SL4 assignment in ^{15} N HSQC spectra optimized for the imino- and the amino group region, respectively. Based on the previous assignment of the aromatic proton spins, ^{1}H,^{13}C-sfHMQC and ^{1}H,^{13}C-HSQC spectra for the aromatic region served to assign 100% of the H2-C2 and H6/8-C6/8 resonances. 3D ^{13}C-NOESY-HSQC spectra confirmed this assignment. All of the adenine N1 and N3 and the purine N7 and N9 resonances were assigned in the lr-^{1}H,^{15}N-HSQC. Nine out of 14 guanine N3 resonances were observed in the HNN-COSY spectrum of 5_SL4. With the exception of C100, all pyrimidine N3 resonances were assigned in the H5(C5C4)N3 spectrum for 5_SL4sh. Out of the additional 15 pyrimidine residues in 5_SL4, N3 signals for 10 were assigned in ^{1}H,^{15}N HSQC and HNN-COSY spectra. All pyrimidine N1 resonances in 5_SL4 were assigned using the carbon detected 3D (H6/8)C6/8N1/9C1′ experiments, which also served link the aromatic carbons to the C1′ resonances for all residues. Using ^{1}H,^{13}C-HSQC and 3D ^{13}C-NOESY-HSQC spectra for the aliphatic region 100% of the H1′–C1′ and the H5–C5 resonances could be assigned. The remaining ribose carbon resonances were assigned using 3D (H)CCH-TOCSY spectra. The complete assignment of the ribose protons was then achieved by 3D HC(C)H-COSY, -TOCSY experiments for 5_SL4 and using 3D ^{13}C-NOESY-HSQC spectra for 5_SL4sh.

Assignment and ribose conformation of the apical loop

5_SL4 is capped by an apical loop comprising nucleotides 104 to 108 with the sequence 5′-U_{104}GCAU_{108}-3′. The sequential assignment of the loop residues solely from NOE contacts was ambiguous. Therefore, we recorded a H(C)P-CCH-TOCSY spectrum for the 5_SL4sh RNA establishing the sequential ribose spin system assignment for this part (Fig. 2a). Furthermore, all 27 ^{31}P resonances for 5_SL4sh were assigned using this spectrum together with the 1D ^{31}P spectrum, which also served to assign the characteristic C25 cyclic phosphate and the α, β and γ phosphate resonance of the 5′-terminal G residue.

Extensive assignment of the ribose carbon spins allows for the extraction of canonical coordinates yielding information about the sugar pucker mode (Cherepanov et al. 2010; Ebrahimi et al. 2001) (Fig. 2b). A C2′-endo conformation was found for the apical loop residues G105 to U108, while

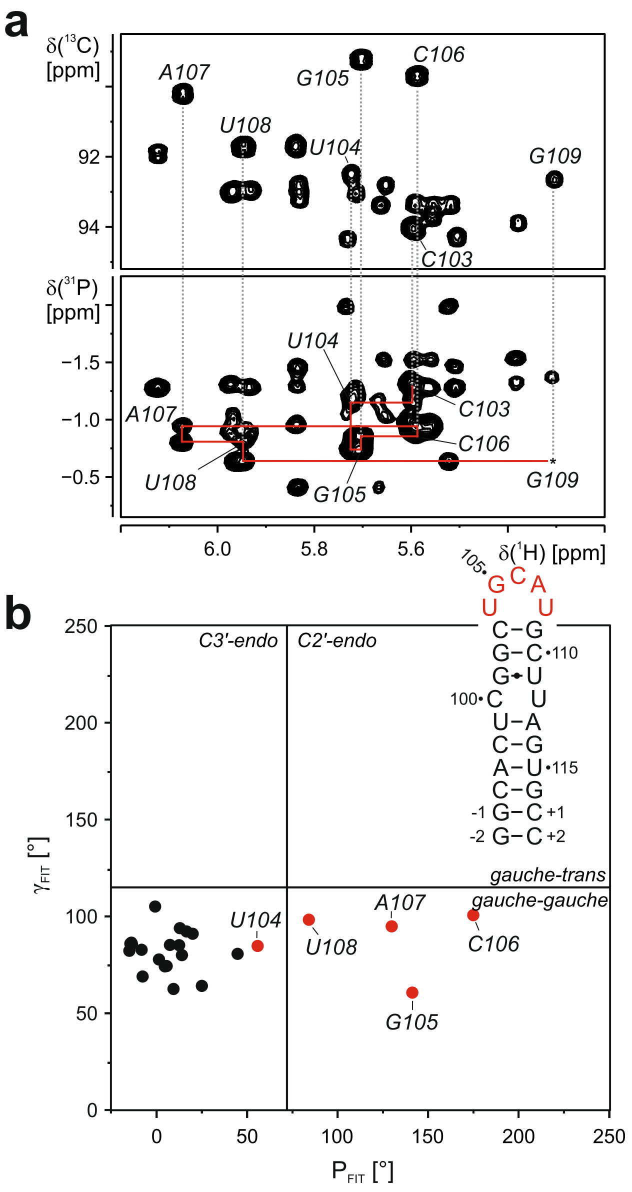


Fig. 2 Assignment and ribose conformation of the 5_SL4 apical loop. **a** H1′C1′ region of a ^{1}H, ^{13}C HSQC spectrum (top) and H(C)P-CCH-TOCSY spectrum (bottom) recorded for the 5_SL4sh RNA. Sequential correlations between the residues in the loop are shown as red lines. **b** Canonical coordinates for all but the 3′-terminal residue of 5_SL4sh The secondary structure of 5_SL4sh is displayed. Residues of the apical loop are highlighted in red and assignments are given

U104 from the loop and all remaining residues in 5_SL4sh adopt a C3′-endo conformation.

Base pairing interactions

In general, the secondary structure of 5_SL4 has been established previously by combination of 2D imino NOESY data with HNN-COSY spectra (Wacker et al. 2020). HNN-COSY spectra verified the identity of the canonical Watson–Crick A–U and G–C base pairs in the stem-loop structure, while of the three potential G–U base pairs, only U87–G124 and G101–U111 could be readily identified by virtue of their strong intra-base-pair imino-imino NOEs in the NOESY spectrum (Wacker et al. 2020). For the remaining G–U base pair (G91–U120), only a very weak guanine imino resonance was identified. For U87 and U111, which possess detectable imino resonances, C2 and C4 shifts could be obtained in an 2D H(N)CO spectrum (data not shown). Downfield C2 and upfield C4 shifts for both U87 and U111 point to a classical wobble-arrangement for the respective G–U base pairs, with the G imino group hydrogen bonded to the U C2. For the G91–U120 base pair, the uridine is missing an imino resonance. In order to investigate the C4 chemical shift of this residue even in the absence of a stable imino resonance, a H5(C5)C4 spectrum was recorded (Fig. 3a). The C4 of U120 has a resonance frequency very similar to both U87 and U111, suggesting a similar Wobble base pair geometry for this G–U base pair.

The carbon and nitrogen chemical shifts for carbon and nitrogen nuclei not directly bound to protons in the C100–U112 mismatch were investigated using 5_SL4sh. Since the imino proton of U112 is not observable the information about this potential base pair has been very limited. The investigation of the smaller construct enabled the additional assignment of quaternary carbon spins in the pyrimidine nucleobases which can be predictive for base functional group hydrogen bonding patterns (Ohlenschläger et al. 2004). Using a 2D H5(C5)C4 spectra, 100% of the pyrimidine C4 carbon atoms were assigned (Fig. 3a). Compared to the C4 carbon chemical shift of U115 and U99 whose C4 carbonyl groups are hydrogen bonded in A–U base pairs, the C4 of U112 is shifted upfield, indicating no involvement of the C4 carbonyl group in hydrogen bonding interactions. A 2D H6(C6N1)C2 experiment was used to identify the C2 resonances of all C and U residues of 5_SL4sh (Fig. 3b). The resonance of U112 is shifted upfield compared to the of U111 in the G–U base pair, for which the C2 carbonyl group is hydrogen bonded to the G101 imino group. Taken together carbonyl chemical shifts for U112 suggests that neither the C2 nor the C4 of this residue is involved in stable hydrogen bond interactions. Using 2D H5(C5C4)N3 spectra, seven out of eight cytidine (Fig. 3c) and 100% of the uridine N3 nitrogen resonances (Fig. 3d) were assigned. The N3 resonance of U112 is shifted upfield compared to the N3 of U99 and U115 that are involved in N–H···N type hydrogen bonds suggesting that the U112 imino group is not involved in such an interaction with C100. The N3 nitrogen of C100 is not detectable. Hence, the structure and putative dynamics of the C100–U112 mismatch are still subject for further investigations.

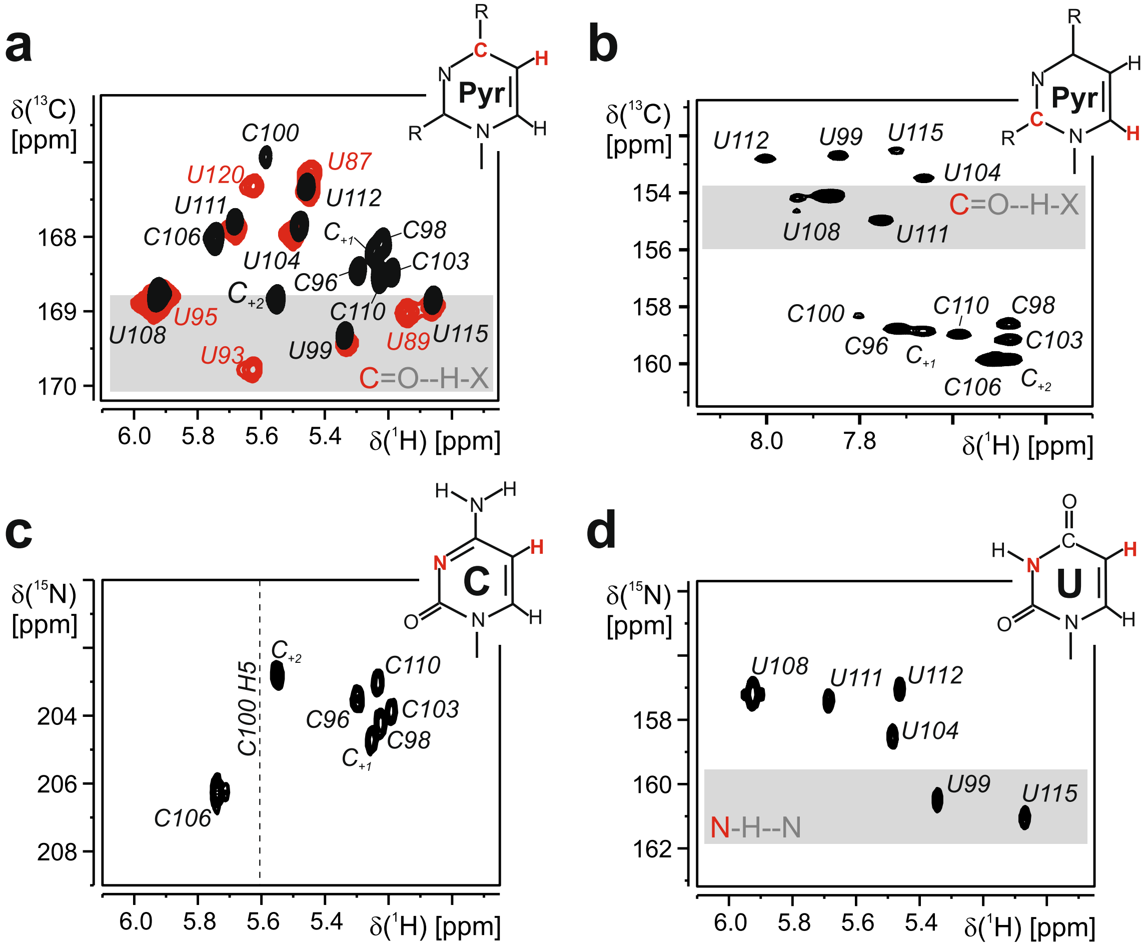


Fig. 3 Carbon and nitrogen chemical shifts of quaternary carbon and the nitrogen spins in the pyrimidine residues. **a** 2D H5(C5)C4 spectrum recorded for 5_SL4 (red) and 5_SL5sh (black). Assignments are given. The chemical shift region of the carbon atoms of carbonyl groups involved in hydrogen bonds is highlighted with a gray bar. The chemical structure of pyrimidine bases is displayed. The H5 and C4 atoms are highlighted in red. **b** H6(C6N1)C2 spectra of 5_SL4sh. Resonance assignments are given. The chemical structure of a pyrimidine base is shown with H6 and C2 highlighted in red. **c, d** 2D H5(C5C4)N3 spectra recorded for the cytidines (**c**) and the uridines (**d**) of 5_SL4sh. The chemical structures of cytidine (C) and uridine (U) bases are displayed. The H5 and N3 atoms are highlighted in red. The chemical shift of the H5 of C100 is shown as a dashed line. The chemical shift region of nitrogen atoms of imino groups that are involved in hydrogen bonds is marked with a gray bar

Data deposition

For 5_SL4, we updated the BMRB deposition with code 50347. For 5_SL4sh a new BMRB entry (50760) was created.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12104-021-10026-7.

Acknowledgements Work at the Center for Biomolecular Magnetic Resonance (BMRZ) at the Goethe-University Frankfurt is supported by the state of Hesse. Work in Covid19-nmr was supported by the Goethe Corona Funds, by the IWB-EFRE-programme 20007375 of the state of Hesse, and the DFG in CRC902: “Molecular Principles of RNA-based regulation.” and infrastructure funds (Project Numbers: 277478796, 277479031, 392682309, 452632086, 70653611). H.S. and B.F. are supported by the DFG in graduate school CLIC (GRK 1986).

Funding Open Access funding enabled and organized by Projekt DEAL.

Declaration

Conflict of interest The authors declare that they have no conflict of interest.

References

Cherepanov AV, Glaubitz C, Schwalbe H (2010) High-resolution studies of uniformly ^{13}C,^{15}N-labeled RNA by solid-state NMR spectroscopy. Angew Chem Int Ed Engl 49:4747–4750. https://doi.org/10.1002/anie.200906885

Ebrahimi M, Rossi P, Rogers C, Harbison GS (2001) Dependence of ^{13}C NMR chemical shifts on conformations of RNA nucleosides and nucleotides. J Magn Reson 150:1–9. https://doi.org/10.1006/jmre.2001.2314

Keller R (2004) The computer aided resonance assignment tutorial. CANTINA verlag, Goldau

Madhugiri R, Fricke M, Marz M, Ziebuhr J (2016) Coronavirus cis-acting RNA elements. Adv Virus Res 96:127–163. https://doi.org/10.1016/bs.aivir.2016.08.007

Maurer T, Kalbitzer HR (1996) Indirect referencing of ^{31}P and ^{19}F NMR Spectra. J Magn Reson B 113:177–178. https://doi.org/10.1006/jmrb.1996.0172

Ohlenschläger O, Wöhnert J, Bucci E, Seitz S, Häfner S, Ramachandran R, Zell R, Görlach M (2004) The structure of the stemloop D subdomain of coxsackievirus B3 cloverleaf RNA and its interaction with the proteinase 3C. Structure 12:237–248. https://doi.org/10.1016/j.str.2004.01.014

Wacker A, Weigand JE, Akabayov SR, Altincekic N, Bains JK, Banijamali E, Binas O, Castillo-Martinez J, Cetiner E, Ceylan B, Chiu L-Y, Davila-Calderon J, Dhamotharan K, Duchardt-Ferner E, Ferner J, Frydman L, Fürtig B, Gallego J, Grün JT, Hacker C, Haddad C, Hähnke M, Hengesbach M, Hiller F, Hohmann KF, Hymon D, de Jesus V, Jonker H, Keller H, Knezic B, Landgraf T, Löhr F, Luo Le, Mertinkus KR, Muhs C, Novakovic M, Oxenfarth A, Palomino-Schätzlein M, Petzold K, Peter SA, Pyper DJ, Qureshi NS, Riad M, Richter C, Saxena K, Schamber T, Scherf T, Schlagnitweit J, Schlundt A, Schnieders R, Schwalbe H, Simba-Lahuasi A, Sreeramulu S, Stirnal E, Sudakov A, Tants J-N, Tolbert BS, Vögele J, Weiß L, Wirmer-Bartoschek J, Wirtz Martin MA, Wöhnert J, Zetzsche H (2020) Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy. Nucleic Acids Res 48:12415–12435. https://doi.org/10.1093/nar/gkaa1013

Wishart DS, Bigam CG, Yao J, Abildgaard F, Dyson HJ, Oldfield E, Markley JL, Sykes BD (1995) 1H, ^{13}C and ^{15}N chemical shift referencing in biomolecular NMR. J Biomol NMR 6:135–140. https://doi.org/10.1007/BF00211777

Wu H-Y, Guan B-J, Su Y-P, Fan Y-H, Brian DA (2014) Reselection of a genomic upstream open reading frame in mouse hepatitis coronavirus 5′-untranslated-region mutants. J Virol 88:846–858. https://doi.org/10.1128/JVI.02831-13

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5.9 **Research article:** Large-Scale Recombinant Production of the SARS-CoV-2 Proteome for High-Throughput and Structural Biology Applications

Nadide Altincekic*, Sophie Marianne Korn*, Nusrat Shahin Qureshi*, Marie Dujardin*, Martí Ninot-Pedrosa*, Rupert Abele, Marie Jose Abi Saad, Caterina Alfano, Fabio C. L. Almeida, Islam Alshamleh, Gisele Cardoso de Amorim, Thomas K. Anderson, Cristiane D. Anobom, Chelsea Anorma, Jasleen Kaur Bains, Adriaan Bax, Martin Blackledge, Julius Blechar, Anja Böckmann, Louis Brigandat, Anna Bula, Matthias Bütikofer, Aldo R. Camacho-Zarco, Teresa Carlomagno, Icaro Putinhon Caruso, Betül Ceylan, Apirat Chaikuad, Feixia Chu, Laura Cole, Marquise G. Crosby, Vanessa de Jesus, Karthikeyan Dhamotharan, Isabella C. Felli, Jan Ferner, Yanick Fleischmann, Marie-Laure Fogeron, Nikolaos K. Fourkiotis, Christin Fuks, Boris Fürtig, Angelo Gallo, Santosh L. Gande, Juan Atilio Gerez, Dhiman Ghosh, Francisco Gomes-Neto, Oksana Gorbatyuk, Serafima Guseva, Carolin Hacker, Sabine Häfner, Bing Hao, Bruno Hargittay, K. Henzler-Wildman, Jeffrey C. Hoch, Katharina F. Hohmann, Marie T. Hutchison, Kristaps Jaudzems, Katarina Jovic, Janina Kaderli, Gints Kalnins, Iveta Kanepe, Robert N. Kirchdoerfer, John Kirkpatrick, Stefan Knapp, Robin Krishnathas, Felicitas Kutz, Susanne zur Lage, Roderick Lambertz, Andras Lang, Douglas Laurents, Lauriane Lecoq, Verena Linhard, Frank Löhr, Anas Malki, Luiza Mamigonian Bessa, Rachel W. Martin, Tobias Matzel, Damien Maurin, Seth W. McNutt, Nathane Cunha Mebus-Antunes, Beat H. Meier, Nathalie Meiser, Miguel Mompeán, Elisa Monaca, Roland Montserret, Laura Mariño Perez, Celine Moser, Claudia Muhle-Goll, Thais Cristtina Neves-Martins, Xiamonin Ni, Brenna Norton-Baker, Roberta Pierattelli, Letizia Pontoriero, Yulia Pustovalova, Oliver Ohlenschläger, Julien Orts, Andrea T. Da Poian, Dennis J. Pyper, Christian Richter, Roland Riek, Chad M. Rienstra, Angus Robertson, Anderson S. Pinheiro, Raffaele Sabbatella, Nicola Salvi, Krishna Saxena, Linda Schulte, Marco Schiavina, Harald Schwalbe, Mara Silber, Marcius da Silva Almeida, Marc A. Sprague-Piercy, Georgios A. Spyroulias, Sridhar Sreeramulu, Jan-Niklas Tants, Kaspars Tars, Felix Torres, Sabrina Töws, Miguel Á. Treviño, Sven Trucks, Aikaterini C. Tsika, Krisztina Varga, Ying Wang, Marco E. Weber, Julia E. Weigand, Christoph Wiedemann, Julia Wirmer-Bartoschek, Maria Alexandra Wirtz Martin, Johannes Zehnder, Martin Hengesbach and Andreas Schlundt; *Frontiers in Molecular Biosciences* **8**, 653148 (2021)

* These authors contributed equally to this work.

This article describes the protocols for the large-scale production of more than 80% of all SARS-CoV-2 proteins for expression and purification of isotope-labeled protein samples for NMR-based applications. The protocols were established for a total of 10 full-length and 13 truncated non-structural protein constructs, as well as nine full-length and seven truncated accessory protein constructs.

The project was designed and coordinated by the Covid-19 NMR consortium. The manuscript was prepared by N. Altincekic, S. M. Korn, N. S. Qureshi, M. Dujardin and M. Ninot-Pedrosa. The author of this thesis was part of the Covid-19 NMR consortium and the protein production team that prepared and purified various isotope-labeled protein samples for NMR spectroscopy.

https://doi.org/10.3389/fmolb.2021.653148

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 10 May 2021
doi: 10.3389/fmolb.2021.653148

Edited by:
Qian Han,
Hainan University, China

Reviewed by:
David Douglas Boehr,
Pennsylvania State University, United States
Luis G. Brieba,
National Polytechnic Institute of Mexico
(CINVESTAV), Mexico

***Correspondence:**
Anja Böckmann
a.bockmann@ibcp.fr
Harald Schwalbe
schwalbe@nmr.uni-frankfurt.de
Martin Hengesbach
hengesbach@nmr.uni-frankfurt.de
Andreas Schlundt
schlundt@bio.uni-frankfurt.de

[†]These authors have contributed equally to this work and share first authorship
[‡]These authors share last authorship

Specialty section:
This article was submitted to
Structural Biology,
a section of the journal
Frontiers in Molecular Biosciences

Received: 13 January 2021
Accepted: 04 February 2021
Published: 10 May 2021

Citation:
Altincekic N, Korn SM, Qureshi NS, Dujardin M, Ninot-Pedrosa M, Abele R, Abi Saad MJ, Alfano C, Almeida FCL, Alshamleh I, de Amorim GC, Anderson TK, Anobom CD, Anorma C, Bains JK, Bax A, Blackledge M, Blechar J, Böckmann A, Brigandat L, Bula A, Bütikofer M, Camacho-Zarco AR, Carlomagno T, Caruso IP, Ceylan B, Chaikuad A, Chu F, Cole L, Crosby MG, de Jesus V, Dhamotharan K, Felli IC, Ferner J, Fleischmann Y, Fogeron M-L, Fourkiotis NK, Fuks C, Fürtig B, Gallo A, Gande SL, Gerez JA, Ghosh D, Gomes-Neto F, Gorbatyuk O, Guseva S, Hacker C, Häfner S, Hao B, Hargittay B, Henzler-Wildman K, Hoch JC, Hohmann KF, Hutchison MT, Jaudzems K, Jović K, Kaderli J, Kalniņš G, Kaņepe I, Kirchdoerfer RN, Kirkpatrick J, Knapp S, Krishnathas R, Kutz F, zur Lage S, Lambertz R, Lang A, Laurents D, Lecoq L, Linhard V, Löhr F, Malki A, Bessa LM, Martin RW, Matzel T, Maurin D, McNutt SW, Mebus-Antunes NC, Meier BH, Meiser N, Mompeán M, Monaca E, Montserret R, Mariño Perez L, Moser C, Muhle-Goll C, Neves-Martins TC, Ni X, Norton-Baker B, Pierattelli R, Pontoriero L, Pustovalova Y, Ohlenschläger O, Orts J, Da Poian AT, Pyper DJ, Richter C, Riek R, Rienstra CM, Robertson A, Pinheiro AS, Sabbatella R, Salvi N, Saxena K, Schulte L, Schiavina M, Schwalbe H, Silber M, Almeida MdS, Sprague-Piercy MA, Spyroulias GA, Sreeramulu S, Tants J-N, Tārs K, Torres F, Töws S, Treviño MÁ, Trucks S, Tsika AC, Varga K, Wang Y, Weber ME, Weigand JE, Wiedemann C, Wirmer-Bartoschek J, Wirtz Martin MA, Zehnder J, Hengesbach M and Schlundt A (2021) Large-Scale Recombinant Production of the SARS-CoV-2 Proteome for High-Throughput and Structural Biology Applications.
Front. Mol. Biosci. 8:653148.
doi: 10.3389/fmolb.2021.653148

Large-Scale Recombinant Production of the SARS-CoV-2 Proteome for High-Throughput and Structural Biology Applications

Nadide Altincekic[1,2†], Sophie Marianne Korn[2,3†], Nusrat Shahin Qureshi[1,2†], Marie Dujardin[4†], Martí Ninot-Pedrosa[4†], Rupert Abele[5], Marie Jose Abi Saad[6], Caterina Alfano[7], Fabio C. L. Almeida[8,9], Islam Alshamleh[1,2], Gisele Cardoso de Amorim[8,10], Thomas K. Anderson[11], Cristiane D. Anobom[8,12], Chelsea Anorma[13], Jasleen Kaur Bains[1,2], Adriaan Bax[14], Martin Blackledge[15], Julius Blechar[1,2], Anja Böckmann[4‡], Louis Brigandat[4], Anna Bula[16], Matthias Bütikofer[6], Aldo R. Camacho-Zarco[15], Teresa Carlomagno[17,18], Icaro Putinhon Caruso[8,9,19], Betül Ceylan[1,2], Apirat Chaikuad[20,21], Feixia Chu[22], Laura Cole[4], Marquise G. Crosby[23], Vanessa de Jesus[1,2], Karthikeyan Dhamotharan[2,3], Isabella C. Felli[24,25], Jan Ferner[1,2], Yanick Fleischmann[6], Marie-Laure Fogeron[4], Nikolaos K. Fourkiotis[26], Christin Fuks[1], Boris Fürtig[1,2], Angelo Gallo[26], Santosh L. Gande[1,2], Juan Atilio Gerez[6], Dhiman Ghosh[6], Francisco Gomes-Neto[8,27], Oksana Gorbatyuk[28], Serafima Guseva[15], Carolin Hacker[29], Sabine Häfner[30], Bing Hao[28], Bruno Hargittay[1,2], K. Henzler-Wildman[11], Jeffrey C. Hoch[28], Katharina F. Hohmann[1,2], Marie T. Hutchison[1,2], Kristaps Jaudzems[16], Katarina Jović[22], Janina Kaderli[6], Gints Kalniņš[31], Iveta Kaņepe[16], Robert N. Kirchdoerfer[11], John Kirkpatrick[17,18], Stefan Knapp[20,21], Robin Krishnathas[1,2], Felicitas Kutz[1,2], Susanne zur Lage[18], Roderick Lambertz[3], Andras Lang[30], Douglas Laurents[32], Lauriane Lecoq[4], Verena Linhard[1,2], Frank Löhr[2,33], Anas Malki[15], Luiza Mamigonian Bessa[15], Rachel W. Martin[13,23], Tobias Matzel[1,2], Damien Maurin[15], Seth W. McNutt[22], Nathane Cunha Mebus-Antunes[8,9], Beat H. Meier[6], Nathalie Meiser[1], Miguel Mompeán[32], Elisa Monaca[7], Roland Montserret[4], Laura Mariño Perez[15], Celine Moser[34], Claudia Muhle-Goll[34], Thais Cristtina Neves-Martins[8,9], Xiamonin Ni[20,21], Brenna Norton-Baker[13], Roberta Pierattelli[24,25], Letizia Pontoriero[24,25], Yulia Pustovalova[28], Oliver Ohlenschläger[30], Julien Orts[6], Andrea T. Da Poian[9], Dennis J. Pyper[1,2], Christian Richter[1,2], Roland Riek[6], Chad M. Rienstra[35], Angus Robertson[14], Anderson S. Pinheiro[8,12], Raffaele Sabbatella[7], Nicola Salvi[15], Krishna Saxena[1,2], Linda Schulte[1,2], Marco Schiavina[24,25], Harald Schwalbe[1,2*‡], Mara Silber[34], Marcius da Silva Almeida[8,9], Marc A. Sprague-Piercy[23], Georgios A. Spyroulias[26], Sridhar Sreeramulu[1,2], Jan-Niklas Tants[2,3], Kaspars Tārs[31], Felix Torres[6], Sabrina Töws[3], Miguel Á. Treviño[32], Sven Trucks[1], Aikaterini C. Tsika[26], Krisztina Varga[22], Ying Wang[17], Marco E. Weber[6], Julia E. Weigand[36], Christoph Wiedemann[37], Julia Wirmer-Bartoschek[1,2], Maria Alexandra Wirtz Martin[1,2], Johannes Zehnder[6], Martin Hengesbach[1*‡] and Andreas Schlundt[2,3*‡]*

[1]Institute for Organic Chemistry and Chemical Biology, Goethe University Frankfurt, Frankfurt am Main, Germany, [2]Center of Biomolecular Magnetic Resonance (BMRZ), Goethe University Frankfurt, Frankfurt am Main, Germany, [3]Institute for Molecular Biosciences, Goethe University Frankfurt, Frankfurt am Main, Germany, [4]Molecular Microbiology and Structural Biochemistry, UMR 5086, CNRS/Lyon University, Lyon, France, [5]Institute for Biochemistry, Goethe University Frankfurt, Frankfurt am Main, Germany, [6]Swiss Federal Institute of Technology, Laboratory of Physical Chemistry, ETH Zurich, Zurich, Switzerland, [7]Structural Biology and Biophysics Unit, Fondazione Ri.MED, Palermo, Italy, [8]National Center of Nuclear Magnetic Resonance (CNRMN, CENABIO), Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, [9]Institute of Medical Biochemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, [10]Multidisciplinary Center for Research in Biology (NUMPEX), Campus Duque de Caxias Federal University of Rio de Janeiro, Duque de Caxias, Brazil, [11]Institute for Molecular Virology, University of Wisconsin-Madison, Madison, WI, United States, [12]Institute of Chemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, [13]Department of Chemistry, University of California, Irvine, CA, United States, [14]LCP, NIDDK, NIH, Bethesda, MD, United States, [15]Univ. Grenoble Alpes, CNRS, CEA, IBS, Grenoble, France, [16]Latvian Institute of Organic Synthesis, Riga, Latvia, [17]BMWZ and Institute of Organic

Chemistry, Leibniz University Hannover, Hannover, Germany, [18]Group of NMR-Based Structural Chemistry, Helmholtz Centre for Infection Research, Braunschweig, Germany, [19]Multiuser Center for Biomolecular Innovation (CMIB), Department of Physics, São Paulo State University (UNESP), São José do Rio Preto, Brazil, [20]Institute of Pharmaceutical Chemistry, Goethe University Frankfurt, Frankfurt am Main, Germany, [21]Structural Genomics Consortium, Buchmann Institute for Molecular Life Sciences, Frankfurt am Main, Germany, [22]Department of Molecular, Cellular, and Biomedical Sciences, University of New Hampshire, Durham, NH, United States, [23]Department of Molecular Biology and Biochemistry, University of California, Irvine, CA, United States, [24]Magnetic Resonance Centre (CERM), University of Florence, Sesto Fiorentino, Italy, [25]Department of Chemistry "Ugo Schiff", University of Florence, Sesto Fiorentino, Italy, [26]Department of Pharmacy, University of Patras, Patras, Greece, [27]Laboratory of Toxinology, Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil, [28]Department of Molecular Biology and Biophysics, UConn Health, Farmington, CT, United States, [29]Signals GmbH & Co. KG, Frankfurt am Main, Germany, [30]Leibniz Institute on Aging—Fritz Lipmann Institute (FLI), Jena, Germany, [31]Latvian Biomedical Research and Study Centre, Riga, Latvia, [32]"Rocasolano" Institute for Physical Chemistry (IQFR), Spanish National Research Council (CSIC), Madrid, Spain, [33]Institute of Biophysical Chemistry, Goethe University Frankfurt, Frankfurt am Main, Germany, [34]IBG-4, Karlsruhe Institute of Technology, Karlsruhe, Germany, [35]Department of Biochemistry and National Magnetic Resonance Facility at Madison, University of Wisconsin-Madison, Madison, WI, United States, [36]Department of Biology, Technical University of Darmstadt, Darmstadt, Germany, [37]Institute of Biochemistry and Biotechnology, Charles Tanford Protein Centre, Martin Luther University Halle-Wittenberg, Halle/Saale, Germany

The highly infectious disease COVID-19 caused by the *Betacoronavirus* SARS-CoV-2 poses a severe threat to humanity and demands the redirection of scientific efforts and criteria to organized research projects. The international *COVID19-NMR* consortium seeks to provide such new approaches by gathering scientific expertise worldwide. In particular, making available viral proteins and RNAs will pave the way to understanding the SARS-CoV-2 molecular components in detail. The research in *COVID19-NMR* and the resources provided through the consortium are fully disclosed to accelerate access and exploitation. NMR investigations of the viral molecular components are designated to provide the essential basis for further work, including macromolecular interaction studies and high-throughput drug screening. Here, we present the extensive catalog of a holistic SARS-CoV-2 protein preparation approach based on the consortium's collective efforts. We provide protocols for the large-scale production of more than 80% of all SARS-CoV-2 proteins or essential parts of them. Several of the proteins were produced in more than one laboratory, demonstrating the high interoperability between NMR groups worldwide. For the majority of proteins, we can produce isotope-labeled samples of HSQC-grade. Together with several NMR chemical shift assignments made publicly available on *covid19-nmr.com*, we here provide highly valuable resources for the production of SARS-CoV-2 proteins in isotope-labeled form.

Keywords: COVID-19, SARS-CoV-2, nonstructural proteins, structural proteins, accessory proteins, intrinsically disordered region, cell-free protein synthesis, NMR spectroscopy

INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, SCoV2) is the cause of the early 2020 pandemic coronavirus lung disease 2019 (COVID-19) and belongs to *Betacoronaviruses*, a genus of the Coronaviridae family covering the α–δ genera (Leao et al., 2020). The large RNA genome of SCoV2 has an intricate, highly condensed arrangement of coding sequences (Wu et al., 2020). Sequences starting with the main start codon contain an open reading frame 1 (ORF1), which codes for two distinct, large polypeptides (pp), whose relative abundance is governed by the action of an RNA pseudoknot structure element. Upon RNA folding, this element causes a −1 frameshift to allow the continuation of translation, resulting in the generation of a 7,096-amino acid 794 kDa polypeptide. If the pseudoknot is not formed, expression of the first ORF generates a 4,405-amino acid 490 kDa polypeptide. Both the short and long polypeptides translated from this ORF (pp1a and pp1ab, respectively) are posttranslationally cleaved by virus-encoded proteases into functional, nonstructural proteins (nsps). ORF1a encodes eleven nsps, and ORF1ab additionally encodes the nsps 12–16. The downstream ORFs encode structural proteins (S, E, M, and N) that are essential components for the synthesis of new virus particles. In between those, additional proteins (accessory/auxiliary factors) are encoded, for which sequences partially overlap (Finkel et al., 2020) and whose identification and classification are a matter of ongoing research (Nelson et al., 2020; Pavesi, 2020). In total, the number of identified peptides or proteins generated from the viral genome is at least 28 on the evidence level, with an additional set of smaller proteins or peptides being predicted with high likelihood.

High-resolution studies of SCoV and SCoV2 proteins have been conducted using all canonical structural biology approaches, such as X-ray crystallography on proteases (Zhang et al., 2020) and methyltransferases (MTase) (Krafcikova et al., 2020), cryo-EM of the RNA polymerase (Gao et al., 2020; Yin et al., 2020), and liquid-state (Almeida et al., 2007; Serrano et al., 2009; Cantini et al., 2020; Gallo et al., 2020; Korn et al., 2020a; Korn et al., 2020b;

TABLE 1 | SCoV2 protein constructs expressed and purified, given with the genomic position and corresponding PDBs for construct design.

Protein genome position (nt)[a]	Trivial name construct expressed	Size (aa)	Boundaries	MW (kDa)	Homol. SCoV (%)[b]	Template PDB[c]	SCoV2 PDB[d]
nsp1 *266–805*	**Leader**	**180**		**19.8**	**84**		
	Full-length	180	1–180	19.8	83		
	Globular domain (GD)	116	13–127	12.7	85	2GDT	7K7P
nsp2 *806–2,719*		**638**		**70.5**	**68**		
	C-terminal IDR (CtDR)	45	557–601	4.9	55		
nsp3 *2,720–8,554*		**1,945**		**217.3**	**76**		
a	Ub-like (Ubl) domain	111	1–111	12.4	79	2IDY	7KAG
a	Ub-like (Ubl) domain + IDR	206	1–206	23.2	58		
b	Macrodomain	170	207–376	18.3	74	6VXS	6VXS
c	SUD-N	140	409–548	15.5	69	2W2G	
c	SUD-NM	267	409–675	29.6	74	2W2G	
c	SUD-M	125	551–675	14.2	82	2W2G	
c	SUD-MC	195	551–743	21.9	79	2KQV	
c	SUD-C	64	680–743	7.4	73	2KAF	
d	Papain-like protease PL^{pro}	318	743–1,060	36	83	6W9C	6W9C
e	NAB	116	1,088–1,203	13.4	87	2K87	
Y	CoV-Y	308	1,638–1,945	34	89		
nsp5 *10,055–10,972*	**Main protease (M^{pro})**	**306**		**33.7**	**96**		
	Full-length[e]	306	1–306	33.7	96	6Y84	6Y84
nsp7 *11,843–12,091*		**83**		**9.2**	**99**		
	Full-length	83	1–83	9.2	99	6WIQ	6WIQ
nsp8 *12,092–12,685*		**198**		**21.9**	**98**		
	Full-length	198	1–198	21.9	97	6WIQ	6WIQ
nsp9 *12,686–13,024*		**113**		**12.4**	**97**		
	Full-length	113	1–113	12.4	97	6W4B	6W4B
nsp10 *13,025–13,441*		**139**		**14.8**	**97**		
	Full-length	139	1–139	14.8	97	6W4H	6W4H
nsp13 *16,237–18,039*	**Helicase**	**601**		**66.9**	**100**		
	Full-length	601	1–601	66.9	100	6ZSL	6ZSL
nsp14 *18,040–19,620*	**Exonuclease/ methyltransferase**	**527**		**59.8**	**95**		
	Full-length	527	1–527	59.8	95	5NFY	
	MTase domain	240	288–527	27.5	95		
nsp15 *19,621–20,658*	**Endonuclease**	**346**		**38.8**	**89**		
	Full-length	346	1–346	38.8	89	6W01	6W01
nsp16 *20,659–21,552*	**Methyltransferase**	**298**		**33.3**	**93**		
	Full-length	298	1–298	33.3	93	6W4H	6W4H
ORF3a *25,393–26,220*		**275**		**31.3**	**72**		
	Full-length	275	1–275	31.3	72	6XDC	6XDC
ORF4 *26,245–26,472*	**Envelope (E) protein**	**75**		**8.4**	**95**		
	Full-length	75	1–75	8.4	95	5X29	7K3G
ORF5 *26,523–27,387*	**Membrane glycoprotein (M)**	**222**		**25.1**	**91**		
	Full-length	222	1–222	25.1	91		
ORF6 *27,202–27,387*		**61**		**7.3**	**69**		
	Full-length	61	1–61	7.3	69		

(Continued on following page)

TABLE 1 | (*Continued*) SCoV2 protein constructs expressed and purified, given with the genomic position and corresponding PDBs for construct design.

Protein genome position (nt)[a]	Trivial name construct expressed	Size (aa)	Boundaries	MW (kDa)	Homol. SCoV (%)[b]	Template PDB[c]	SCoV2 PDB[d]
ORF7a *27,394–27,759*		**121**		**13.7**	**85**		
	Ectodomain (ED)	66	16–81	7.4	85	1XAK	6W37
ORF7b *27,756–27,887*		**43**		**5.2**	**85**		
	Full-length	43	1–43	5.2	85		
ORF8 *27,894–28,259*		**121**		**13.8**	**32**		
ORF8	Full-length	121	1–121	13.8	32		
ΔORF8	w/o signal peptide	106	16–121	12	41	7JTL	7JTL
ORF9a *28,274–29,533*	**Nucleocapsid (N)**	**419**		**45.6**	**91**		
	IDR1-NTD-IDR2	248	1–248	26.5	90		
	NTD-SR	169	44–212	18.1	92		
	NTD	136	44–180	14.9	93	6YI3	6YI3
	CTD	118	247–364	13.3	96	2JW8	7C22
ORF9b *28,284–28,574*		**97**		**10.8**	**72**		
	Full-length	97	1–97	10.8	72	6Z4U	6Z4U
ORF14 *28,734–28,952*		**73**		**8**	**n.a**		
	Full-length	73	1–73	8	n.a		
ORF10 *29,558–29,674*		**38**		**4.4**	**29**		
	Full-length	38	1–38	4.4	29		

[a] *Genome position in nt corresponding to SCoV2 NCBI reference genome entry NC_045512.2, identical to GenBank entry MN908947.3.*
[b] *Sequence identities to SCoV are calculated from an alignment with corresponding protein sequences based on the genome sequence of NCBI Reference NC_004718.3.*
[c] *Representative PDB that was available at the beginning of construct design, either SCoV or SCoV2.*
[d] *Representative PDB available for SCoV2 (as of December 2020).*
[e] *Additional point mutations in fl-construct have been expressed.*
n.a.: not applicable.

Kubatova et al., 2020; Tonelli et al., 2020) and solid-state NMR spectroscopy of transmembrane (TM) proteins (Mandala et al., 2020). These studies have significantly improved our understanding on the functions of molecular components, and they all rely on the recombinant production of viral proteins in high amount and purity.

Apart from structures, purified SCoV2 proteins are required for experimental and preclinical approaches designed to understand the basic principles of the viral life cycle and processes underlying viral infection and transmission. Approaches range from studies on immune responses (Esposito et al., 2020), antibody identification (Jiang et al., 2020), and interactions with other proteins or components of the host cell (Bojkova et al., 2020; Gordon et al., 2020). These examples highlight the importance of broad approaches for the recombinant production of viral proteins.

The research consortium *COVID19-NMR* founded in 2020 seeks to support the search for antiviral drugs using an NMR-based screening approach. This requires the large-scale production of all druggable proteins and RNAs and their NMR resonance assignments. The latter will enable solution structure determination of viral proteins and RNAs for rational drug design and the fast mapping of compound binding sites. We have recently produced and determined secondary structures of SCoV2 RNA *cis*-regulatory elements in near completeness by NMR spectroscopy, validated by DMS-MaPseq (Wacker et al., 2020), to provide a basis for RNA-oriented fragment screens with NMR.

We here compile a compendium of more than 50 protocols (see **Supplementary Tables SI1–SI23**) for the production and purification of 23 of the 30 SCoV2 proteins or fragments thereof (summarized in **Tables 1**, **2**). We defined those 30 proteins as existing or putative ones to our current knowledge (see later discussion). This compendium has been generated in a coordinated and concerted effort between >30 labs worldwide (**Supplementary Table S1**), with the aim of providing pure mg amounts of SCoV2 proteins. Our protocols include the rational strategy for construct design (if applicable, guided by available homolog structures), optimization of expression, solubility, yield, purity, and suitability for follow-up work, with a focus on uniform stable isotope-labeling.

We also present protocols for a number of accessory and structural E and M proteins that could only be produced using wheat-germ cell-free protein synthesis (WG-CFPS). In SCoV2, accessory proteins represent a class of mostly small and relatively poorly characterized proteins, mainly due to their difficult behavior in classical expression systems. They are often found in inclusion bodies and difficult to purify in quantities adequate for structural studies. We thus here exploit cell-free synthesis, mainly based on previous reports on production and purification of viral membrane proteins in general (Fogeron et al., 2015b; Fogeron et al., 2017; Jirasko

TABLE 2 | Summary of SCoV2 protein production results in *Covid19-NMR*.

Construct expressed	Yields (mg/L)[a] or *(mg/ml)*[b]	Results	Comments	BMRB	Supplementary Material
nsp1					SI1
fl	5	NMR assigned	Expression only at >20°C; after 7 days at 25°C partial proteolysis	50620[d]	
GD	>0.5	HSQC	High expression; mainly insoluble; higher salt increases stability (>250 mM)		
nsp2					SI2
CtDR	0.7–1.5	NMR assigned	Assignment with His-tag shown in (Mompean et al., 2020)	50687[c]	
nsp3					SI3
UBl	0.7	HSQC	Highly stable over weeks; spectrum overlays with Ubl + IDR		
UBl + IDR	2–3	NMR assigned	Highly stable for >2 weeks at 25°C	50446[d]	
Macrodomain	9	NMR assigned	Highly stable for >1 week at 25°C and > 2 weeks at 4°C	50387[d] 50388[d]	
SUD-N	14	NMR assigned	Highly stable for >10 days at 25°C	50448[d]	
SUD-NM	17	HSQC	Stable for >1 week at 25°C		
SUD-M	8.5	NMR assigned	Significant precipitation during measurement; tendency to dimerize	50516[d]	
SUD-MC	12	HSQC	Stable for >1 week at 25°C		
SUD-C	4.7	NMR assigned	Stable for >10 days at 25°C	50517[d]	
PL^{pro}	12	HSQC	Solubility-tag essential for expression; tendency to aggregate		
NAB	3.5	NMR assigned	Highly stable for >1 week at 25°C; stable for >5 weeks at 4°C	50334[d]	
CoV-Y	12	HSQC	Low temperature (<25°C) and low concentrations (<0.2 mM) favor stability; gradual degradation at 25°C; lithium bromide in final buffer supports solubility		
nsp5					SI4
fl	55	HSQC	Impaired dimerization induced by artificial N-terminal residues		
nsp7					SI5
fl	17	NMR assigned	Stable for several days at 35°C; stable for >1 month at 4°C	50337[d]	
nsp8					SI6
fl	17	HSQC	Concentration dependent aggregation; low concentrations favor stability		
nsp9					SI7
fl	4.5	NMR assigned	Stable dimer for >4 months at 4°C and >2 weeks at 25°C	50621[d] 50622[d] 50513	
nsp10					SI8
fl	15	NMR assigned	Zn^{2+} addition during expression and purification increases protein stability; stable for >1 week at 25°C	50392	
nsp13					SI9
fl	0.5	HSQC	Low expression; protein unstable; concentration above 20 µM not possible		
nsp14					SI10
fl	6	Pure protein	Not above 50 µM; best storage: with 50% (v/v) glycerol; addition of reducing agents		
MTase	10	Pure protein	As fl nsp14; high salt (>0.4 M) for increased stability; addition of reducing agents		
nsp15					SI11
fl	5	HSQC	Tendency to aggregate at 25°C		
nsp16					SI12
fl	10	Pure protein	Addition of reducing agents; 5% (v/v) glycerol favorable; highly unstable		
ORF3a					SI13
fl	*0.6*	Pure protein	Addition of detergent during expression (0.05% Brij-58); stable protein		
E protein					SI14
fl	*0.45*	Pure protein	Addition of detergent during expression (0.05% Brij-58); stable protein		

(Continued on following page)

TABLE 2 | (*Continued*) Summary of SCoV2 protein production results in *Covid19-NMR*.

Construct expressed	Yields (mg/L)[a] or *(mg/ml)*[b]	Results	Comments	BMRB	Supplementary Material
M Protein					SI15
fl	*0.33*	Pure protein	Addition of detergent during expression (0.05% Brij-58); stable protein		
ORF6					SI16
fl	*0.27*	HSQC	Soluble expression without detergent; stable protein; no expression with STREP-tag at N-terminus		
ORF7a					SI17
ED	0.4	HSQC	Unpurified protein tends to precipitate during refolding, purified protein stable for 4 days at 25°C		
ORF7b					SI18
fl	0.6	HSQC	Tendency to oligomerize; solubilizing agents needed		
fl	*0.27*	HSQC	Addition of detergent during expression (0.1% MNG-3); stable protein		
ORF8					SI19
fl	*0.62*	HSQC	Tendency to oligomerize		
ΔORF8	0.5	Pure protein			
N protein					SI20
IDR1-NTD-IDR2	12	NMR assigned	High salt (>0.4 M) for increased stability	50618, 50619, 50558, 50557[d]	
NTD-SR	3	HSQC			
NTD	3	HSQC		34511	
CTD	2	NMR assigned	Stable dimer for >4 months at 4°C and >3 weeks at 30°C	50518[d]	
ORF9b					SI21
fl	*0.64*	HSQC	Expression without detergent, protein is stable		
ORF14					SI22
fl	*0.43*	HSQC	Addition of detergent during expression (0.05% Brij-58); stable in detergent but unstable on lipid reconstitution		
ORF10					SI23
fl	2	HSQC	Tendency to oligomerize; unstable upon tag cleavage		

[a] *Yields from bacterial expression represent the minimal protein amount in mg/L independent of the cultivation medium. Italic values indicate yields from CFPS.*

[b] *Yields from CFPS represent the minimal protein amount in mg/ml of wheat-germ extract.*

[c] COVID19-nmr *BMRB depositions yet to be released.*

[d] COVID19-nmr *BMRB depositions.*

et al., 2020b). Besides yields compatible with structural studies, ribosomes in WG extracts further possess an increased folding capacity (Netzer and Hartl, 1997), favorable for those more complicated proteins.

We exemplify in more detail the optimization of protein production, isotope-labeling, and purification for proteins with different individual challenges: the nucleic acid–binding (NAB) domain of nsp3e, the main protease nsp5, and several auxiliary proteins. For the majority of produced and purified proteins, we achieve >95% purity and provide ^{15}N-HSQC spectra as the ultimate quality measure. We also provide additional suggestions for challenging proteins, where our protocols represent a unique resource and starting point exploitable by other labs.

MATERIALS AND METHODS

Strains, Plasmids, and Cloning

The rationale of construct design for all proteins can be found within the respective protocols in **Supplementary Tables SI1–SI23**. For bacterial production, *E. coli* strains and expression plasmids are given; for WG-CFPS, template vectors are listed. Protein coding sequences of interest have been obtained as either commercial, codon-optimized genes or, for shorter ORFs and additional sequences, annealed from oligonucleotides prior to insertion into the relevant vector. Subcloning of inserts, adjustment of boundaries, and mutations of genes have been carried out by standard molecular biology techniques. All expression plasmids can be obtained upon request from the *COVID19-NMR* consortium (https://covid19-nmr.com/), including information about coding sequences, restriction sites, fusion tags, and vector backbones.

Protein Production and Purification

For SCoV2 proteins, we primarily used heterologous production in *E. coli*. Detailed protocols of individual full-length (fl) proteins, separate domains, combinations, or particular expression constructs as listed in **Table 1** can be found in the (**Supplementary Tables SI1–SI23**).

The ORF3a, ORF6, ORF7b, ORF8, ORF9b, and ORF14 accessory proteins and the structural proteins M and E were produced by WG-CFPS as described in the **Supplementary Material**. In brief, transcription and translation steps have

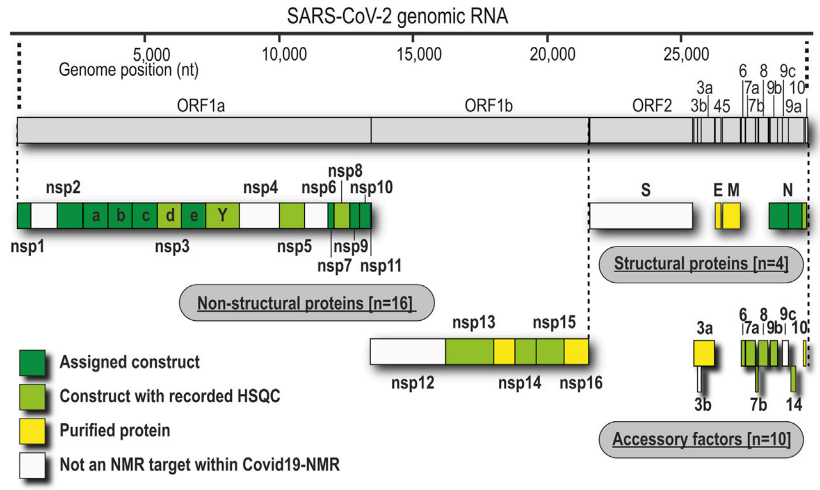


FIGURE 1 | Genomic organization of proteins and current state of analysis or purification. Boxes represent the domain boundaries as outlined in the text and in **Table 1**. Their position corresponds with the genomic loci. Colors indicate whether the pure proteins were purified (yellow), analyzed by NMR using only HSQC (lime), or characterized in detail, including NMR resonance assignments (green).

been performed separately, and detergent has been added for the synthesis of membrane proteins as described previously (Takai et al., 2010; Fogeron et al., 2017).

NMR Spectroscopy

All amide correlation spectra, either HSQC- or TROSY-based, are representative examples. Details on their acquisition parameters and the raw data are freely accessible through https://covid19-nmr.de or upon request.

RESULTS

In the following, we provide protocols for the purification of SCoV2 proteins sorted into 1) nonstructural proteins and 2) structural proteins together with accessory ORFs. **Table 1** shows an overview of expression constructs. We use a consequent terminology of those constructs, which is guided by domains, intrinsically disordered regions (IDRs) or other particularly relevant sequence features within them. This study uses the SCoV2 NCBI reference genome entry NC_045512.2, identical to GenBank entry MN908947.3 (Wu et al., 2020), unless denoted differently in the respective protocols. Any relevant definition of boundaries can also be found in the SI protocols.

As applicable for a major part of our proteins, we further define a standard procedure for the purification of soluble His-tagged proteins that are obtained through the sequence of **I**MAC, TEV/Ulp1 **P**rotease cleavage, **R**everse IMAC, and **S**ize-exclusion chromatography, eventually with individual alterations, modifications, or additional steps. For convenient reading, we will thus use the abbreviation IPRS to avoid redundant protocol description. Details for every protein, including detailed expression conditions, buffers, incubation times, supplements, storage conditions, yields, and stability, can be found in the respective **Supplementary Tables SI1–SI23** (see also **Supplementary Tables S1, S2**) and **Tables 1**, **2**.

Nonstructural Proteins

We have approached and challenged the recombinant production of a large part of the SCoV2 nsps (**Figure 1**), with great success (**Table 2**). We excluded nsp4 and nsp6 (TM proteins), which are little characterized and do not reveal soluble, folded domains by prediction (Oostra et al., 2007; Oostra et al., 2008). The function of the very short (13 aa) nsp11 is unknown, and it seems to be a mere copy of the nsp12 amino-terminal residues, remaining as a protease cleavage product of ORF1a. Further, we left out the RNA-dependent RNA polymerase nsp12 in our initial approach because of its size (>100 kDa) and known unsuitability for heterologous recombinant production in bacteria. Work on NMR-suitable nsp12 bacterial production is ongoing, while other expert labs have succeeded in purifying nsp12 for cryo-EM applications in different systems (Gao et al., 2020; Hillen et al., 2020). For the remainder of nsps, we here provide protocols for fl-proteins or relevant fragments of them.

nsp1

nsp1 is the very N-terminus of the polyproteins pp1a and pp1ab and one of the most enigmatic viral proteins, expressed only in α- and β-CoVs (Narayanan et al., 2015). Interestingly, nsp1 displays the highest divergence in sequence and size among different CoVs, justifying it as a genus-specific marker (Snijder et al., 2003). It functions as a host shutoff factor by suppressing innate immune functions and host gene expression (Kamitani et al.,

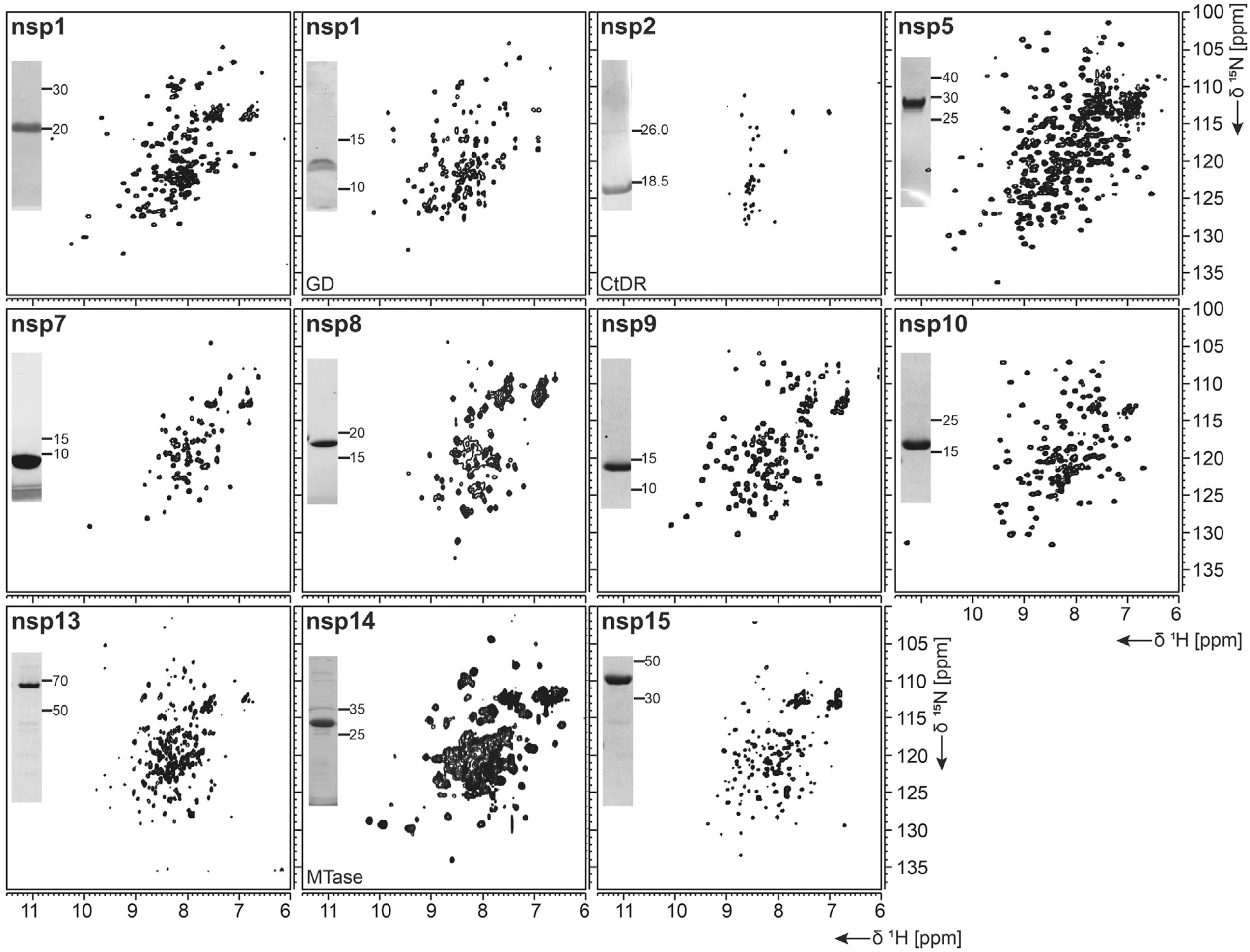


FIGURE 2 | ^{1}H, ^{15}N-correlation spectra of investigated nonstructural proteins. Construct names according to **Table 1** are indicated unless fl-proteins are shown. A representative SDS-PAGE lane with final samples is included as inset. Spectra for nsp3 constructs are collectively shown in **Figure 3**.

2006; Narayanan et al., 2008; Schubert et al., 2020). This suppression is achieved by an interaction of the nsp1 C-terminus with the mRNA entry tunnel within the 40 S subunit of the ribosome (Schubert et al., 2020; Thoms et al., 2020).

As summarized in **Table 1**, fl-domain boundaries of nsp1 were chosen to contain the first 180 amino acids, in analogy to its closest homolog from SCoV (Snijder et al., 2003). In addition, a shorter construct was designed, encoding only the globular core domain (GD, aa 13–127) suggested by the published SCoV nsp1 NMR structure (Almeida et al., 2007). His-tagged fl nsp1 was purified using the IPRS approach. Protein quality was confirmed by the available HSQC spectrum (**Figure 2**). Despite the flexible C-terminus, we were able to accomplish a near-complete backbone assignment (Wang et al., 2021).

Interestingly, the nsp1 GD was found to be problematic in our hands despite good expression. We observed insolubility, although buffers were used according to the homolog SCoV nsp1 GD (Almeida et al., 2007). Nevertheless, using a protocol comparable to the one for fl nsp1, we were able to record an HSQC spectrum proving a folded protein (**Figure 2**).

nsp2

nsp2 has been suggested to interact with host factors involved in intracellular signaling (Cornillez-Ty et al., 2009; Davies et al., 2020). The precise function, however, is insufficiently understood. Despite its potential dispensability for viral replication in general, it might be a valuable model to gain insights into virulence due to its possible involvement in the regulation of global RNA synthesis (Graham et al., 2005). We provide here a protocol for the purification of the C-terminal IDR (CtDR) of nsp2 from residues 557 to 601, based on disorder predictions [PrDOS (Ishida and Kinoshita, 2007)]. The His-Trx-tagged peptide was purified by IPRS. Upon dialysis, two IEC steps were performed: first anionic and then cationic, with good final yields (**Table 1**). Stability and

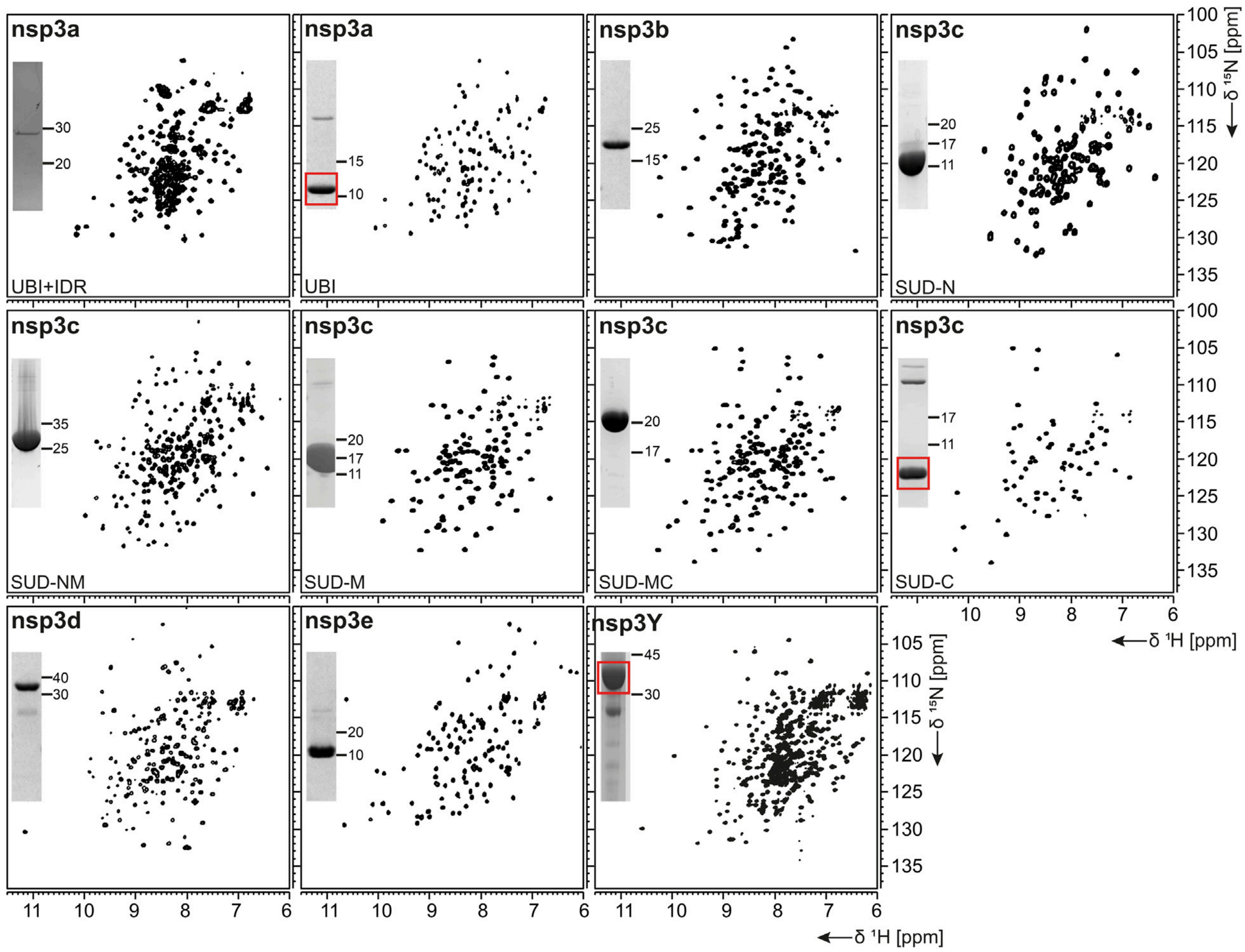


FIGURE 3 | ^{1}H, ^{15}N-correlation spectra of investigated constructs from nonstructural protein 3. Construct names of subdomains according to **Table 1** are indicated unless fl-domains are shown. A representative SDS-PAGE lane with final samples is included as inset. Red boxes indicate protein bands of interest.

purity were confirmed by an HSQC spectrum (**Figure 2**) and a complete backbone assignment (Mompean et al., 2020; **Table 2**).

nsp3

nsp3, the largest nsp (Snijder et al., 2003), is composed of a plethora of functionally related, yet independent, subunits. After cleavage of nsp3 from the fl ORF1-encoded polypeptide chain, it displays a 1945-residue multidomain protein, with individual functional entities that are subclassified from nsp3a to nsp3e followed by the ectodomain embedded in two TM regions and the very C-terminal CoV-Y domain. The soluble nsp3a-3e domains are linked by various types of linkers with crucial roles in the viral life cycle and are located in the so-called viral cytoplasm, which is separated from the host cell after budding off the endoplasmic reticulum and contains the viral RNA (Wolff et al., 2020). Remarkably, the nsp3c substructure comprises three subdomains, making nsp3 the most complex SCoV2 protein. The precise function and eventual RNA-binding specificities of nsp3 domains are not yet understood. We here focus on the nsp3 domains a–e and provide elaborated protocols for additional constructs carrying relevant linkers or combinations of domains (**Table 1**). Moreover, we additionally present a convenient protocol for the purification of the C-terminal CoV-Y domain.

nsp3a

The N-terminal portion of nsp3 is comprised of a ubiquitin-like (Ubl) structured domain and a subsequent acidic IDR. Besides its ability to bind ssRNA (Serrano et al., 2007), nsp3a has been reported to interact with the nucleocapsid (Hurst et al., 2013; Khan et al., 2020), playing a potential role in virus replication. We here provide protocols for the purification of both the Ubl (aa 1–111) and fl nsp3a (aa 1–206), including the acidic IDR (Ubl + IDR **Table 1**). Domain boundaries were defined similar to the published NMR structure of SCoV nsp3a (Serrano et al., 2007). His-tagged nsp3a Ubl + IDR and GST-tagged nsp3a Ubl were each purified via the IPRS approach. nsp3a Ubl yielded mM sample concentrations and displayed a well-dispersed HSQC spectrum (**Figure 3**). Notably, the herein described protocol also enables purification of fl nsp3a (Ubl + IDR) (**Tables 1**, **2**). Despite the unstructured IDR overhang, the excellent protein quality and stability allowed for near-complete backbone assignment [**Figure 3**, (Salvi et al., 2021)].

nsp3b

nsp3b is an ADP-ribose phosphatase macrodomain and potentially plays a key role in viral replication. Moreover, the de-ADP ribosylation function of nsp3b protects SCoV2 from antiviral host immune response, making nsp3b a promising drug target (Frick et al., 2020). As summarized in **Table 1**, the domain boundaries of the herein investigated nsp3b are residues 207–376 of the nsp3 primary sequence and were identical to available crystal structures with PDB entries 6YWM and 6YWL (unpublished). For purification, we used the IPRS approach, which yielded pure fl nsp3b (**Table 2**). Fl nsp3b displays well-dispersed HSQC spectra, making this protein an amenable target for NMR structural studies. In fact, we recently reported near-to-complete backbone assignments for nsp3b in its apo and ADP-ribose–bound form (Cantini et al., 2020).

nsp3c

The SARS unique domain (SUD) of nsp3c has been described as a distinguishing feature of SCoVs (Snijder et al., 2003). However, similar domains in more distant CoVs, such as MHV or MERS, have been reported recently (Chen et al., 2015; Kusov et al., 2015). nsp3c comprises three distinct globular domains, termed SUD-N, SUD-M, and SUD-C, according to their sequential arrangement: N-terminal (N), middle (M), and C-terminal (C). SUD-N and SUD-M develop a macrodomain fold similar to nsp3b and are described to bind G-quadruplexes (Tan et al., 2009), while SUD-C preferentially binds to purine-containing RNA (Johnson et al., 2010). Domain boundaries for SUD-N and SUD-M and for the tandem-domain SUD-NM were defined in analogy to the SCoV homolog crystal structure (Tan et al., 2009). Those for SUD-C and the tandem SUD-MC were based on NMR solution structures of corresponding SCoV homologs (**Table 1**) (Johnson et al., 2010). SUD-N, SUD-C, and SUD-NM were purified using GST affinity chromatography, whereas SUD-M and SUD-MC were purified using His affinity chromatography. Removal of the tag was achieved by thrombin cleavage and final samples of all domains were prepared subsequent to size-exclusion chromatography (SEC). Except for SUD-M, all constructs were highly stable (**Table 2**). Overall protein quality allowed for the assignment of backbone chemical shifts for the three single domains (Gallo et al., 2020) amd good resolved HSQC spectra also for the tandem domains (**Figure 3**).

nsp3d

nsp3d comprises the papain-like protease (PL^{pro}) domain of nsp3 and, hence, is one of the two SCoV2 proteases that are responsible for processing the viral polypeptide chain and generating functional proteins (Shin et al., 2020). The domain boundaries of PL^{pro} within nsp3 are set by residues 743 and 1,060 (**Table 1**). The protein is particularly challenging, as it is prone to misfolding and rapid precipitation. We prepared His-tagged and His-SUMO-tagged PL^{pro}. The His-tagged version mainly remained in the insoluble fraction. Still, mg quantities could be purified from the soluble fraction, however, greatly misfolded. Fusion to SUMO significantly enhanced protein yield of soluble PL^{pro}. The His-SUMO-tag allowed simple IMAC purification, followed by cleavage with Ulp1 and isolation of cleaved PL^{pro} via a second IMAC. A final purification step using gel filtration led to pure PL^{pro} of both unlabeled and 15N-labeled species (**Table 2**). The latter has allowed for the acquisition of a promising amide correlation spectrum (**Figure 3**).

nsp3e

nsp3e is unique to *Betacoronaviruses* and consists of a nucleic acid–binding domain (NAB) and the so-called group 2-specific marker (G2M) (Neuman et al., 2008). Structural information is rare; while the G2M is predicted to be intrinsically disordered (Lei et al., 2018); the only available experimental structure of the nsp3e NAB was solved from SCoV by the Wüthrich lab using solution NMR (Serrano et al., 2009). We here used this structure for a sequence-based alignment to derive reasonable domain boundaries for the SCoV2 nsp3e NAB (**Figures 4A,B**). The high sequence similarity suggested using nsp3 residues 1,088–1,203 (**Table 1**). This polypeptide chain was encoded in expression vectors comprising His- and His-GST tags, both cleavable by TEV protease. Both constructs showed excellent expression, suitable for the IPRS protocol (**Figure 4C**). Finally, a homogenous NAB species, as supported by the final gel of pooled samples (**Figure 4D**), was obtained. The excellent protein quality and stability are supported by the available HSQC (**Figure 3**) and a published backbone assignment (Korn et al., 2020a).

nsp3Y

nsp3Y is the most C-terminal domain of nsp3 and exists in all coronaviruses (Neuman et al., 2008; Neuman, 2016). Together, though, with its preceding regions G2M, TM 1, the ectodomain, TM2, and the Y1-domain, it has evaded structural investigations so far. The precise function of the CoV-Y domain remains unclear, but, together with the Y1-domain, it might affect binding to nsp4 (Hagemeijer et al., 2014). We were able to produce and purify nsp3Y (CoV-Y) comprising amino acids 1,638–1,945 (**Table 1**), yielding 12 mg/L with an optimized protocol that keeps the protein in a final NMR buffer containing HEPES and lithium bromide. Although the protein still shows some tendency to aggregate and degrade (**Table 2**), and despite its relatively large size, the spectral quality is excellent (**Figure 3**). nsp3 CoV-Y appears suitable for an NMR backbone

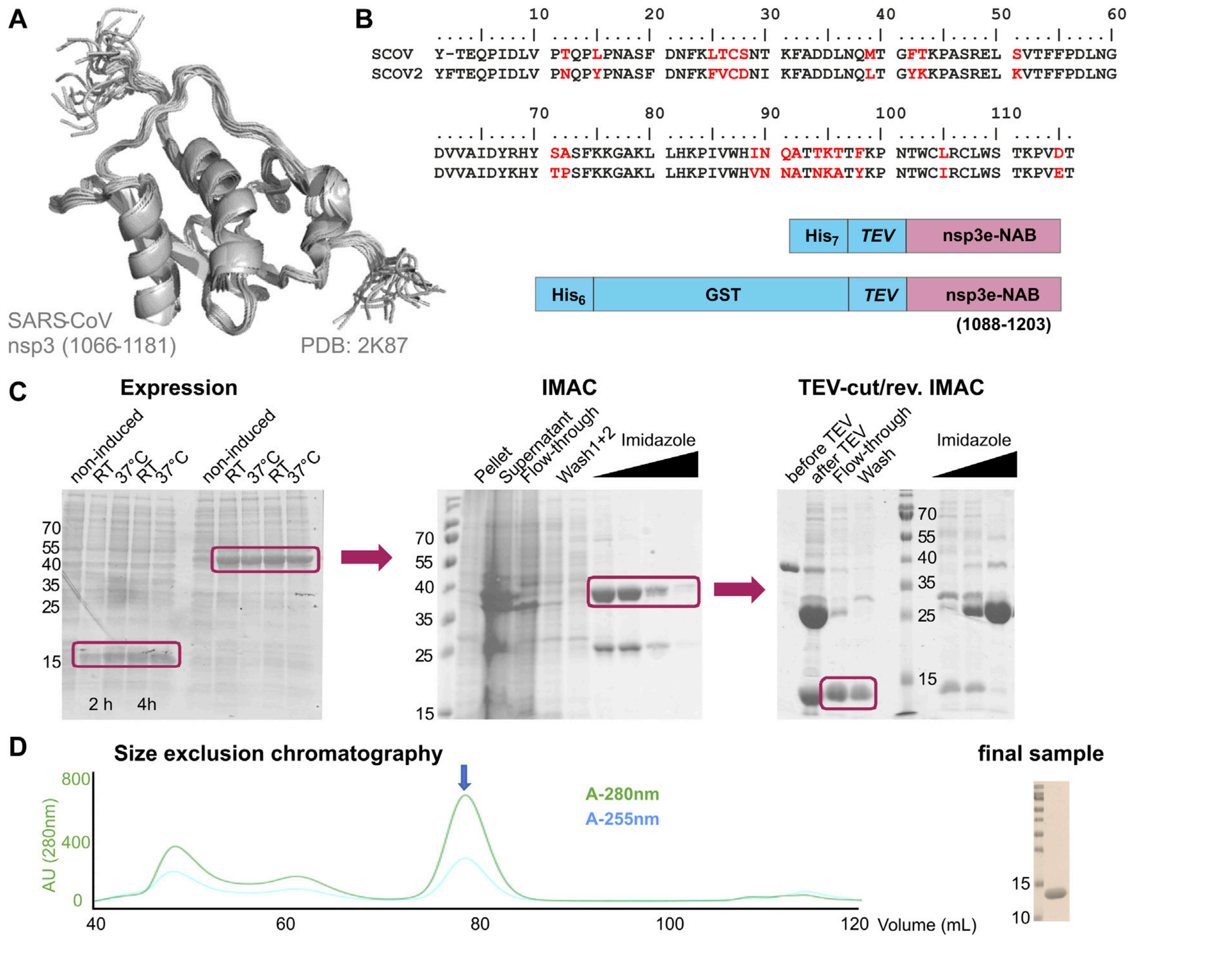


FIGURE 4 | Rationale of construct design, expression, and IPRS purification of the nsp3e nucleic acid–binding domain (NAB). **(A)** NMR structural ensemble of the homologous SCoV nsp3e (Serrano et al., 2009). The domain boundaries as displayed are given. **(B)** Sequence alignment of SCoV and SCoV2 regions representing the nsp3e locus. Arrows indicate the sequence stretch as used for the structure in panel **(A)**. The analogous region was used for the design of the two protein expression constructs shown **(C)**. Left, SDS-PAGE showing the expression of nsp3e constructs from panel **(B)** over 4 h at two different temperatures. Middle, SDS-PAGE showing the subsequent steps of IMAC. Right, SDS-PAGE showing steps and fractions obtained before and after TEV/dialysis and reverse IMAC. Boxes highlight the respective sample species of interest for further usage **(D)** SEC profile of nsp3e following steps in panel **(C)** performed with a Superdex 75 16/600 (GE Healthcare) column in the buffer as denoted in **Supplementary Table SI3**. The arrow indicates the protein peak of interest containing monomeric and homogenous nsp3e NAB devoid of significant contaminations of nucleic acids as revealed by the excellent 280/260 ratio. Right, SDS-PAGE shows 0.5 µL of the final NMR sample used for the spectrum in **Figure 3** after concentrating relevant SEC fractions.

assignment carried out at lower concentrations in a deuterated background (ongoing).

nsp5

The functional main protease nsp5 (M^{pro}) is a dimeric cysteine protease (Ullrich and Nitsche, 2020). Amino acid sequence and 3D structure of SCoV [PDB 1P9U (Anand et al., 2003)] and SCoV2 (PDB 6Y2E [Zhang et al., 2020)] homologs are highly conserved (**Figures 5A,B**). The dimer interface involves the N-termini of both monomers, which puts considerable constraints on the choice of protein sequence for construct design regarding the N-terminus.

We thus designed different constructs differing in the N-terminus: the native N-terminus (wt), a GS mutant with the additional N-terminal residues glycine and serine as His-SUMO fusion, and a GHM mutant with the amino acids glycine, histidine, and methionine located at the N-terminus with His-tag and TEV cleavage site (**Figure 5C**). Purification of all proteins via the IPRS approach (**Figures 5D,E**) yielded homogenous and highly pure protein, analyzed by PAGE (**Figure 5G**), mass spectrometry, and 2D [^{15}N, ^{1}H]-BEST TROSY spectra (**Figure 5H**). Final yields are summarized in **Table 2**.

nsp7 and nsp8

Both nsp7 and nsp8 are auxiliary factors of the polymerase complex together with the RNA-dependent RNA polymerase nsp12 and have high sequence homology with SCoV (100% and 99%, respectively) (Gordon et al., 2020). For nsp7 in complex with nsp8 or for nsp8 alone, additional functions in RNA synthesis priming have been proposed (Tvarogova et al., 2019;

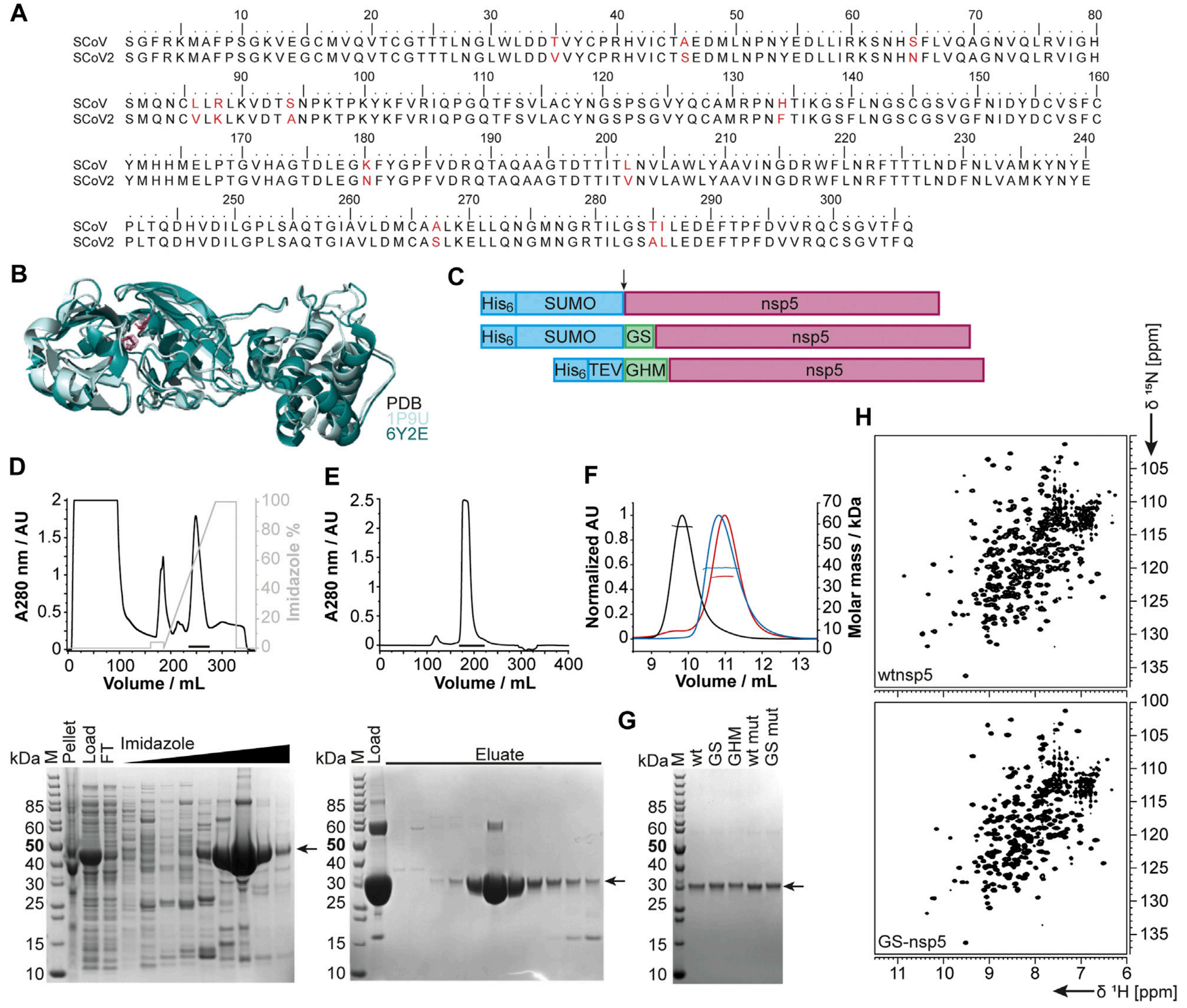


FIGURE 5 | Rationale of construct design, expression, and purification of different nsp5 constructs. **(A)** Sequence alignment of SCoV and SCoV2 fl nsp5. **(B)** X-ray structural overlay of the homologous SCoV (PDB 1P9U, light blue) and SCoV2 nsp5 (PDB 6Y2E, green) in cartoon representation. The catalytic dyad (H41 and C145) is shown in stick representation (magenta). **(C)** Schematics of nsp5 expression constructs involving purification and solubilization tags (blue), different N-termini and additional aa after cleavage (green), and nsp5 (magenta). Cleavage sites are indicated by an arrow. **(D, E)** An exemplary purification is shown for wtnsp5. IMAC **(D)** and SEC **(E)** chromatograms (upper panels) and the corresponding SDS PAGE (lower panels). Black bars in the chromatograms indicate pooled fractions. Gel samples are as follows: M: MW standard; pellet/load: pellet/supernatant after cell lysis; FT: IMAC flow-through; imidazole: eluted fractions with linear imidazole gradient; eluate: eluted SEC fractions from input (load). **(F)** SEC-MALS analysis with ~0.5 µg of wtnsp5 without additional aa (wtnsp5, black) with GS (GS-nsp5, blue) and with GHM (GHM-nsp5, red)) in NMR buffer on a Superdex 75, 10/300 GL (GE Healthcare) column. Horizontal lines indicate fractions of monodisperse nsp5 used for MW determination. **(G)** A SDS-PAGE showing all purified nsp5 constructs. The arrow indicates nsp5. **(H)** Exemplary [^{15}N, ^{1}H]-BEST-TROSY spectra measured at 298 K for the dimeric wtnsp5 (upper spectrum) and monomeric GS-nsp5 (lower spectrum). See **Supplementary Table SI4** for technical details regarding this figure.

Konkolova et al., 2020). In a recent study including an RNA-substrate-bound structure (Hillen et al., 2020), both proteins (with two molecules of nsp8 and one molecule of nsp7 for each nsp12 RNA polymerase) were found to be essential for polymerase activity in SCoV2. For both fl-proteins, a previously established expression and IPRS purification strategy for the SCoV proteins (Kirchdoerfer and Ward, 2019) was successfully transferred, which resulted in decent yields of reasonably stable proteins (**Table 2**). Driven by its intrinsically oligomeric state, nsp8 showed some tendency toward aggregation, limiting the available sample concentration. The higher apparent molecular weight and limited solubility are also reflected in the success of NMR experiments. While we succeeded in a complete NMR backbone assignment of nsp7 (Tonelli et al., 2020), the quality of the spectra obtained for nsp8 is currently limited to the HSQC presented in **Figure 2**.

nsp9

The 12.4 kDa ssRNA-binding nsp9 is highly conserved among *Betacoronaviruses*. It is a crucial part of the viral replication machinery (Miknis et al., 2009), possibly targeting the 3'-end stem-loop II (s2m) of the genome (Robertson et al., 2005). nsp9 adopts a fold similar to oligonucleotide/oligosaccharide-binding proteins (Egloff et al., 2004), and structural data consistently uncovered nsp9 to be dimeric in solution (Egloff et al., 2004; Sutton et al., 2004; Miknis et al., 2009; Littler et al., 2020). Dimer formation seems to be a prerequisite for viral replication (Miknis et al., 2009) and influences RNA-binding (Sutton et al., 2004), despite a moderate affinity for RNA *in vitro* (Littler et al., 2020).

Based on the early available crystal structure of SCoV2 nsp9 (PDB 6W4B, unpublished), we used the 113 aa fl sequence of nsp9 for our expression construct (**Table 1**). Production of either His- or His-GST-tagged fl nsp9 yielded high amounts of soluble protein in both natural abundance and ^{13}C- and ^{15}N-labeled form. Purification *via* the IPRS approach enabled us to separate fl nsp9 in different oligomer states. The earliest eluted fraction represented higher oligomers, was contaminated with nucleic acids and was not possible to concentrate above 2 mg/ml. This was different for the subsequently eluting dimeric fl nsp9 fraction, which had a A260/280 ratio of below 0.7 and could be concentrated to >5 mg/ml (**Table 2**). The excellent protein quality and stability are supported by the available HSQC (**Figure 2**), and a near-complete backbone assignment (Dudas et al., 2021).

nsp10

The last functional protein encoded by ORF1a, nsp10, is an auxiliary factor for both the methyltransferase/exonuclease nsp14 and the 2′-O-methyltransferase (MTase) nsp16. However, it is required for the MTase activity of nsp16 (Krafcikova et al., 2020), it confers exonuclease activity to nsp14 in the RNA polymerase complex in SCoV (Ma et al., 2015). It contains two unusual zinc finger motifs (Joseph et al., 2006) and was initially proposed to comprise RNA-binding properties. We generated a construct (**Table 1**) containing an expression and affinity purification tag on the N-terminus as reported for the SCoV variant (Joseph et al., 2006). Importantly, additional Zn^{2+} ions present during expression and purification stabilize the protein significantly (Kubatova et al., 2020). The yield during isotope-labeling was high (**Table 2**), and tests in unlabeled rich medium showed the potential for yields exceeding 100 mg/L. These characteristics facilitated in-depth NMR analysis and a backbone assignment (Kubatova et al., 2020).

nsp13

nsp13 is a conserved ATP-dependent helicase that has been characterized as part of the RNA synthesis machinery by binding to nsp12 (Chen et al., 2020b). It represents an interesting drug target, for which the available structure (PDB 6ZSL) serves as an excellent basis (**Table 1**). The precise molecular function, however, has remained enigmatic since it is not clear whether the RNA unwinding function is required for making ssRNA accessible for RNA synthesis (Jia et al., 2019) or whether it is required for proofreading and backtracking (Chen et al., 2020b). We obtained pure protein using a standard expression vector, generating a His-SUMO-tagged protein. Following Ulp1 cleavage, the protein showed limited protein stability in the solution (**Table 2**).

nsp14

nsp14 contains two domains: an N-terminal exonuclease domain and a C-terminal MTase domain (Ma et al., 2015). The exonuclease domain interacts with nsp10 and provides part of the proofreading function that supports the high fidelity of the RNA polymerase complex (Robson et al., 2020). Several unusual features, such as the unusual zinc finger motifs, set it apart from other DEDD-type exonucleases (Chen et al., 2007), which are related to both nsp10 binding and catalytic activity. The MTase domain modifies the N7 of the guanosine cap of genomic and subgenomic viral RNAs, which is essential for the translation of viral proteins (Thoms et al., 2020). The location of this enzymatic activity within the RNA synthesis machinery ensures that newly synthesized RNA is rapidly capped and thus stabilized. As a strategy, we used constructs, which allow coexpression of both nsp14 and nsp10 (pRSFDuet and pETDuet, respectively). Production of isolated fl nsp14 was successful, however, with limited yield and stability (**Table 2**). Expression of the isolated MTase domain resulted in soluble protein with 27.5 kDa mass that was amenable to NMR characterization (**Figure 2**), although only under reducing conditions and in the presence of high (0.4 M) salt concentration.

nsp15

The poly-U-specific endoribonuclease nsp15 was one of the very first SCoV2 structures deposited in the PDB [6VWW, (Kim et al., 2020)]. Its function has been suggested to be related to the removal of U-rich RNA elements, preventing recognition by the innate immune system (Deng et al., 2017), even though the precise mechanism remains to be established. The exact role of the three domains (N-terminal, middle, and C-terminal catalytic domain) also remains to be characterized in more detail (Kim et al., 2020). Here, the sufficient yield of fl nsp15 during expression supported purification of pure protein, which, however, showed limited stability in solution (**Table 2**).

nsp16

The MTase reaction catalyzed by nsp16 is dependent on nsp10 as a cofactor (Krafcikova et al., 2020). In this reaction, the 2'-OH group of nucleotide +1 in genomic and subgenomic viral RNA is methylated, preventing recognition by the innate immune system. Since both nsp14 and nsp16 are in principle susceptible to inhibition by MTase inhibitors, a drug targeting both enzymes would be highly desirable (Bouvet et al., 2010). nsp16 is the last protein being encoded by ORF1ab, and only its N-terminus is formed by cleavage by the M^{pro} nsp5. Employing a similar strategy to that for nsp14, nsp16 constructs were designed with the possibility of nsp10 coexpression. Expression of fl nsp16 resulted in good yields, when expressed both isolated and together with nsp10. The protein, however, is in either case

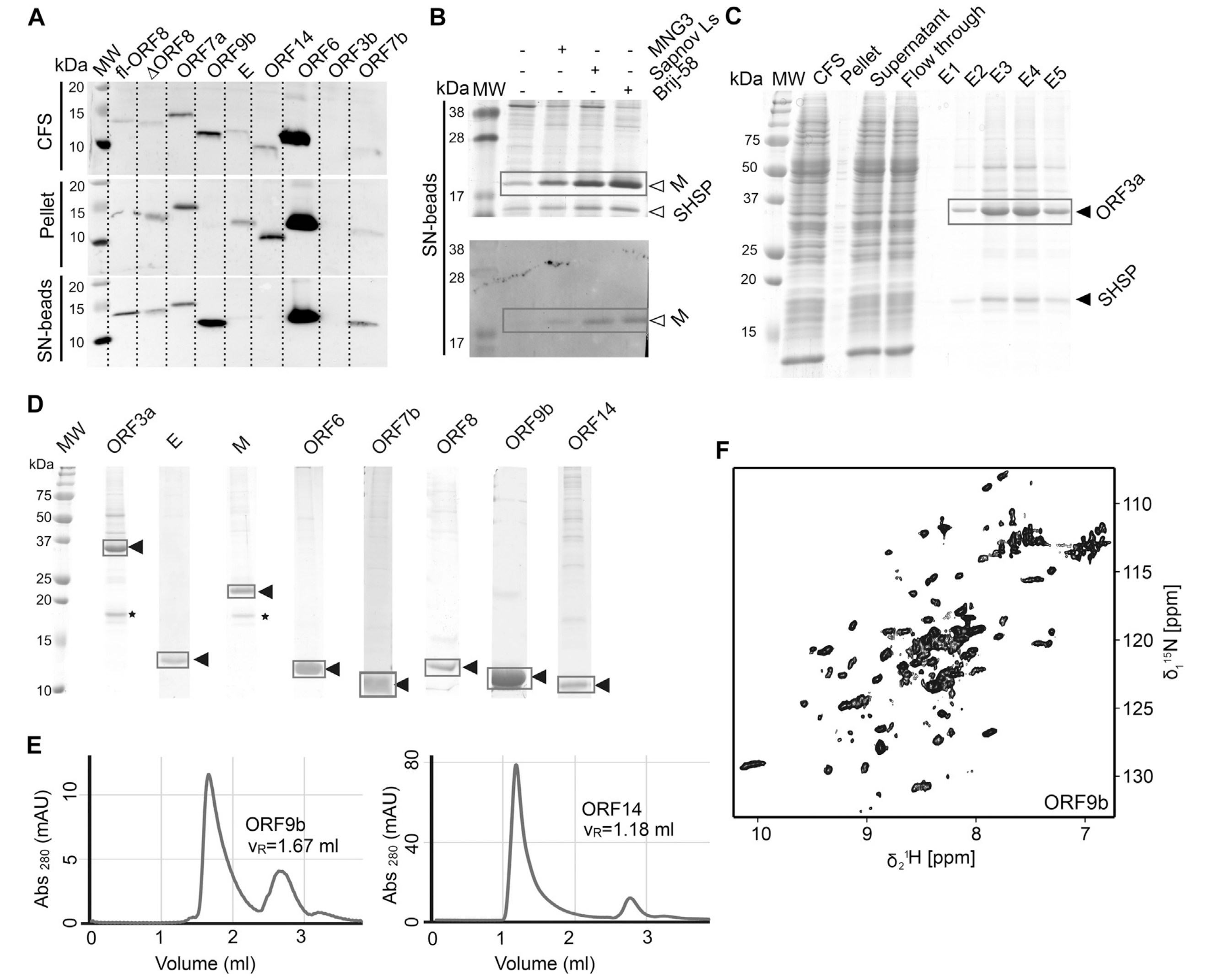


FIGURE 6 | Cell-free protein synthesis of accessory ORFs and structural proteins E and M. **(A)** Screening for expression and solubility of different ORFs using small-scale reactions. The total cell-free reaction (CFS), the pellet after centrifugation, and the supernatant (SN) captured on magnetic beads coated with Strep-Tactin were analyzed. All tested proteins were synthesized, with the exception of ORF3b. MW, MW standard. **(B)** Detergent solubilization tests using three different detergents, here at the example of the M protein, shown by SDS-PAGE and Western Blot. **(C)** Proteins are purified in a single step using a Strep-Tactin column. For ORF3a (and also for M), a small heat-shock protein of the HSP20 family is copurified, as identified by mass spectrometry (see also * in Panel **D**). **(D)** SDS-PAGE of the ^{2}H, ^{13}C, ^{15}N-labeled proteins used as NMR samples. Yields were between 0.2 and 1 mg protein per mL wheat-germ extract used. **(E)** SEC profiles for two ORFs. Left, ORF9b migrates as expected for a dimer. Right, OFR14 shows large assemblies corresponding to approximately 9 protein units and the DDM detergent micelle. **(F)** 2D [^{15}N, ^{1}H]-BEST-TROSY spectrum of ORF9b, recorded at 900 MHz in 1 h at 298 K, on less than 1 mg of protein. See **Supplementary Tables SI13–SI19** and **Supplementary Tables SI19, SI20** for technical and experimental details regarding this figure.

unstable in solution and highly dependent on reducing buffer conditions (**Table 2**). The purification procedures of nsp16 were adapted with minor modifications from a previous X-ray crystallography study (Rosas-Lemus et al., 2020).

Structural Proteins and Accessory ORFs

Besides establishing expression and purification protocols for the nsps, we also developed protocols and obtained pure mg quantities of the SCoV2 structural proteins E, M, and N, as well as literally all accessory proteins. With the exception of the relatively well-behaved nucleocapsid (N) protein, SCoV2 E, M, and the remaining accessory proteins represent a class of mostly small and relatively poorly characterized proteins, mainly due to their difficult behavior in classical expression systems.

We used wheat-germ cell-free protein synthesis (WG-CFPS) for the successful production, solubilization, purification, and, in part, initial NMR spectroscopic investigation of ORF3a, ORF6, ORF7b, ORF8, ORF9b, and ORF14 accessory proteins, as well as E and M in mg quantities using the highly efficient translation machinery extracted from wheat-germs (**Figures 6A–D**).

ORF3a

The protein from ORF3a in SCoV2 corresponds to the accessory protein 3a in SCoV, with homology of more than 70% (**Table 1**). It has 275 amino acids, and its structure has recently been determined (Kern et al., 2020). The structure of SCoV2 3a displays a dimer, but it can also form higher oligomers. Each monomer has three TM helices and a cytosolic β-strand rich domain. SCoV2 ORF3a is a cation channel, and its structure has been solved by electron microscopy in nanodiscs. In SCoV, 3a is a structural component and was found in recombinant virus-like particles (Liu et al., 2014), but is not explicitly needed for their formation. The major challenge for NMR studies of this largest accessory protein is its size, independent of its employment in solid state or solution NMR spectroscopy.

As most other accessory proteins described in the following, ORF3a has been produced using WG-CFPS and was expressed in soluble form in the presence of Brij-58 (**Figure 6C**). It is copurified with a small heat-shock protein of the HSP20 family from the wheat-germ extract. The protocol described here is highly similar to that of the other cell-free synthesized accessory proteins. Where NMR spectra have been reported, the protein has been produced in a ^{2}H, ^{13}C, ^{15}N uniformly labeled form; otherwise, natural abundance amino acids were added to the reaction. The proteins were further affinity-purified in one step using Strep-Tactin resin, through the Strep-tag II fused to their N- or C-terminus. For membrane proteins, protein synthesis and also purification were done in the presence of detergent.

About half a milligram of pure protein was generally obtained per mL of extract, and up to 3 ml wheat-germ extract have been used to prepare NMR samples.

ORF3b

The ORF3b protein is a putative protein stemming from a short ORF (57 aa) with no homology to existing SCoV proteins (Chan et al., 2020). Indeed, ORF3b gene products of SCoV2 and SCoV are considerably different, with one of the distinguishing features being the presence of premature stop codons, resulting in the expression of a drastically shortened ORF3b protein (Konno et al., 2020). However, the SCoV2 nucleotide sequence after the stop codon shows a high similarity to the SCoV ORF3b. Different C-terminal truncations seem to play a role in the interferon-antagonistic activity of ORF3b (Konno et al., 2020). ORF3b is the only protein that, using WG-CFPS, was not synthesized at all; i.e., it was neither observed in the total cell-free reaction nor in supernatant or pellet. This might be due to the premature stop codon, which was not considered. Constructs of ORF3b thus need to be redesigned.

ORF4 (Envelope Protein, E)

The SCoV2 envelope (E) protein is a small (75 amino acids), integral membrane protein involved in several aspects of the virus' life cycle, such as assembly, budding, envelope formation, and pathogenicity, as recently reviewed in (Schoeman and Fielding, 2020). Structural models for SCoV (Surya et al., 2018) and the TM helix of SCoV2 (Mandala et al., 2020) E have been established. The structural models show a pentamer with a TM helix. The C-terminal part is polar, with charged residues interleaved, and is positioned on the membrane surface in SCoV. E was produced in a similar manner to ORF3a, using the addition of detergent to the cell-free reaction.

ORF5 (Membrane Glycoprotein, M)

The M protein is the most abundant protein in the viral envelope and is believed to be responsible for maintaining the virion in its characteristic shape (Huang et al., 2004). M is a glycoprotein and sequence analyses predict three domains: A C-terminal endodomain, a TM domain with three predicted helices, and a short N-terminal ectodomain. M is essential for viral particle assembly. Intermolecular interactions with the other structural proteins, N and S to a lesser extent, but most importantly E (Vennema et al., 1996), seem to be central for virion envelope formation in coronaviruses, as M alone is not sufficient. Evidence has been presented that M could adopt two conformations, elongated and compact, and that the two forms fulfill different functions (Neuman et al., 2011). The lack of more detailed structural information is in part due to its small size, close association with the viral envelope, and a tendency to form insoluble aggregates when perturbed (Neuman et al., 2011). The M protein is readily produced using cell-free synthesis in the presence of detergent; as ORF3a, it is copurified with a small heat-shock protein of the HSP20 family (**Figure 6B**). Membrane-reconstitution will likely be necessary to study this protein.

ORF6

The ORF6 protein is incorporated into viral particles and is also released from cells (Huang et al., 2004). It is a small protein (61 aa), which has been found to concentrate at the endoplasmic reticulum and Golgi apparatus. In a murine coronavirus model, it was shown that expressing ORF6 increased virulence in mice (Zhao et al., 2009), and results indicate that ORF6 may serve an important role in the pathogenesis during SCoV infection (Liu et al., 2014). Also, it showed to inhibit the expression of certain STAT1-genes critical for the host immune response and could contribute to the immune evasion. ORF6 is expressed very well in WG-CFPS; the protein was fully soluble with detergents and partially soluble without them and was easily purified in the presence of detergent, but less efficiently in the absence thereof. Solution NMR spectra in the presence of detergent display narrow but few resonances, which correspond, in addition to the C-terminal STREP-tag, to the very C-terminal ORF6 protein residues.

ORF7a

SCoV2 protein 7a (121 aa) shows over 85% homology with the SCoV protein 7a. While the SCoV2 7a protein is produced and retained intracellularly, SCoV protein 7a has also been shown to be a structural protein incorporated into mature virions (Liu et al., 2014). 7a is one of the accessory proteins, of which a (partial) structure has been determined at high

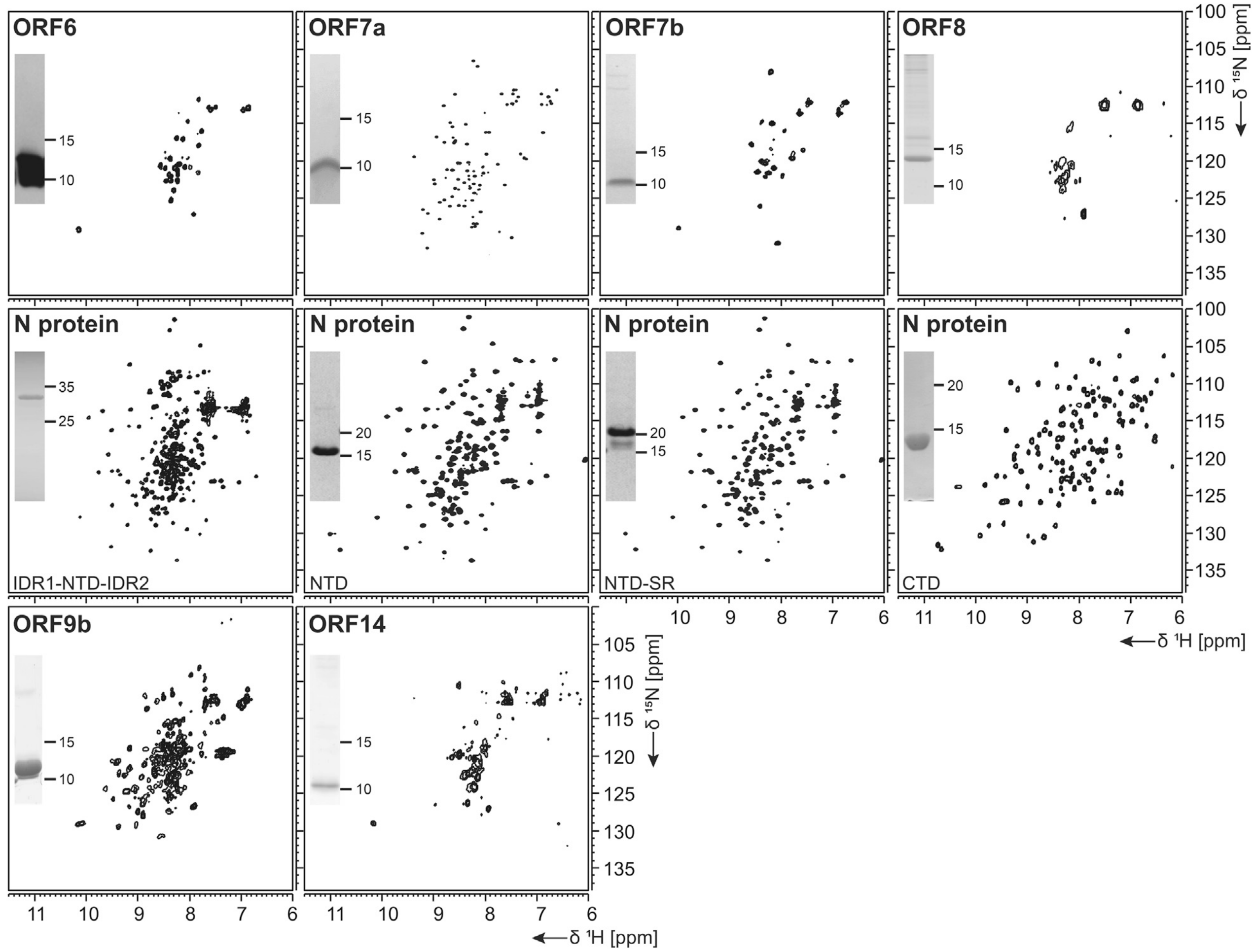


FIGURE 7 | ^{1}H, ^{15}N-correlation spectra of investigated structural and accessory proteins. Construct names according to **Table 1** are indicated unless fl-proteins are shown. A representative SDS-PAGE lane with final samples is included as inset.

resolution for SCoV2 (PDB 6W37). However, the very N-terminal signal peptide and the C-terminal membrane anchor, both highly hydrophobic, have not been determined experimentally yet.

Expression of the ORF7a ectodomain (ED) with a GB1 tag (Bogomolovas et al., 2009) was expected to produce reasonable yields. The IPRS purification resulted in a highly stable protein, as evidenced by the NMR data obtained (**Figure 7**).

ORF7b

Protein ORF7b is associated with viral particles in a SARS context (Liu et al., 2014). Protein 7b is one of the shortest ORFs with 43 residues. It shows a long hydrophobic stretch, which might correspond to a TM segment. It shows over 93% sequence homology with a bat coronavirus 7b protein (Liu et al., 2014). There, the cysteine residue in the C-terminal part is not conserved, which might facilitate structural studies. ORF7b has been synthesized successfully both from bacteria and by WG-CFPS in the presence of detergent and could be purified using a STREP-tag (**Table 2**). Due to the necessity of solubilizing agent and its obvious tendency to oligomerize, structure determination, fragment screening, and interaction studies are challenging. However, we were able to record the first promising HSQC, as shown in **Figure 7**.

ORF8

ORF 8 is believed to be responsible for the evolution of *Betacoronaviruses* and their species jumps (Wu et al., 2016) and to have a role in repressing the host response (Tan et al., 2020). ORF 8 (121 aa) from SCoV2 does not apparently exist in SCoV on the protein level, despite the existence of a putative ORF. The sequences of the two homologs only show limited identity, with the exception of a small 7 aa segment, where, in SCoV, the glutamate is replaced with an aspartate. It, however, aligns very well with several coronaviruses endemic to animals, including Paguma and Bat (Chan et al., 2020). The protein comprises a hydrophobic peptide at its very N-terminus, likely corresponding to a signal peptide; the remaining part does not show any specific sequence features. Its structure has been determined (PDB 7JTL) and shows a similar fold to ORF7a (Flower et al., 2020). In this study, ORF8 has been used both with (fl) and without signal peptide (ΔORF8). We first tested the production of ORF8 in *E. coli*, but yields were low because of insolubility. Both ORF8 versions have then been synthesized in the cell-free system and were soluble in the presence of detergent. Solution NMR spectra, however, indicate that the protein is forming either oligomers or aggregates.

ORF9a (Nucleocapsid Protein, N)

The nucleocapsid protein (N) is important for viral genome packaging (Luo et al., 2006). The multifunctional RNA-binding protein plays a crucial role in the viral life cycle (Chang et al., 2014) and its domain architecture is highly conserved among coronaviruses. It comprises the N-terminal intrinsically disordered region (IDR1), the N-terminal RNA-binding globular domain (NTD), a central serine/arginine-(SR-) rich intrinsically disordered linker region (IDR2), the C-terminal dimerization domain (CTD), and a C-terminal intrinsically disordered region (IDR3) (Kang et al., 2020).

N represents a highly promising drug target. We thus focused our efforts not exclusively on the NTD and CTD alone, but, in addition, also provide protocols for IDR-containing constructs within the N-terminal part.

N-Terminal Domain

The NTD is the RNA-binding domain of the nucleocapsid (Kang et al., 2020). It is embedded within IDRs, functions of which have not yet been deciphered. Recent experimental and bioinformatic data indicate involvement in liquid-liquid phase separation (Chen et al., 2020a).

For the NTD, several constructs were designed, also considering the flanking IDRs (**Table 1**). In analogy to the available NMR [PDB 6YI3, (Dinesh et al., 2020)] and crystal [PDB 6M3M, (Kang et al., 2020)] structures of the SCoV2 NTD, boundaries for the NTD and the NTD-SR domains were designed to span residues 44–180 and 44–212, respectively. In addition, an extended IDR1-NTD-IDR2 (residues 1–248) construct was designed, including the N-terminal disordered region (IDR1), the NTD domain, and the central disordered linker (IDR2) that comprises the SR region. His-tagged NTD and NTD-SR were purified using IPRS and yielded approx. 3 mg/L in ^{15}N-labeled minimal medium. High protein quality and stability are supported by the available HSQC spectra (**Figure 7**).

The untagged IDR1-NTD-IDR2 was purified by IEC and yielded high amounts of ^{13}C, ^{15}N-labeled samples of 12 mg/L for further NMR investigations. The quality of our purification is confirmed by the available HSQC (**Figure 7**), and a near-complete backbone assignment of the two IDRs was achieved (Guseva et al., 2021; Schiavina et al., 2021). Notably, despite the structurally and dynamically heterogeneous nature of the N protein, the mentioned N constructs revealed a very good long-term stability, as shown in **Table 2**.

C-Terminal Domain

Multiple studies on the SCoV2 CTD, including recent crystal structures (Ye et al., 2020; Zhou et al., 2020), confirm the domain as dimeric. Its ability to self-associate seems to be necessary for viral replication and transcription (Luo et al., 2006). In addition, the CTD was shown to, presumably nonspecifically, bind ssRNA (Zhou et al., 2020).

Domain boundaries for the CTD were defined to comprise amino acids 247–364 (**Table 1**), in analogy to the NMR structure of the CTD from SCoV (PDB 2JW8, [Takeda et al., 2008)]. Gene expression of His- or His-GST-tagged CTD yielded high amounts of soluble protein. Purification was achieved via IPRS. The CTD eluted as a dimer judged by its retention volume on the size-exclusion column and yielded good amounts (**Table 2**). The excellent protein quality and stability are supported by the available HSQC spectrum (**Figure 7**) and a near-complete backbone assignment (Korn et al., 2020b).

ORF9b

Protein 9b (97 aa) shows 73% sequence homology to the SCoV and also to bat virus (bat-SL-CoVZXC21) 9b protein (Chan et al., 2020). The structure of SCoV2 ORF9b has been determined at high resolution (PDB 6Z4U). Still, a significant portion of the structure was not found to be well ordered. The protein shows a β-sheet-rich structure and a hydrophobic tunnel, in which bound lipid was identified. How this might relate to membrane binding is not fully understood at this point. The differences in sequence between SCoV and SCoV2 are mainly located in the very N-terminus, which was not resolved in the structure (PDB 6Z4U). Another spot of deviating sequence not resolved in the structure is a solvent-exposed loop, which presents a potential interacting segment. ORF9b has been synthesized as a dimer (**Figure 6E**) using WG-CFPS in its soluble form. Spectra show a well-folded protein, and assignments are underway (**Figure 6F**).

ORF14 (ORF9c)

ORF14 (73 aa) remains, at this point in time, hypothetical. It shows 89% homology with a bat virus protein (bat-SL-CoVZXC21). It shows a highly hydrophobic part in its C-terminal region, comprising two negatively charged residues and a charged/polar N-terminus. The C-terminus is likely mediating membrane interaction. While ORF14 has been synthesized in the wheat-germ cell-free system in the presence of detergent and solution NMR spectra have been recorded, they hint at an aggregated protein (**Figure 6E**). Membrane-reconstitution of ORF14 revealed an unstable protein, which had been degraded during detergent removal.

ORF10

The ORF10 protein is comprised of 38 aa and is a hypothetical protein with unknown function (Yoshimoto, 2020). SCoV2 ORF10 displays 52.4% homology to SCoV ORF9b. The protein sequence is rich in hydrophobic residues, rendering expression and purification challenging. Expression of ORF10 as His-Trx-tagged or His-SUMO tagged fusion protein was possible; however, the ORF10 protein is poorly soluble and shows partial unfolding, even as an uncleaved fusion protein. Analytical SEC hints at oligomerization under the current conditions.

DISCUSSION

The ongoing SCoV2 pandemic and its manifestation as the COVID-19 disease call for an urgent provision of therapeutics that will specifically target viral proteins and their interactions with each other and RNAs, which are crucial for viral propagation. Two "classical" viral targets have been addressed in comprehensive approaches soon after the outbreak in December 2019: the viral protease nsp5 and the RNA-dependent RNA polymerase (RdRp) nsp12. While the latter turned out to be a suitable target using the repurposed compound Remdesivir (Hillen et al., 2020), nsp5 is undergoing a broad structure-based screen against a battery of inhibitors in multiple places (Jin et al., 2020; Zhang et al., 2020), but with, as of yet, the limited outcome for effective medication. Hence, a comprehensive, reliable treatment of COVID-19 at any stage after the infection has remained unsuccessful.

Further viral protein targets will have to be taken into account in order to provide inhibitors with increased specificity and efficacy and preparative starting points for following potential generations of (SARS-)CoVs. Availability of those proteins in a recombinant, pure, homogenous, and stable form in milligrams is, therefore, a prerequisite for follow-up applications like vaccination, high-throughput screening campaigns, structure determination, and mapping of viral protein interaction networks. We here present, for the first time, a near-complete compendium of SCoV2 protein purification protocols that enable the production of large amounts of pure proteins.

The *COVID19-NMR* consortium was launched with the motivation of providing NMR assignments of all SCoV2 proteins and RNA elements, and enormous progress has been made since the outbreak of COVID-19 for both components [see **Table 2** and (Wacker et al., 2020)]. Consequently, we have put our focus on producing proteins in stable isotope-labeled forms for NMR-based applications, e.g., the site-resolved mapping of interactions with compounds (Li and Kang, 2020). Relevant to a broad scientific community, we here report our protocols to suite perfectly any downstream biochemical or biomedical application.

Overall Success and Protein Coverage

As summarized in **Table 2**, we have successfully purified 80% of the SCoV2 proteins either in fl or providing relevant fragments of the parent protein. Those include most of the nsps, where all of the known/predicted soluble domains have been addressed (**Figure 1**). For a very large part, we were able to obtain protein samples of high purity, homogeneity, and fold for NMR-based applications. We would like to point out a number of CoV proteins that, evidenced by their HSQCs, for the first time, provide access to structural information, e.g., the PL^{pro} nsp3d and nsp3Y. Particularly for the nsp3 multidomain protein, we here present soluble samples of almost the complete cytosolic region with more than 120 kDa in the form of excellent 2D NMR spectra (**Figure 3**), a major part of which fully backbone-assigned. We thus enable the exploitation of the largest and most enigmatic multifunctional SCoV2 protein through individual domains in solution, allowing us to study their concerted behavior with single residue resolution. Similarly, for nsp2, we provide a promising starting point for studying the so far neglected, often uncharacterized, and apparently unstructured proteins.

Driven by the fast-spreading COVID-19, we initially left out proteins that require advanced purification procedures (e.g., nsp12 and S) or where *a priori* information was limited (nsp4 and nsp6). This procedure seems justified with the time-saving approach of our effort in favor of the less attended proteins. However, we are in the process of collecting protocols for the missing proteins.

Different Complexities and Challenges

The compilation of protein production protocols, initially guided by information from CoV homologs (**Table 1**), has confronted us with very different levels of complexity. With some prior expectation toward this, we have shared forces to quickly "work off" the highly conserved soluble and small proteins and soon put focus into the processing of the challenging ones. The difficulties in studying this second class of proteins are due to their limited sequence conservation, no prior information, large molecular weights, insolubility, and so forth.

The nsp3e NAB represents one example where the available NMR structure of the SCoV homolog provided a *bona fide* template for selecting initial domain boundaries (**Figure 4**). The transfer of information derived from SCoV was straightforward; the transferability included the available protocol for the production of comparable protein amounts and quality, given the high sequence identity. In such cases, we found ourselves merely to adapt protocols and optimize yields based on slightly different expression vectors and *E. coli* strains.

However, in some cases, such transfer was unexpectedly not successful, e.g., for the short nsp1 GD. Despite intuitive domain boundaries with complete local sequence identity seen from the SCoV nsp1 NMR structure, it took considerable efforts to purify an analogous nsp1 construct, which is likely related to the impaired stability and solubility caused by a number of impacting amino acid exchanges within the domain's flexible loops. In line with that, currently available structures of SCoV2 nsp1 have been obtained by crystallography or cryo-EM and include different buffers. As such, our initial design was insufficient in terms of taking into account the parameters mentioned above. However, one needs to consider those particular differences between the nsp1 homologs as one of the most promising target sites for potential drugs as they appear to be hotspots in the CoV evolution and will have essential effects for the molecular networks, both in the virus and with the host (Zust et al., 2007; Narayanan et al., 2015; Shen et al., 2019; Thoms et al., 2020).

A special focus was put on the production of the SCoV2 main protease nsp5, for which NMR-based screenings are ongoing. The main protease is critical in terms of inhibitor design as it appears under constant selection, and novel mutants remarkably influence the structure and biochemistry of the protein (Cross et al., 2020). In the present study, the expression of the different constructs allowed us to characterize the protein in both its monomeric and dimeric forms. Comparison of NMR spectra reveals that the constructs with additional amino acids (GS and GHM mutant) display marked structural differences to the wild-type protein while being structurally similar among themselves (**Figure 5H**). The addition of two residues (GS) interferes with the dimerization interface, despite being similar to its native N-terminal amino acids (SGFR). We also introduced an active site mutation that replaces cysteine 145 with alanine (Hsu et al., 2005). Intriguingly, this active site mutation C145A, known to stabilize the dimerization of the main protease (Chang et al., 2007), supports dimer formation of the GS added construct (GS-nsp5 C145A) shown by its 2D NMR spectrum overlaying with the one of wild-type nsp5 (**Supplementary Table SI4**). The NMR results are in line with SEC-MALS analyses (**Figure 5F**). Indeed, the additional amino acids at the N-terminus shift the dimerization equilibrium toward the monomer, whereas the mutation shifts it toward the dimer despite the N-terminal aa additions. This example underlines the need for a thorough and precise construct design and the detailed biochemical and NMR-based characterization of the final sample state. The presence of monomers vs. dimers will play an essential role in the inhibitor search against SCoV2 proteins, as exemplified by the particularly attractive nsp5 main protease target.

Exploiting Nonbacterial Expression

As a particular effort within this consortium, we included the so far neglected accessory proteins using a structural genomics procedure supported by wheat-germ cell-free protein synthesis. This approach allowed us previously to express a variety of difficult viral proteins in our hands (Fogeron et al., 2015a; Fogeron et al., 2015b; Fogeron et al., 2016; Fogeron et al., 2017; Wang et al., 2019; Jirasko et al., 2020a). Within the workflow, we especially highlight the straightforward solubilization of the membrane proteins through the addition of detergent to the cell-free reaction, which allowed the production of soluble protein in milligram amounts compatible with NMR studies. While home-made extracts were used here, very similar extracts are available commercially (Cell-Free Sciences, Japan) and can thus be implemented by any lab without prior experience. Also, a major benefit of the WG-CFPS system for NMR studies lies in the high efficiency and selectivity of isotopic labeling. In contrast to cell-based expression systems, only the protein of interest is produced (Morita et al., 2003), which allows bypassing extensive purification steps. In fact, one-step affinity purification is in most cases sufficient, as shown for the different ORFs in this study. Samples could be produced for virtually all proteins, with the exception of the ORF3b construct used. With new recent insight into the stop codons present in this ORF, constructs will be adapted, which shall overcome the problems of ORF3b production (Konno et al., 2020).

For two ORFs, 7b and 8, we exploited a paralleled production strategy, i.e., both in bacteria and via cell-free synthesis. For those challenging proteins, we were, in principle, able to obtain pure samples from either expression system. However, for ORF7b, we found a strict dependency on detergents for follow-up work from both approaches. ORF8 showed significantly better solubility when produced in WG extracts compared to bacteria. This shows the necessity of parallel routes to take, in particular, for the understudied, biochemically nontrivial ORFs that might represent yet unexplored but highly specific targets to consider in the treatment of COVID-19.

Downstream structural analysis of ORFs produced with CFPS remains challenging but promising progress is being made in the light of SCoV2. Some solution NMR spectra show the expected number of signals with good resolution (e.g., ORF9b). As expected, however, most proteins cannot be straightforwardly analyzed by solution NMR in their current form, as they exhibit too large objects after insertion into micelles and/or by inherent oligomerization. Cell-free synthesized proteins can be inserted into membranes through reconstitution (Fogeron et al., 2015a; Fogeron et al., 2015b; Fogeron et al., 2016; Jirasko et al., 2020a; Jirasko et al., 2020b). Reconstitution will thus be the next step for many accessory proteins, but also for M and E, which were well produced by WG-CFPS. We will also exploit the straightforward deuteration in WG-CFPS (David et al., 2018; Wang et al., 2019; Jirasko et al., 2020a) that circumvents proton back-exchange, rendering denaturation and refolding steps obsolete (Tonelli et al., 2011). Nevertheless, the herein presented protocols for the production of non-nsps by WG-CFPS instantly enable their employment in binding studies and screening campaigns and thus provide a significant contribution to soon-to-come studies on SCoV2 proteins beyond the classical and convenient drug targets.

Altogether and judged by the ultimate need of exploiting recombinant SCoV2 proteins in vaccination and highly paralleled screening campaigns, we optimized sample amount, homogeneity, and long-term stability of samples. Our freely accessible protocols and accompanying NMR spectra now offer a great resource to be exploited for the unambiguous and reproducible production of SCoV2 proteins for the intended applications.

DATA AVAILABILITY STATEMENT

Assignments of backbone chemical shifts have been deposited at BMRB for proteins, as shown in **Table 2**, indicated by their respective BMRB IDs. All expression constructs are available as plasmids from https://covid19-nmr.de/.

AUTHOR CONTRIBUTIONS

NA, SK, NQ, MD, MN, ABö, HS, MH, and AS designed the study, compiled the protocols and NMR data, and wrote the manuscript. All authors contributed coordinative or practical work to the study. All authors contributed to the creation and collection of protein protocols and NMR spectra.

FUNDING

This work was supported by Goethe University (Corona funds), the DFG-funded CRC: "Molecular Principles of RNA-Based Regulation," DFG infrastructure funds (project numbers: 277478796, 277479031, 392682309, 452632086, 70653611), the state of Hesse (BMRZ), the Fondazione CR Firenze (CERM), and the IWB-EFRE-program 20007375. This project has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 871037. AS is supported by DFG Grant SCHL 2062/2-1 and by the JQYA at Goethe through project number 2019/AS01. Work in the lab of KV was supported by a CoRE grant from the University of New Hampshire. The FLI is a member of the Leibniz Association (WGL) and financially supported by the Federal Government of Germany and the State of Thuringia. Work in the lab of RM was supported by NIH (2R01EY021514) and NSF (DMR-2002837). BN-B was supported by the NSF GRFP. MC was supported by NIH (R25 GM055246 MBRS IMSD), and MS-P was supported by the HHMI Gilliam Fellowship. Work in the labs of KJ and KT was supported by Latvian Council of Science Grant No. VPP-COVID 2020/1-0014. Work in the UPAT's lab was supported by the INSPIRED (MIS 5002550) project, which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure," funded by the Operational Program "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014–2020) and cofinanced by Greece and the EU (European Regional Development Fund) and the FP7 REGPOT CT-2011-285950–"SEE-DRUG" project (purchase of UPAT's 700 MHz NMR equipment). Work in the CM-G lab was supported by the Helmholtz society. Work in the lab of ABö was supported by the CNRS, the French National Research Agency (ANR, NMR-SCoV2- ORF8), the Fondation de la Recherche Médicale (FRM, NMR-SCoV2-ORF8), and the IR-RMN-THC Fr3050 CNRS. Work in the lab of BM was supported by the Swiss National Science Foundation (Grant number 200020_188711), the Günthard Stiftung für Physikalische Chemie, and the ETH Zurich. Work in the labs of ABö and BM was supported by a common grant from SNF (grant 31CA30_196256). This work was supported by the ETH Zurich, the grant ETH 40 18 1, and the grant Krebsliga KFS 4903 08 2019. Work in the lab of the IBS Grenoble was supported by the Agence Nationale de Recherche (France) RA-COVID SARS2NUCLEOPROTEIN and European Research Council Advanced Grant DynamicAssemblies. Work in the CA lab was supported by Patto per il Sud della Regione Siciliana–CheMISt grant (CUP G77B17000110001). Part of this work used the platforms of the Grenoble Instruct-ERIC center (ISBG; UMS 3518 CNRS-CEA-UGA-EMBL) within the Grenoble Partnership for Structural Biology (PSB), supported by FRISBI (ANR-10-INBS-05-02) and GRAL, financed within the University Grenoble Alpes graduate school (Ecoles Universitaires de Recherche) CBH-EUR-GS (ANR-17-EURE-0003). Work at the UW-Madison was supported by grant numbers NSF MCB2031269 and NIH/NIAID AI123498. MM is a Ramón y Cajal Fellow of the Spanish AEI-Ministry of Science and Innovation (RYC2019-026574-I), and a "La Caixa" Foundation (ID 100010434) Junior Leader Fellow (LCR/BQ/PR19/11700003). Funded by project COV20/00764 from the Carlos III Institute of Health and the Spanish Ministry of Science and Innovation to MM and DVL. VDJ was supported by the Boehringer Ingelheim Fonds. Part of this work used the resources of the Italian Center of Instruct-ERIC at the CERM/CIRMMP infrastructure, supported by the Italian Ministry for University and Research (FOE funding). CF was supported by the Stiftung Polytechnische Gesellschaft. Work in the lab of JH was supported by NSF (RAPID 2030601) and NIH (R01GM123249).

ACKNOWLEDGMENTS

The authors thank Leonardo Gonnelli and Katharina Targaczewski for the valuable technical assistance. IBS acknowledges integration into the Interdisciplinary Research Institute of Grenoble (IRIG CEA). They acknowledge the Advanced Technologies Network Center of the University of Palermo to support infrastructures.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.653148/full#supplementary-material

REFERENCES

Almeida, M. S., Johnson, M. A., Herrmann, T., Geralt, M., and Wüthrich, K. (2007). Novel beta-barrel fold in the nuclear magnetic resonance structure of the replicase nonstructural protein 1 from the severe acute respiratory syndrome coronavirus. *J. Virol.* 81 (7), 3151–3161. doi:10.1128/JVI.01939-06

Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R., and Hilgenfeld, R. (2003). Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* 300 (5626), 1763–1767. doi:10.1126/science.1085658

Bogomolovas, J., Simon, B., Sattler, M., and Stier, G. (2009). Screening of fusion partners for high yield expression and purification of bioactive viscotoxins. *Protein Expr. Purif.* 64 (1), 16–23. doi:10.1016/j.pep.2008.10.003

Bojkova, D., Klann, K., Koch, B., Widera, M., Krause, D., Ciesek, S., et al. (2020). Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* 583 (7816), 469–472. doi:10.1038/s41586-020-2332-7

Bouvet, M., Debarnot, C., Imbert, I., Selisko, B., Snijder, E. J., Canard, B., et al. (2010). *In vitro* reconstitution of SARS-coronavirus mRNA cap methylation. *PLoS Pathog.* 6 (4), e1000863. doi:10.1371/journal.ppat.1000863

Cantini, F., Banci, L., Altincekic, N., Bains, J. K., Dhamotharan, K., Fuks, C., et al. (2020). (1)H, (13)C, and (15)N backbone chemical shift assignments of the apo and the ADP-ribose bound forms of the macrodomain of SARS-CoV-2 non-structural protein 3b. *Biomol. NMR Assign.* 14 (2), 339–346. doi:10.1007/s12104-020-09973-4

Chan, J. F., Kok, K. H., Zhu, Z., Chu, H., To, K. K., Yuan, S., et al. (2020). Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg. Microbes Infect.* 9 (1), 221–236. doi:10.1080/22221751.2020.1719902

Chang, C. K., Hou, M. H., Chang, C. F., Hsiao, C. D., and Huang, T. H. (2014). The SARS coronavirus nucleocapsid protein--forms and functions. *Antivir. Res.* 103, 39–50. doi:10.1016/j.antiviral.2013.12.009

Chang, H. P., Chou, C. Y., and Chang, G. G. (2007). Reversible unfolding of the severe acute respiratory syndrome coronavirus main protease in guanidinium chloride. *Biophys. J.* 92 (4), 1374–1383. doi:10.1529/biophysj.106.091736

Chen, H., Cui, Y., Han, X., Hu, W., Sun, M., Zhang, Y., et al. (2020a). Liquid-liquid phase separation by SARS-CoV-2 nucleocapsid protein and RNA. *Cell Res.* 30, 1143. doi:10.1038/s41422-020-00408-2

Chen, J., Malone, B., Llewellyn, E., Grasso, M., Shelton, P. M. M., Olinares, P. D. B., et al. (2020b). Structural basis for helicase-polymerase coupling in the SARS-CoV-2 replication-transcription complex. *Cell* 182 (6), 1560–1573. doi:10.1016/j.cell.2020.07.033

Chen, P., Jiang, M., Hu, T., Liu, Q., Chen, X. S., and Guo, D. (2007). Biochemical characterization of exoribonuclease encoded by SARS coronavirus. *J. Biochem. Mol. Biol.* 40 (5), 649–655. doi:10.5483/bmbrep.2007.40.5.649

Chen, Y., Savinov, S. N., Mielech, A. M., Cao, T., Baker, S. C., and Mesecar, A. D. (2015). X-ray structural and functional studies of the three tandemly linked domains of non-structural protein 3 (nsp3) from murine hepatitis virus reveal conserved functions. *J. Biol. Chem.* 290 (42), 25293–25306. doi:10.1074/jbc.M115.662130

Cornillez-Ty, C. T., Liao, L., Yates, J. R., 3rd, Kuhn, P., and Buchmeier, M. J. (2009). Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling. *J. Virol.* 83 (19), 10314–10318. doi:10.1128/JVI.00842-09

Cross, T. J., Takahashi, G. R., Diessner, E. M., Crosby, M. G., Farahmand, V., Zhuang, S., et al. (2020). Sequence characterization and molecular modeling of clinically relevant variants of the SARS-CoV-2 main protease. *Biochemistry* 59 (39), 3741–3756. doi:10.1021/acs.biochem.0c00462

David, G., Fogeron, M. L., Schledorn, M., Montserret, R., Haselmann, U., Penzel, S., et al. (2018). Structural studies of self-assembled subviral particles: combining cell-free expression with 110 kHz MAS NMR spectroscopy. *Angew. Chem. Int. Ed. Engl.* 57 (17), 4787–4791. doi:10.1002/anie.201712091

Davies, J. P., Almasy, K. M., McDonald, E. F., and Plate, L. (2020). Comparative multiplexed interactomics of SARS-CoV-2 and homologous coronavirus non-structural proteins identifies unique and shared host-cell dependencies. *bioRxiv* [Epub ahead of print]. doi:10.1101/2020.07.13.201517

Deng, X., Hackbart, M., Mettelman, R. C., O'Brien, A., Mielech, A. M., Yi, G., et al. (2017). Coronavirus nonstructural protein 15 mediates evasion of dsRNA sensors and limits apoptosis in macrophages. *Proc. Natl. Acad. Sci. U.S.A.* 114 (21), E4251–E4260. doi:10.1073/pnas.1618310114

Dinesh, D. C., Chalupska, D., Silhan, J., Koutna, E., Nencka, R., Veverka, V., et al. (2020). Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. *PLoS Pathog.* 16 (12), e1009100. doi:10.1371/journal.ppat.1009100

Dudas, F. D., Puglisi, R., Korn, S. M., Alfano, C., Kelly, G., Monaca, E., et al. (2021). Backbone chemical shift spectral assignments of coronavirus-2 non-structural protein nsp9. *Biomol. NMR Assign.* 2021, 1–10. doi:10.1007/s12104-020-09992-1

Egloff, M. P., Ferron, F., Campanacci, V., Longhi, S., Rancurel, C., Dutartre, H., et al. (2004). The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world. *Proc. Natl. Acad. Sci. U.S.A.* 101 (11), 3792–3796. doi:10.1073/pnas.0307877101

Esposito, D., Mehalko, J., Drew, M., Snead, K., Wall, V., Taylor, T., et al. (2020). Optimizing high-yield production of SARS-CoV-2 soluble spike trimers for serology assays. *Protein Expr. Purif.* 174, 105686. doi:10.1016/j.pep.2020.105686

Finkel, Y., Mizrahi, O., Nachshon, A., Weingarten-Gabbay, S., Morgenstern, D., Yahalom-Ronen, Y., et al. (2020). The coding capacity of SARS-CoV-2. *Nature* 589, 125. doi:10.1038/s41586-020-2739-1

Flower, T. G., Buffalo, C. Z., Hooy, R. M., Allaire, M., Ren, X., and Hurley, J. H. (2020). Structure of SARS-CoV-2 ORF8, a rapidly evolving coronavirus protein implicated in immune evasion. *bioRxiv* [Epub ahead of print]. doi:10.1101/2020.08.27.270637

Fogeron, M. L., Badillo, A., Jirasko, V., Gouttenoire, J., Paul, D., Lancien, L., et al. (2015a). Wheat germ cell-free expression: two detergents with a low critical micelle concentration allow for production of soluble HCV membrane proteins. *Protein Expr. Purif.* 105, 39–46. doi:10.1016/j.pep.2014.10.003

Fogeron, M. L., Badillo, A., Penin, F., and Böckmann, A. (2017). Wheat germ cell-free overexpression for the production of membrane proteins. *Methods Mol. Biol.* 1635, 91–108. doi:10.1007/978-1-4939-7151-0_5

Fogeron, M. L., Jirasko, V., Penzel, S., Paul, D., Montserret, R., Danis, C., et al. (2016). Cell-free expression, purification, and membrane reconstitution for NMR studies of the nonstructural protein 4B from hepatitis C virus. *J. Biomol. NMR* 65 (2), 87–98. doi:10.1007/s10858-016-0040-2

Fogeron, M. L., Paul, D., Jirasko, V., Montserret, R., Lacabanne, D., Molle, J., et al. (2015b). Functional expression, purification, characterization, and membrane reconstitution of non-structural protein 2 from hepatitis C virus. *Protein Expr. Purif.* 116, 1–6. doi:10.1016/j.pep.2015.08.027

Frick, D. N., Virdi, R. S., Vuksanovic, N., Dahal, N., and Silvaggi, N. R. (2020). Molecular basis for ADP-ribose binding to the Mac1 domain of SARS-CoV-2 nsp3. *Biochemistry* 59 (28), 2608–2615. doi:10.1021/acs.biochem.0c00309

Gallo, A., Tsika, A. C., Fourkiotis, N. K., Cantini, F., Banci, L., Sreeramulu, S., et al. (2020). 1H, 13C and 15N chemical shift assignments of the SUD domains of SARS-CoV-2 non-structural protein 3c: "the N-terminal domain-SUD-N". *Biomol. NMR Assign.* 2020, 1–5. doi:10.1007/s12104-020-09987-y

Gao, Y., Yan, L., Huang, Y., Liu, F., Zhao, Y., Cao, L., et al. (2020). Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* 368 (6492), 779–782. doi:10.1126/science.abb7498

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583 (7816), 459–468. doi:10.1038/s41586-020-2286-9

Graham, R. L., Sims, A. C., Brockway, S. M., Baric, R. S., and Denison, M. R. (2005). The nsp2 replicase proteins of murine hepatitis virus and severe acute respiratory syndrome coronavirus are dispensable for viral replication. *J. Virol.* 79 (21), 13399–13411. doi:10.1128/JVI.79.21.13399-13411.2005

Guseva, S., Perez, L. M., Camacho-Zarco, A., Bessa, L. M., Salvi, N., Malki, A., et al. (2021). (1)H, (13)C and (15)N Backbone chemical shift assignments of the n-terminal and central intrinsically disordered domains of SARS-CoV-2 nucleoprotein. *Biomol NMR Assign*. doi:10.1007/s12104-021-10014-x

Hagemeijer, M. C., Monastyrska, I., Griffith, J., van der Sluijs, P., Voortman, J., van Bergen en Henegouwen, P. M., et al. (2014). Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4. *Virology* 458–459, 125–135. doi:10.1016/j.virol.2014.04.027

Hillen, H. S., Kokic, G., Farnung, L., Dienemann, C., Tegunov, D., and Cramer, P. (2020). Structure of replicating SARS-CoV-2 polymerase. *Nature* 584 (7819), 154–156. doi:10.1038/s41586-020-2368-8

Hsu, M. F., Kuo, C. J., Chang, K. T., Chang, H. C., Chou, C. C., Ko, T. P., et al. (2005). Mechanism of the maturation process of SARS-CoV 3CL protease. *J. Biol. Chem.* 280 (35), 31257–31266. doi:10.1074/jbc.M502577200

Huang, Y., Yang, Z. Y., Kong, W. P., and Nabel, G. J. (2004). Generation of synthetic severe acute respiratory syndrome coronavirus pseudoparticles: implications for assembly and vaccine production. *J. Virol.* 78 (22), 12557–12565. doi:10.1128/JVI.78.22.12557-12565.2004

Hurst, K. R., Koetzner, C. A., and Masters, P. S. (2013). Characterization of a critical interaction between the coronavirus nucleocapsid protein and nonstructural protein 3 of the viral replicase-transcriptase complex. *J. Virol.* 87 (16), 9159–9172. doi:10.1128/JVI.01275-13

Ishida, T., and Kinoshita, K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.* 35, W460–W464. doi:10.1093/nar/gkm363

Jia, Z., Yan, L., Ren, Z., Wu, L., Wang, J., Guo, J., et al. (2019). Delicate structural coordination of the severe acute respiratory syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Res.* 47 (12), 6538–6550. doi:10.1093/nar/gkz409

Jiang, H. W., Li, Y., Zhang, H. N., Wang, W., Yang, X., Qi, H., et al. (2020). SARS-CoV-2 proteome microarray for global profiling of COVID-19 specific IgG and IgM responses. *Nat. Commun.* 11 (1), 3581. doi:10.1038/s41467-020-17488-8

Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., et al. (2020). Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* 582 (7811), 289–293. doi:10.1038/s41586-020-2223-y

Jirasko, V., Lakomek, N. A., Penzel, S., Fogeron, M. L., Bartenschlager, R., Meier, B. H., et al. (2020a). Proton-detected solid-state NMR of the cell-free synthesized α-helical transmembrane protein NS4B from hepatitis C virus. *Chembiochem*. 21 (10), 1453–1460. doi:10.1002/cbic.201900765

Jirasko, V., Lends, A., Lakomek, N. A., Fogeron, M. L., Weber, M. E., Malär, A. A., et al. (2020b). Dimer organization of membrane-associated NS5A of hepatitis C virus as determined by highly sensitive 1 H-detected solid-state NMR. *Angew. Chem. Int. Ed.* 60 (10), 5339–5347. doi:10.1002/anie.202013296

Johnson, M. A., Chatterjee, A., Neuman, B. W., and Wüthrich, K. (2010). SARS coronavirus unique domain: three-domain molecular architecture in solution and RNA binding. *J. Mol. Biol.* 400 (4), 724–742. doi:10.1016/j.jmb.2010.05.027

Joseph, J. S., Saikatendu, K. S., Subramanian, V., Neuman, B. W., Brooun, A., Griffith, M., et al. (2006). Crystal structure of nonstructural protein 10 from the severe acute respiratory syndrome coronavirus reveals a novel fold with two zinc-binding motifs. *J. Virol.* 80 (16), 7894–7901. doi:10.1128/JVI.00467-06

Kamitani, W., Narayanan, K., Huang, C., Lokugamage, K., Ikegami, T., Ito, N., et al. (2006). Severe acute respiratory syndrome coronavirus nsp1 protein suppresses host gene expression by promoting host mRNA degradation. *Proc. Natl. Acad. Sci. U.S.A.* 103 (34), 12885–12890. doi:10.1073/pnas.0603144103

Kang, S., Yang, M., Hong, Z., Zhang, L., Huang, Z., Chen, X., et al. (2020). Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharm. Sin. B* 10 (7), 1228–1238. doi:10.1016/j.apsb.2020.04.009

Kern, D. M., Sorum, B., Mali, S. S., Hoel, C. M., Sridharan, S., Remis, J. P., et al. (2020). Cryo-EM structure of the SARS-CoV-2 3a ion channel in lipid nanodiscs. *bioRxiv* 17, 156554. doi:10.1101/2020.06.17.156554

Khan, M. T., Zeb, M. T., Ahsan, H., Ahmed, A., Ali, A., Akhtar, K., et al. (2020). SARS-CoV-2 nucleocapsid and Nsp3 binding: an in silico study. *Arch. Microbiol.* 203, 59. doi:10.1007/s00203-020-01998-6

Kim, Y., Jedrzejczak, R., Maltseva, N. I., Wilamowski, M., Endres, M., Godzik, A., et al. (2020). Crystal structure of Nsp15 endoribonuclease NendoU from SARS-CoV-2. *Protein Sci.* 29 (7), 1596–1605. doi:10.1002/pro.3873

Kirchdoerfer, R. N., and Ward, A. B. (2019). Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nat. Commun.* 10 (1), 2342. doi:10.1038/s41467-019-10280-3

Konkolova, E., Klima, M., Nencka, R., and Boura, E. (2020). Structural analysis of the putative SARS-CoV-2 primase complex. *J. Struct. Biol.* 211 (2), 107548. doi:10.1016/j.jsb.2020.107548

Konno, Y., Kimura, I., Uriu, K., Fukushi, M., Irie, T., Koyanagi, Y., et al. (2020). SARS-CoV-2 ORF3b is a potent interferon antagonist whose activity is increased by a naturally occurring elongation variant. *Cell Rep*. 32 (12), 108185. doi:10.1016/j.celrep.2020.108185

Korn, S. M., Dhamotharan, K., Fürtig, B., Hengesbach, M., Löhr, F., Qureshi, N. S., et al. (2020a). 1H, 13C, and 15N backbone chemical shift assignments of the nucleic acid-binding domain of SARS-CoV-2 non-structural protein 3e. *Biomol. NMR Assign.* 14 (2), 329–333. doi:10.1007/s12104-020-09971-6

Korn, S. M., Lambertz, R., Fürtig, B., Hengesbach, M., Löhr, F., Richter, C., et al. (2020b). 1H, 13C, and 15N backbone chemical shift assignments of the C-terminal dimerization domain of SARS-CoV-2 nucleocapsid protein. *Biomol. NMR Assign.* 2020, 1–7. doi:10.1007/s12104-020-09995-y

Krafcikova, P., Silhan, J., Nencka, R., and Boura, E. (2020). Structural analysis of the SARS-CoV-2 methyltransferase complex involved in RNA cap creation bound to sinefungin. *Nat. Commun.* 11 (1), 3717. doi:10.1038/s41467-020-17495-9

Kubatova, N., Qureshi, N. S., Altincekic, N., Abele, R., Bains, J. K., Ceylan, B., et al. (2020). 1H, 13C, and 15N backbone chemical shift assignments of coronavirus-2 non-structural protein Nsp10. *Biomol. NMR Assign.* 2020, 1–7. doi:10.1007/s12104-020-09984-1

Kusov, Y., Tan, J., Alvarez, E., Enjuanes, L., and Hilgenfeld, R. (2015). A G-quadruplex-binding macrodomain within the "SARS-unique domain" is essential for the activity of the SARS-coronavirus replication-transcription complex. *Virology* 484, 313–322. doi:10.1016/j.virol.2015.06.016

Leao, J. C., Gusmao, T. P. L., Zarzar, A. M., Leao Filho, J. C., Barkokebas Santos de Faria, A., Morais Silva, I. H., et al. (2020). Coronaviridae-old friends, new enemy! *Oral Dis.* 2020, 13447. doi:10.1111/odi.13447

Lei, J., Kusov, Y., and Hilgenfeld, R. (2018). Nsp3 of coronaviruses: structures and functions of a large multi-domain protein. *Antivir. Res.* 149, 58–74. doi:10.1016/j.antiviral.2017.11.001

Li, Q., and Kang, C. (2020). A practical perspective on the roles of solution NMR spectroscopy in drug discovery. *Molecules* 25 (13), 2974. doi:10.3390/molecules25132974

Littler, D. R., Gully, B. S., Colson, R. N., and Rossjohn, J. (2020). Crystal structure of the SARS-CoV-2 non-structural protein 9, Nsp9. *iScience* 23 (7), 101258. doi:10.1016/j.isci.2020.101258

Liu, D. X., Fung, T. S., Chong, K. K., Shukla, A., and Hilgenfeld, R. (2014). Accessory proteins of SARS-CoV and other coronaviruses. *Antivir. Res.* 109, 97–109. doi:10.1016/j.antiviral.2014.06.013

Luo, H., Chen, J., Chen, K., Shen, X., and Jiang, H. (2006). Carboxyl terminus of severe acute respiratory syndrome coronavirus nucleocapsid protein: self-association analysis and nucleic acid binding characterization. *Biochemistry* 45 (39), 11827–11835. doi:10.1021/bi0609319

Ma, Y., Wu, L., Shaw, N., Gao, Y., Wang, J., Sun, Y., et al. (2015). Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. *Proc. Natl. Acad. Sci. U.S.A.* 112 (30), 9436–9441. doi:10.1073/pnas.1508686112

Mandala, V. S., McKay, M. J., Shcherbakov, A. A., Dregni, A. J., Kolocouris, A., and Hong, M. (2020). Structure and drug binding of the SARS-CoV-2 envelope protein transmembrane domain in lipid bilayers. *Nat. Struct. Mol. Biol.* 27, 1202. doi:10.1038/s41594-020-00536-8

Miknis, Z. J., Donaldson, E. F., Umland, T. C., Rimmer, R. A., Baric, R. S., and Schultz, L. W. (2009). Severe acute respiratory syndrome coronavirus nsp9 dimerization is essential for efficient viral growth. *J. Virol.* 83 (7), 3007–3018. doi:10.1128/JVI.01505-08

Mompean, M., Trevino, M. A., and Laurents, D. V. (2020). Towards targeting the disordered SARS-CoV-2 nsp2 C-terminal region: partial structure and dampened mobility revealed by NMR spectroscopy. *bioRxiv* [Epub ahead of print]. doi:10.1101/2020.11.09.374173

Morita, E. H., Sawasaki, T., Tanaka, R., Endo, Y., and Kohno, T. (2003). A wheat germ cell-free system is a novel way to screen protein folding and function. *Protein Sci.* 12 (6), 1216–1221. doi:10.1110/ps.0241203

Narayanan, K., Huang, C., Lokugamage, K., Kamitani, W., Ikegami, T., Tseng, C. T., et al. (2008). Severe acute respiratory syndrome coronavirus nsp1 suppresses host gene expression, including that of type I interferon, in infected cells. *J. Virol.* 82 (9), 4471–4479. doi:10.1128/JVI.02472-07

Narayanan, K., Ramirez, S. I., Lokugamage, K. G., and Makino, S. (2015). Coronavirus nonstructural protein 1: common and distinct functions in the regulation of host and viral gene expression. *Virus. Res.* 202, 89–100. doi:10.1016/j.virusres.2014.11.019

Nelson, C. W., Ardern, Z., Goldberg, T. L., Meng, C., Kuo, C. H., Ludwig, C., et al. (2020). Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic. *Elife* 9, 59633. doi:10.7554/eLife.59633

Netzer, W. J., and Hartl, F. U. (1997). Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* 388 (6640), 343–349. doi:10.1038/41024

Neuman, B. W., Joseph, J. S., Saikatendu, K. S., Serrano, P., Chatterjee, A., Johnson, M. A., et al. (2008). Proteomics analysis unravels the functional repertoire of coronavirus nonstructural protein 3. *J. Virol.* 82 (11), 5279–5294. doi:10.1128/JVI.02631-07

Neuman, B. W., Kiss, G., Kunding, A. H., Bhella, D., Baksh, M. F., Connelly, S., et al. (2011). A structural analysis of M protein in coronavirus assembly and morphology. *J. Struct. Biol.* 174 (1), 11–22. doi:10.1016/j.jsb.2010.11.021

Neuman, B. W. (2016). Bioinformatics and functional analyses of coronavirus nonstructural proteins involved in the formation of replicative organelles. *Antivir. Res.* 135, 97–107. doi:10.1016/j.antiviral.2016.10.005

Oostra, M., Hagemeijer, M. C., van Gent, M., Bekker, C. P., te Lintelo, E. G., Rottier, P. J., et al. (2008). Topology and membrane anchoring of the coronavirus replication complex: not all hydrophobic domains of nsp3 and nsp6 are membrane spanning. *J. Virol.* 82 (24), 12392–12405. doi:10.1128/JVI.01219-08

Oostra, M., te Lintelo, E. G., Deijs, M., Verheije, M. H., Rottier, P. J., and de Haan, C. A. (2007). Localization and membrane topology of coronavirus nonstructural protein 4: involvement of the early secretory pathway in replication. *J. Virol.* 81 (22), 12323–12336. doi:10.1128/JVI.01506-07

Pavesi, A. (2020). New insights into the evolutionary features of viral overlapping genes by discriminant analysis. *Virology* 546, 51–66. doi:10.1016/j.virol.2020.03.007

Robertson, M. P., Igel, H., Baertsch, R., Haussler, D., Ares, M., Jr., and Scott, W. G. (2005). The structure of a rigorously conserved RNA element within the SARS virus genome. *PLoS Biol.* 3 (1), e5. doi:10.1371/journal.pbio.0030005

Robson, F., Khan, K. S., Le, T. K., Paris, C., Demirbag, S., Barfuss, P., et al. (2020). Coronavirus RNA proofreading: molecular basis and therapeutic targeting. *Mol. Cel* 79 (5), 710–727. doi:10.1016/j.molcel.2020.07.027

Rosas-Lemus, M., Minasov, G., Shuvalova, L., Inniss, N. L., Kiryukhina, O., Wiersum, G., et al. (2020). The crystal structure of nsp10-nsp16 heterodimer from SARS-CoV-2 in complex with S-adenosylmethionine. *bioRxiv* [Epub ahead of print]. doi:10.1101/2020.04.17.047498

Salvi, N., Bessa, L. M., Guseva, S., Camacho-Zarco, A., Maurin, D., Perez, L. M., et al. (2021). 1H, 13C and 15N backbone chemical shift assignments of SARS-CoV-2 nsp3a. *Biomol. NMR Assign.* 2021, 1–4. doi:10.1007/s12104-020-10001-8

Schoeman, D., and Fielding, B. C. (2020). Is there a link between the pathogenic human coronavirus envelope protein and immunopathology? A review of the literature. *Front. Microbiol.* 11, 2086. doi:10.3389/fmicb.2020.02086

Schubert, K., Karousis, E. D., Jomaa, A., Scaiola, A., Echeverria, B., Gurzeler, L. A., et al. (2020). SARS-CoV-2 Nsp1 binds the ribosomal mRNA channel to inhibit translation. *Nat. Struct. Mol. Biol.* 27 (10), 959–966. doi:10.1038/s41594-020-0511-8

Serrano, P., Johnson, M. A., Almeida, M. S., Horst, R., Herrmann, T., Joseph, J. S., et al. (2007). Nuclear magnetic resonance structure of the N-terminal domain of nonstructural protein 3 from the severe acute respiratory syndrome coronavirus. *J. Virol.* 81 (21), 12049–12060. doi:10.1128/JVI.00969-07

Serrano, P., Johnson, M. A., Chatterjee, A., Neuman, B. W., Joseph, J. S., Buchmeier, M. J., et al. (2009). Nuclear magnetic resonance structure of the nucleic acid-binding domain of severe acute respiratory syndrome coronavirus nonstructural protein 3. *J. Virol.* 83 (24), 12998–13008. doi:10.1128/JVI.01253-09

Schiavina, M., Pontoriero, L., Uversky, V. N., Felli, I. C., and Pierattelli, R. (2021). The highly flexible disordered regions of the SARS-CoV-2 nucleocapsid N protein within the 1-248 residue construct: Sequence-specific resonance assignments through NMR. *Biomol NMR Assign.* [in press].

Shen, Z., Wang, G., Yang, Y., Shi, J., Fang, L., Li, F., et al. (2019). A conserved region of nonstructural protein 1 from alphacoronaviruses inhibits host gene expression and is critical for viral virulence. *J. Biol. Chem.* 294 (37), 13606–13618. doi:10.1074/jbc.RA119.009713

Shin, D., Mukherjee, R., Grewe, D., Bojkova, D., Baek, K., Bhattacharya, A., et al. (2020). Papain-like protease regulates SARS-CoV-2 viral spread and innate immunity. *Nature* 587 (7835), 657–662. doi:10.1038/s41586-020-2601-5

Snijder, E. J., Bredenbeek, P. J., Dobbe, J. C., Thiel, V., Ziebuhr, J., Poon, L. L., et al. (2003). Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J. Mol. Biol.* 331 (5), 991–1004. doi:10.1016/s0022-2836(03)00865-9

Surya, W., Li, Y., and Torres, J. (2018). Structural model of the SARS coronavirus E channel in LMPG micelles. *Biochim. Biophys. Acta Biomembr.* 1860 (6), 1309–1317. doi:10.1016/j.bbamem.2018.02.017

Sutton, G., Fry, E., Carter, L., Sainsbury, S., Walter, T., Nettleship, J., et al. (2004). The nsp9 replicase protein of SARS-coronavirus, structure and functional insights. *Structure* 12 (2), 341–353. doi:10.1016/j.str.2004.01.016

Takai, K., Sawasaki, T., and Endo, Y. (2010). Practical cell-free protein synthesis system using purified wheat embryos. *Nat. Protoc.* 5 (2), 227–238. doi:10.1038/nprot.2009.207

Takeda, M., Chang, C. K., Ikeya, T., Güntert, P., Chang, Y. H., Hsu, Y. L., et al. (2008). Solution structure of the c-terminal dimerization domain of SARS coronavirus nucleocapsid protein solved by the SAIL-NMR method. *J. Mol. Biol.* 380 (4), 608–622. doi:10.1016/j.jmb.2007.11.093

Tan, J., Vonrhein, C., Smart, O. S., Bricogne, G., Bollati, M., Kusov, Y., et al. (2009). The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes. *Plos Pathog.* 5 (5), e1000428. doi:10.1371/journal.ppat.1000428

Tan, Y., Schneider, T., Leong, M., Aravind, L., and Zhang, D. (2020). Novel immunoglobulin domain proteins provide insights into evolution and pathogenesis of SARS-CoV-2-related viruses. *mBio* 11 (3). doi:10.1128/mBio.00760-20

Thoms, M., Buschauer, R., Ameismeier, M., Koepke, L., Denk, T., Hirschenberger, M., et al. (2020). Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. *Science* 369 (6508), 1249–1255. doi:10.1126/science.abc8665

Tonelli, M., Singarapu, K. K., Makino, S., Sahu, S. C., Matsubara, Y., Endo, Y., et al. (2011). Hydrogen exchange during cell-free incorporation of deuterated amino acids and an approach to its inhibition. *J. Biomol. NMR* 51 (4), 467–476. doi:10.1007/s10858-011-9575-4

Tonelli, M., Rienstra, C., Anderson, T. K., Kirchdoerfer, R., and Henzler-Wildman, K. (2020). 1H, 13C, and 15N backbone and side chain chemical shift assignments of the SARS-CoV-2 non-structural protein 7. *Biomol. NMR Assign.* 2020, 1–5. doi:10.1007/s12104-020-09985-0

Tvarogová, J., Madhugiri, R., Bylapudi, G., Ferguson, L. J., Karl, N., and Ziebuhr, J. (2019). Identification and characterization of a human coronavirus 229E nonstructural protein 8-associated RNA 3′-terminal adenylyltransferase activity. *J. Virol.* 93 (12), e00291–e00319. doi:10.1128/JVI.00291-19

Ullrich, S., and Nitsche, C. (2020). The SARS-CoV-2 main protease as drug target. *Bioorg. Med. Chem. Lett.* 30 (17), 127377. doi:10.1016/j.bmcl.2020.127377

Vennema, H., Godeke, G. J., Rossen, J. W., Voorhout, W. F., Horzinek, M. C., Opstelten, D. J., et al. (1996). Nucleocapsid-independent assembly of coronavirus-like particles by co-expression of viral envelope protein genes. *EMBO J.* 15 (8), 2020–2028. doi:10.1002/j.1460-2075.1996.tb00553.x

Wacker, A., Weigand, J. E., Akabayov, S. R., Altincekic, N., Bains, J. K., Banijamali, E., et al. (2020). Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy. *Nucleic Acids Res.* 48, 12415. doi:10.1093/nar/gkaa1013

Wang, Y., Kirkpatrick, J., Zur Lage, S., Korn, S. M., Neissner, K., Schwalbe, H., et al. (2021). (1)H, (13)C, and (15)N backbone chemical-shift assignments of SARS-CoV-2 non-structural protein 1 (leader protein). *Biomol NMR Assign.* doi:10.1007/s12104-021-10019-6

Wang, S., Fogeron, M. L., Schledorn, M., Dujardin, M., Penzel, S., Burdette, D., et al. (2019). Combining cell-free protein synthesis and NMR into a tool to study capsid assembly modulation. *Front. Mol. Biosci.* 6, 67. doi:10.3389/fmolb.2019.00067

Wolff, G., Limpens, R. W. A. L., Zevenhoven-Dobbe, J. C., Laugks, U., Zheng, S., de Jong, A. W. M., et al. (2020). A molecular pore spans the double membrane of the coronavirus replication organelle. *Science* 369 (6509), 1395–1398. doi:10.1126/science.abd3629

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579 (7798), 265–269. doi:10.1038/s41586-020-2008-3

Wu, Z., Yang, L., Ren, X., Zhang, J., Yang, F., Zhang, S., et al. (2016). ORF8-related genetic evidence for Chinese horseshoe bats as the source of human severe acute respiratory syndrome coronavirus. *J. Infect. Dis.* 213 (4), 579–583. doi:10.1093/infdis/jiv476

Ye, Q., West, A. M. V., Silletti, S., and Corbett, K. D. (2020). Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein. *Protein Sci.* 29, 1890. doi:10.1002/pro.3909

Yin, W., Mao, C., Luan, X., Shen, D. D., Shen, Q., Su, H., et al. (2020). Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science* 368 (6498), 1499–1504. doi:10.1126/science.abc1560

Yoshimoto, F. K. (2020). The proteins of severe acute respiratory syndrome coronavirus-2 (SARS CoV-2 or n-COV19), the cause of COVID-19. *Protein J.* 39 (3), 198–216. doi:10.1007/s10930-020-09901-4

Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., et al. (2020). Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* 368 (6489), 409–412. doi:10.1126/science.abb3405

Zhao, J., Falcón, A., Zhou, H., Netland, J., Enjuanes, L., Pérez Breña, P., et al. (2009). Severe acute respiratory syndrome coronavirus protein 6 is required for optimal replication. *J. Virol.* 83 (5), 2368–2373. doi:10.1128/JVI.02371-08

Zhou, R., Zeng, R., Von Brunn, A., and Lei, J. (2020). Structural characterization of the C-terminal domain of SARS-CoV-2 nucleocapsid protein. *Mol. Biomed.* 1 (2), 1–11. doi:10.1186/s43556-020-00001-4

Züst, R., Cervantes-Barragán, L., Kuri, T., Blakqori, G., Weber, F., Ludewig, B., et al. (2007). Coronavirus non-structural protein 1 is a major pathogenicity factor: implications for the rational design of coronavirus vaccines. *PLoS Pathog.* 3 (8), e109. doi:10.1371/journal.ppat.0030109

Conflict of Interest: CH was employed by Signals GmbH & Co. KG.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

GLOSSARY

aa Amino acid

BEST Band-selective excitation short-transient

BMRB Biomagnetic resonance databank

CFPS Cell-free protein synthesis

CoV Coronavirus

CTD C-terminal domain

DEDD Asp-Glu-Glu-Asp

DMS Dimethylsulfate

E Envelope protein

ED Ectodomain

fl Full-length

GB1 Protein G B1 domain

GD Globular domain

GF Gel filtration

GST Glutathione-S-transferase

His Hisx-tag

HSP Heat-shock protein

HSQC Heteronuclear single quantum coherence

IDP Intrinsically disordered protein

IDR Intrinsically disordered region

IEC Ion exchange chromatography

IMAC Immobilized metal ion affinity chromatography

IPRS IMAC-protease cleavage-reverse IMAC-SEC;

M Membrane protein

MERS Middle East Respiratory Syndrome

MHV Murine hepatitis virus

M^{pro} Main protease

MTase Methyltransferase

N Nucleocapsid protein

NAB Nucleic acid–binding domain

nsp Nonstructural protein

NTD N-terminal domain

PL^{pro} Papain-like protease

RdRP RNA-dependent RNA polymerase

S Spike protein

SARS Severe Acute Respiratory Syndrome

SEC Size-exclusion chromatography

SUD SARS unique domain

SUMO Small ubiquitin-related modifier

TEV Tobacco etch virus

TM Transmembrane

TROSY Transverse relaxation-optimized spectroscopy

Trx Thioredoxin

Ubl Ubiquitin-like domain

Ulp1 Ubiquitin-like specific protease 1

WG Wheat-germ.

5.10 **Research article:** ^{1}H, ^{13}C and ^{15}N chemical shift assignment of the stem-loop 5a from the 5'-UTR of SARS-CoV-2

Robbin Schnieders*, Stephen A. Peter*, Elnaz Banijamali, Magdalena Riad, Nadide Altincekic, Jasleen Kaur Bains, Betül Ceylan, Boris Fürtig, J. Tassilo Grün, Martin Hengesbach, Katharina F. Hohmann, Daniel Hymon, Bozana Knezic, Andreas Oxenfarth, Katja Petzold, Nusrat S. Qureshi, Christian Richter, Judith Schlagnitweit, Andreas Schlundt, Harald Schwalbe, Elke Stirnal, Alexey Sudakov, Jennifer Vögele, Anna Wacker, Julia E. Weigand, Julia Wirmer-Bartoschek and Jens Wöhnert; *Biomolecular NMR Assignments* **15**, 203-211 (2021)

* These authors contributed equally to this work.

In this work, the stem-loop 5a (SL5a) located in the 5'-UTR of the SARS-CoV-2 genome was characterized by NMR spectroscopy. ^{1}H, ^{13}C, and ^{15}N chemical shift assignments of SL1 were presented as a basis for further in-depth structural and ligand screening studies.

The project was designed and coordinated by the Covid-19 NMR consortium. R. Schnieders and S.A. Peter assigned the NMR spectra and wrote the manuscript. The author of this thesis was part of the Covid-19 NMR consortium and the RNA production team that prepared and purified various isotope-labeled RNA samples for NMR spectroscopy.

https://doi.org/10.1007/s12104-021-10007-w

ARTICLE

^{1}H, ^{13}C and ^{15}N chemical shift assignment of the stem-loop 5a from the 5′-UTR of SARS-CoV-2

Robbin Schnieders[2,4] · Stephen A. Peter[1] · Elnaz Banijamali[6] · Magdalena Riad[6] · Nadide Altincekic[2,4] · Jasleen Kaur Bains[2,4] · Betül Ceylan[2,4] · Boris Fürtig[2,4] · J. Tassilo Grün[2,4] · Martin Hengesbach[2,4] · Katharina F. Hohmann[2,4] · Daniel Hymon[2,4] · Bozana Knezic[2,4] · Andreas Oxenfarth[2,4] · Katja Petzold[6] · Nusrat S. Qureshi[5] · Christian Richter[2,4] · Judith Schlagnitweit[6] · Andreas Schlundt[3,4] · Harald Schwalbe[2,4] · Elke Stirnal[2,4] · Alexey Sudakov[2,4] · Jennifer Vögele[3,4] · Anna Wacker[2,4] · Julia E. Weigand[1] · Julia Wirmer-Bartoschek[2,4] · Jens Wöhnert[3,4]

Received: 9 December 2020 / Accepted: 8 January 2021

Abstract

The SARS-CoV-2 (SCoV-2) virus is the causative agent of the ongoing COVID-19 pandemic. It contains a positive sense single-stranded RNA genome and belongs to the genus of Betacoronaviruses. The 5′- and 3′-genomic ends of the 30 kb SCoV-2 genome are potential antiviral drug targets. Major parts of these sequences are highly conserved among Betacoronaviruses and contain *cis*-acting RNA elements that affect RNA translation and replication. The 31 nucleotide (nt) long highly conserved stem-loop 5a (SL5a) is located within the 5′-untranslated region (5′-UTR) important for viral replication. SL5a features a U-rich asymmetric bulge and is capped with a 5′-UUUCGU-3′ hexaloop, which is also found in stem-loop 5b (SL5b). We herein report the extensive ^{1}H, ^{13}C and ^{15}N resonance assignment of SL5a as basis for in-depth structural studies by solution NMR spectroscopy.

Keywords SARS-CoV-2 · 5′-UTR · SL5a · Solution NMR spectroscopy · COVID19-NMR

Biological context

SCoV-2 is a member of the Betacoronavirus family and contains a large single-stranded (+) RNA genome with a length of approx. 30,000 nucleotides (nts) (Hu et al. 2020; V'kovski et al. 2020). The RNA genome of the virus not only contains the coding regions for the viral proteins, but also extended and highly structured 5′- and 3′-UTRs, as well as internal structured RNA elements with important functional roles

Robbin Schnieders and Stephen A. Peter have equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12104-021-10007-w.

✉ Harald Schwalbe
Schwalbe@nmr.uni-frankfurt.de;
covid19-nmr@dlist.server.uni-frankfurt.de

[1] Department of Biology, Technical University of Darmstadt, Schnittspahnstr. 10, 64287 Darmstadt, Germany

[2] Institute for Organic Chemistry and Chemical Biology, Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue-Str. 7, 60438 Frankfurt/M., Germany

[3] Institute for Molecular Biosciences, Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue-Str. 9, 60438 Frankfurt/M., Germany

[4] Center for Biomolecular Magnetic Resonance (BMRZ), Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue-Str. 9, 60438 Frankfurt/M., Germany

[5] Present Address: EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany

[6] Department of Medical Biochemistry and Biophysics, Karolinska Institute, Biomedicum, Solnavägen 9, 17177 Stockholm, Sweden

Published online: 23 January 2021

in genome replication, transcription of subgenomic (sg) mRNAs and the balanced translation of viral proteins (Madhugiri et al. 2016; Kelly et al. 2020; Tidu et al. 2020). While the development of antiviral therapeutics against COVID-19 is primarily focused on the viral proteins, the highly structured RNA elements provide an extensive reservoir of additional drug targets to be exploited. The architecture of the RNA genome of SCoV2 and related viruses has so far been investigated mainly by sequence-based computational predictions and by chemical probing approaches in vitro and in vivo (e.g. Manfredonia et al. 2020; Rangan et al. 2020). Although structural probing methods have been established to map RNA-small molecule interactions even in cells (Martin et al. 2019), these tools are unable to define the tertiary structure and dynamics of the RNA-elements in the SCoV-2 genome with sufficiently high resolution to enable structure-based drug design by virtual screening.

While the sequences of the individual structural elements vary between different Coronaviruses, their ubiquitous presence and highly conserved secondary structures suggest that these elements are critically important for viral viability and pathogenesis (reviewed in Madhugiri et al. 2016). One example of such an important structure is stem-loop 5 (SL5). SL5 is structurally conserved in the genomes of Alpha- and Betacoronaviruses and has been shown to be crucial for efficient viral replication (Chen and Olsthoorn 2010; Guan et al. 2011).

In SCoV-2, SL5 consists of four helices including nts 149–297 of the 5′-UTR and the first 29 nts of the Nsp1 coding region (Suppl. Figure 1A). Sub-elements are joined to the SL5 basal stem by a four-helix junction. These sub-elements are termed SLs 5a, 5b and 5c. SL5a consists of 31 nucleotides and represents the largest of the three stem-loops. Intriguingly, the apical loop sequences of SL5a and SL5b are identical (5′-UUUCGU-3′) and belong to the 5′-UUYCGU-3′ motif, which is also found in Alphacoronaviruses. This high level of sequence conservation suggests functional importance, e.g. in viral packaging (Masters 2019). Thus, we have recently obtained secondary structure models of SL5a-c and the basal stem segment of SL5 based on initial ^{1}H and ^{15}N assignments (Wacker et al. 2020). In order to characterize SL5a further, we provide here a near complete ^{1}H, ^{13}C and ^{15}N chemical shift assignment.

Methods and experiments

Sample preparation

RNA synthesis for NMR experiments: For DNA template production, the sequence of SL5a together with the T7 promoter was generated by hybridization of complementary oligonucleotides and introduced into the *Eco*RI and *Nco*I sites of an HDV ribozyme encoding plasmid (Schürer et al. 2002), based on the pSP64 vector (Promega). RNAs were transcribed as HDV ribozyme fusions to obtain a homogeneous 3′-end. The recombinant vector pHDV-5_SL5a was transformed and amplified in the *Escherichia coli* strain DH5α. Plasmid-DNA was purified using a large scale DNA isolation kit (Gigaprep; Qiagen) according to the manufacturer's instructions and linearized with *Hind*III prior to in-vitro transcription using the T7 RNA polymerase P266L mutant, which was prepared as described in (Guillerez et al. 2005). 15 ml transcription reactions [20 mM DTT, 2 mM spermidine, 200 ng/µl template, 200 mM Tris/glutamate (pH 8.1), 40 mM $Mg(OAc)_2$, 12 mM NTPs, 32 µg/ml T7 RNA Polymerase, 20% DMSO] were performed to obtain sufficient amounts of SL5a RNA (5′-pppGGGCUGCUUACGGUU UCGUCCGUGUUGCAGCCC-3′). Preparative transcription reactions (6 h at 37 °C and 70 rpm) were terminated by addition of 150 mM EDTA. SL5a RNA was purified as follows: RNAs were precipitated with one sample volume of ice-cold 2-propanol. RNA fragments were separated on 15% denaturing polyacrylamide (PAA) gels and visualized by UV shadowing at 254 nm. SL5a RNA was excised from the gel and eluted using the following protocol: The gel fragments were granulated in two gel volumes 0.3 M NaOAc solution, incubated for 30 min at − 80 °C, followed by 15 min at 65 °C. The RNA was further eluted from gel fragments overnight by passive diffusion into 0.3 M NaOAc, precipitated with EtOH and desalted via PD10 columns (GE Healthcare). Residual PAA was removed by reversed-phase HPLC using a Kromasil RP 18 column and a gradient of 0–40% 0.1 M acetonitrile/triethylammonium acetate. After freeze-drying of RNA-containing fractions and cation exchange by $LiClO_4$ precipitation (2% in acetone), the RNA was folded in water by heating to 80 °C followed by rapid cooling on ice. Buffer exchange to NMR buffer (25 mM potassium phosphate buffer, pH 6.2, 50 mM potassium chloride) was performed using Vivaspin centrifugal concentrators (2 kDa molecular weight cut-off). Purity of SL5a was verified by denaturing PAA gel electrophoresis and homogenous folding was monitored by native PAA gel electrophoresis, loading the same RNA concentration as used in NMR experiments.

Using this protocol, two NMR samples of SL5a, an 810 µM uniformly ^{15}N- and a 680 µM uniformly ^{13}C,^{15}N-labeled sample, were prepared and used for the assignment presented herein.

NMR experiments

NMR experiments using the ^{15}N-labeled RNA were carried out at the Karolinska Institute (KI) using a Bruker AVANCEIII 600 MHz NMR spectrometer equipped with a 5 mm, z-axis gradient ^{1}H [^{13}C, ^{15}N, ^{31}P]-QCI cryogenic probe. All NMR experiments with the ^{13}C,^{15}N-labeled RNA were conducted at the Center for Biomolecular Magnetic Resonance (BMRZ) at the Goethe University (GU) Frankfurt using Bruker AVIIIHD NMR spectrometers from 600 to 800 MHz, which are equipped with the following cryogenic probes: 5 mm, z-axis gradient ^{1}H [^{13}C,^{31}P]-TCI cryogenic probe (600 MHz), 5 mm, z-axis gradient ^{1}H [^{13}C, ^{15}N, ^{31}P]-QCI cryogenic probe (700 MHz) and ^{13}C-optimized 5 mm, z-axis gradient ^{13}C, ^{15}N [^{1}H]-TXO cryogenic probe (800 MHz).

At BMRZ and KI, experiments were performed at 298 K if not indicated otherwise. NMR spectra were processed and analyzed using Topspin versions 4.0.8 (GU) and 3.6.2 (KI). The chemical shift assignment was conducted using Sparky (Lee et al. 2015). NMR data were managed and archived using the platform LOGS (2020, version 2.1.54, Signals GmbH & Co KG, www.logs.repository.com). ^{1}H chemical shifts were referenced externally to DSS, and ^{13}C and ^{15}N chemical shifts were indirectly referenced from the ^{1}H chemical shift as described earlier (Wishart et al. 1995).

We have previously reported the imino and cytidine amino resonance assignment of SL5a (Wacker et al. 2020) that allowed us to determine the base pairing in this RNA element. The location of stable base pairs is confirmed by through space $^{2h}J_{NN}$ coupling constants (Dingley et al. 2008) reported in Suppl. Table S1. These assignments were available from experiments conducted on a ^{15}N-labeled RNA sample and provided starting points of the aromatic proton resonance assignment using ^{1}H,^{1}H-NOESY (Tables 1 I, 2 I) and (H)C(CCN)H (Tables 1 IV, 2 V) experiments linking the imino proton resonances to the aromatic protons and carbons (Fig. 1a and b). The remaining H6/8–C6/8 resonances in the aromatic ^{1}H,^{13}C-HSQC spectrum (Tables 1 II, 2 III) were assigned using a 3D ^{13}C-NOESY-HSQC experiment (Table 1 VII), which was selective for the aromatic region. Cytidine and uridine C5-H5 resonances were assigned using ^{1}H,^{1}H-TOCSY (Table 1 VI, Fig. 1e) and ^{1}H,^{13}C-HSQC spectra (Table 1 III, Fig. 1d). Furthermore, quaternary carbon atoms were assigned using an HNCO type experiment (Table 2 IV) and the TROSY relayed HCCH-COSY experiment (Table 1 VIII). The ^{13}C-detected 3D CNC spectrum (Table 1 V, Fig. 1c) linked the aromatic carbons to the anomeric C1′ resonances, where the nitrogen dimension aided in distinguishing between purine and pyrimidine nucleotides as well as between uridines and cytidines. Also, by correlating C6/8 to C1′, resonance overlap is minimized given the broader signal distribution in the carbon as opposed to the respective proton dimensions. Based on C1′ resonances obtained from the CNC spectrum and from sequential assignment in the NOESY spectra, H1′–C1′ correlations were assigned in the ^{1}H,^{13}C-HSQC spectrum (Table 1 III, Fig. 1f). A continuous sequential walk of H1′-to-H6/H8 was possible for both helices (Fig. 1c). The H1′–C1′ assignment was further confirmed with a 3D ^{13}C-NOESY-HSQC experiment (Table 1 IX), which was selective for the C1′ resonances. Using two different 3D HCCH TOCSY experiments (Table 1 X, XI and XII), the remaining ribose carbon resonances C2′–C5′ were assigned. The two experiments differed in the TOCSY mixing time such that with a short mixing time of 6 ms, C2′ and C3′ resonances could be distinguished by intensity differences, while with a long mixing time of 18 ms also C4′ and C5′ carbons were correlated to the C1′ resonances.

The U-rich bulge

One of the structural features of the SL5a RNA is an asymmetric U-rich bulge (Fig. 1c). In this likely more dynamic part of the RNA, a near to complete sequential walk (H6/8 to H6/8 or H1′) was possible and thus, all aromatic H6/8–C6/8 correlations were assigned. With the aromatic assignment at hand, the strong imino resonance of a uridine involved in non-canonical base pairing was assigned to residue U194 using the (H)C(CCN)H experiment at 283 K. From observation of this signal, the formation of a base pairing involving U194 and likely either U211 or U212 is suggested. This is further supported by an imino-to-imino NOE contact between U194 and a non-canonical uridine at 273 K. Furthermore, from the U194 carbon chemical shifts in the HNCO experiment, we conclude that the hydrogen bonding interaction is mediated through the C2 carbonyl group (Fürtig et al. 2003; Ohlenschläger et al. 2004). The existence of a GU- wobble base pair involving residues U195 and G210 has not been confirmed, yet. However, broadened imino proton resonances for an additional guanosine and uridine, which are taking part in non-canonical interactions, are observed at low temperature (283 K).

The 5′-UUUCGU-3′ hexaloop

In addition to the U-rich asymmetric bulge (Fig. 1c), SL5a features a 5′-UUUCGU-3′ hexaloop, which also caps the helix of SL5b in the 5′-UTR. Except for residue U205, all aromatic loop assignments were derived from sequential NOE correlations, e.g. H6/8 to H5 or H1′ to H6/8 sequential contacts. Since the central residues of this loop sequence, 5′-UUCG-3′, resemble a highly abundant and well-characterized tetraloop sequence (Cheong et al. 1990; Fürtig et al. 2004; Nozinovic et al. 2010), we asked, whether structural features of this UUCG tetraloop are also found within the 5′-UUUCGU-3′ hexaloop of SL5a. While the

Table 1 List of NMR experiments conducted at KI and BMRZ at 298 K

#	NMR experiment	Experimental parameters	Characteristic parameters
I	*^{1}H,^{1}H-NOESY^{KI}* Jump-return water suppression	A Aromatics: 600 MHz, ns: 64, sw(f2): 21.0 ppm, sw(f1): 11.9 ppm, aq(f2): 81 ms, aq(f1): 56 ms, o1(^{1}H): 4.7 ppm, o3(^{15}N): 153 ppm, rel. delay: 1.5 s, time: 25 h B Iminos: 600 MHz, ns: 128, sw(f2): 21.0 ppm, sw(f1): 11.9 ppm, aq(f2): 81 ms, aq(f1): 25 ms, o1(^{1}H): 4.7 ppm, o3(^{15}N): 153 ppm, rel. delay: 1.0 s, time: 16 h	A NOE mixing time 150 ms, JR-delay 167 μs B NOE mixing time 150 ms, JR-delay 52 μs
II	*^{1}H,^{13}C-HSQC^{BMRZ}* Aromatic region (Bodenhausen and Ruben 1980)	800 MHz, ns: 8, sw(f2): 8.6 ppm, sw(f1): 24.8 ppm, aq(f2): 75 ms, aq(f1): 38 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 143 ppm, o3(^{15}N): 153 ppm, rel. delay: 1.0 s, time: 1 h	INEPT transfer time 2.7 ms (^{1}J$_{CH}$ 185 Hz), off-resonant Q3 shaped pulse for C5 decoupling at 95 ppm with 25 ppm bandwidth
III	*^{1}H,^{13}C-CT-HSQC^{BMRZ}* Full (Vuister and Bax 1992)	700 MHz, ns: 40, sw(f2): 10.0 ppm, sw(f1): 95.6 ppm, aq(f2): 73 ms, aq(f1): 17 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 107 ppm, o3(^{15}N): 153 ppm, rel. delay: 1.0 s, time: 7 h	INEPT transfer time 2.8 ms (^{1}J$_{CH}$ 180 Hz), CT period 25 ms (^{1}J$_{CC}$ 40 Hz)
IV	*(H)C(CCN)H^{BMRZ}* Aromatics-to-imino (Piotto et al. 1992; Sklenár et al. 1996; Wöhnert et al. 2003)	800 MHz, ns: 256, sw(f3): 20.9 ppm, sw(f2): 9.9 ppm, aq(f3): 67 ms, aq(f2): 32 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 137 ppm, o3(^{15}N): 154 ppm, rel. delay: 1.5 s, time: 16 h	CC-TOCSY mixing time 28 ms
V	*3D ^{13}C-detected CNC^{BMRZ}* C6/8-to-C1′ (Sklenar et al. 1993a)	800 MHz, ns: 24, sw(f3): 24.5 ppm, sw(f2): 34.7 ppm, sw(f1): 12.0 ppm, aq(f3): 67 ms, aq(f2): 6 ms, aq(f1): 20 ms, o1(^{13}C): 90 ppm, o2(^{1}H): 4.7 ppm, o3(^{15}N): 157 ppm, rel. delay: 0.5 s, time: 48d h	C6/8-N1/9 transfer time 30 ms, C–H transfer time 2.9 ms (1′) and 2.6 ms (6/8)
VI	*^{1}H,^{1}H-TOCSY^{KI}* Excitation sculpting water suppression (Shaka et al. 1988; Shaka and Hwang 1995)	600 MHz, ns: 16, sw(f2): 8.8 ppm, sw(f1): 6.2 ppm, aq(f2): 100 ms, aq(f1): 51 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 99 ppm, o3(^{15}N): 86 ppm, rel. delay: 1.5 s, time: 3 h	CC-TOCSY mixing time 40 ms
VII	*3D ^{13}C-NOESY-HSQC^{BMRZ}* Aromatics (Piotto et al. 1992; Sklenar et al. 1993b)	800 MHz, ns: 16, sw(f3,^{1}H): 8.8 ppm, sw(f2,^{13}C): 22.0 ppm, sw(f1,^{1}H): 6.2 ppm, aq(f3): 73 ms, aq(f2): 7 ms, aq(f1): 20 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 142.5 ppm, o3(^{15}N): 185 ppm, rel. delay: 1.0 s, time: 3 days 4 h	NOE mixing time 200 ms
VIII	*3D TROSY-HCCH-COSY^{BMRZ}* Adenine C2-to-C8 (Simon et al. 2001)	800 MHz, ns: 16, sw(f3,^{1}H): 9.0 ppm, sw(f2,^{13}C): 24.8 ppm, sw(f1,^{13}C): 58.5 ppm, aq(f3): 71 ms, aq(f2): 6 ms, aq(f1): 5 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 142.5 ppm, o3(^{15}N): 150 ppm, rel. delay: 1.0 s, time: 1 day 19 h	Bruker standard parameter set
IX	*3D ^{13}C-NOESY-HSQC^{BMRZ}* Ribose (Piotto et al. 1992; Sklenar et al. 1993b)	600 MHz, ns: 16, sw(f3,^{1}H): 10.4 ppm, sw(f2,^{13}C): 43.0 ppm, sw(f1,^{1}H): 8.3 ppm, aq(f3): 82 ms, aq(f2): 5 ms, aq(f1): 13 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 82.5 ppm, o3(^{31}P): – 1 ppm, rel. delay: 1.3 s, time: 2 days 10 h	NOE mixing time 200 ms, HSQC transfer time 2.9 ms (^{1}J$_{CH}$ 170 Hz)
X	*3D (H)CCH-TOCSY^{BMRZ}* ribose C1′-to-C2′ (Kay et al. 1993; Richter et al. 2010)	600 MHz, ns: 8, sw(f3,^{1}H): 10.4 ppm, sw(f2,^{13}C): 10.0 ppm, sw(f1,^{13}C): 35.4 ppm, aq(f3): 82 ms, aq(f2): 21 ms, aq(f1): 12 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 76.5 ppm, o3(^{31}P): – 1 ppm, rel. delay: 1.3 s, time: 1 day 3 h	CC-TOCSY mixing time 6 ms
XI	*3D (H)CCH-TOCSY^{BMRZ}* ribose C1′-to-C5′ (Kay et al. 1993; Richter et al. 2010)	600 MHz, ns: 8, sw(f3,^{1}H): 10.4 ppm, sw(f2,^{13}C): 10.0 ppm, sw(f1,^{13}C): 35.4 ppm, aq(f3): 82 ms, aq(f2): 21 ms, aq(f1): 12 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 76.5 ppm, o3(^{31}P): – 1 ppm, rel. delay: 1.3 s, time: 1 day 3 h	CC-TOCSY mixing time 18 ms

Table 1 (continued)

#	NMR experiment	Experimental parameters	Characteristic parameters
XII	*3D H(C)CH-TOCSY^{BMRZ}* Ribose C1′-to-H5′ (Kay et al. 1993; Richter et al. 2010)	700 MHz, ns: 16, sw(f3,^{1}H): 8.6 ppm, sw(f2,^{13}C): 10.0 ppm, sw(f1,^{1}H): 3.2 ppm, aq(f3): 85 ms, aq(f2): 25 ms, aq(f1): 39 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 76.5 ppm, o3(^{15}N): 153 ppm, rel. delay: 1.0 s, time: 20 h	CC-TOCSY mixing time 16 ms

Experimental parameters and experiment-specific parameters are given

ns number of scans, *sw* spectral width, *aq* acquisition time, *o1/2/3* carrier frequencies on channels 1/2/3, *rel. delay* relaxation delay, *CT* constant time, *JR* jump-return

Table 2 List of NMR experiments conducted at KI and BMRZ at 283 K

#	NMR experiment	Experimental parameters	Characteristic parameters
I	*^{1}H,^{1}H-NOESY^{KI}* Jump-return water suppression	600 MHz, ns: 128, sw(f2): 21.0 ppm, sw(f1): 11.9 ppm, aq(f2): 81 ms, aq(f1): 25 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 99 ppm, o3(^{15}N): 153 ppm, rel. delay: 1.0 s, time: 16 h	NOE mixing time 150 ms, JR-delay 52 µs
II	*BEST-TROSY-HNN-COSY^{BMRZ}* Across hydrogen bond (Dingley and Grzesiek 1998; Wöhnert et al. 1999; Dingley et al. 2008)	600 MHz, ns: 64, sw(f3): 21.5 ppm, sw(f1): 98.9 ppm, aq(f3): 60 ms, aq(f1): 10 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 140 ppm, o3(^{15}N): 184 ppm, rel. delay: 0.3 s, time: 1.5 h	NN-transfer time 30 ms
III	*^{1}H,^{13}C-HSQC^{BMRZ}* Aromatic region (Bodenhausen and Ruben 1980)	950 MHz, ns: 4, sw(f2): 8.6 ppm, sw(f1): 26.1 ppm, aq(f2): 63 ms, aq(f1): 31 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 143 ppm, o3(^{15}N): 153 ppm, rel. delay: 1.0 s, time: 0.5 h	INEPT transfer time 2.7 ms (^{1}J$_{CH}$ 185 Hz), off-resonant Q3 shaped pulse for C5 decoupling at 95 ppm with 25 ppm bandwidth
IV	*H(N)CO^{BMRZ}* Imino-to-carbon (Favier and Brutscher 2011; Solyom et al. 2013)	950 MHz, ns: 32, sw(f3): 21.0 ppm, sw(f1): 27.9 ppm, aq(f3): 60 ms, aq(f1): 10 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 157.5 ppm, o3(^{15}N): 153 ppm, rel. delay: 0.3 s, time: 0.5 h	NC-INEPT transfer time 18 ms (^{1}J$_{CN}$ 28 Hz)
V	*(H)C(CCN)H^{BMRZ}* Aromatics-to-imino (Piotto et al. 1992; Sklenár et al. 1996; Wöhnert et al. 2003)	700 MHz, ns: 288, sw(f3): 23.0 ppm, sw(f2): 10.0 ppm, aq(f3): 70 ms, aq(f2): 36 ms, o1(^{1}H): 4.7 ppm, o2(^{13}C): 137 ppm, o3(^{15}N): 154 ppm, rel. delay: 1.8 s, time: 21 h	CC-TOCSY mixing time 28 ms

Experimental parameters and experiment-specific parameters are given

ns number of scans, *sw* spectral width, *aq* acquisition time, *o1/2/3* carrier frequencies on channels 1/2/3, *rel. delay* relaxation delay, *JR* jump-return

characteristic imino proton resonances of the sheared GU base pair in the 5′-UUCG-3′ tetraloop remained elusive in SL5a spectra (e.g. ^{1}H 1D or ^{1}H,^{15}N-HSQC), ^{1}H,^{13}C-HSQC spectra of the ribose region of SL5a and a 14 nt RNA with a 5′-cUUCGg-3′ tetraloop (secondary structure Suppl. Figure 1B) yielded a similar peak pattern (Fig. 2a and b). Here, it is evident that the chemical shifts of the central two nucleotides of the 5′-UUUCGU-3′ hexaloop, U202 and C203, are in good agreement with the respective counterparts in the 5′-cUUCGg-3′ tetraloop. This observation is also reflected in the canonical coordinates (Ebrahimi et al. 2001; Cherepanov et al. 2010), which suggest the ribofuranosyl ring to adopt the C2′-endo conformation for U202 and C203, while the remaining nucleotides (with a complete ribose carbon assignment) adopt the canonical C3′-endo conformation (Fig. 2c). These spectral data suggest a structural similarity between the middle part of the 5′-UUUCGU-3′ hexa- and 5′-cUUCGg-3′ tetraloop. This might not hold true to the same extent for the flanking residues U201 and G204 as characteristic resonances are absent in the ^{1}H,^{13}C-HSQC

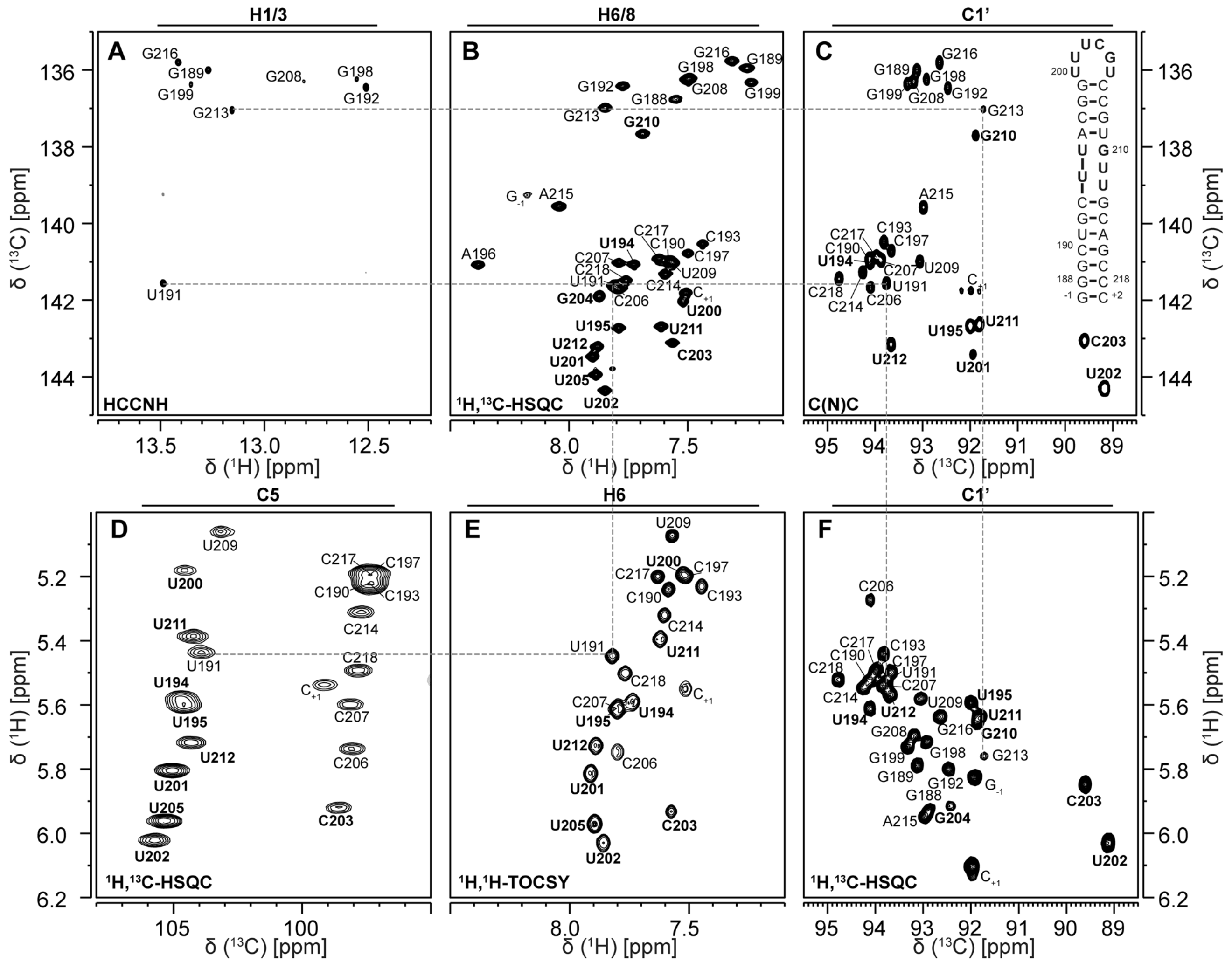


Fig. 1 Resonance assignment of aromatic protons and carbons and the linkage to the ribose. **a** HCCNH experiment correlating the imino protons of guanosines and uridines to the corresponding intranucleobase C8 and C6 resonances, respectively. **b** ^{1}H,^{13}C-HSQC spectrum showing the aromatic H6/8–C6/8 correlations. **c** 2D plane of the ^{13}C-detected CNC-experiment correlating C6/8 to C1′. **d** Transposed ^{1}H,^{13}C-HSQC spectrum showing the H5–C5 correlations for uridines and cytidines. **e** ^{1}H,^{1}H-TOCSY spectrum linking H5 and H6 in pyrimidines. **f** Transposed ^{1}H,^{13}C-HSQC spectrum of the H1′–C1′ region. Panel **c** further shows the secondary structure of SL5a with genomic numbering. Positive contours are given in black, negative contours are held in red. Experimental details are given in Table 1. Exemplary connections between the displayed spectra are demonstrated with the gray dashed lines for residues G213 and U191. Assignments of the asymmetric bulge and the apical loop are highlighted with bold font

spectrum of the ribose region (Fig. 2a and b). Thus, the detailed loop architecture remains subject to further structural investigation.

Assignment and data deposition

The nearly complete resonance assignment of SL5a builds on the imino resonance assignment published earlier (Wacker et al. 2020). Starting from this assignment, all 33 aromatic H6–C6 and H8–C8 correlations were unambiguously assigned. Furthermore, the H2–C2 correlations of the two adenosines present in this RNA as well as all of the H5–C5 correlations of the uridines and cytidines were unambiguously assigned. In addition, the quaternary carbon atoms of the nucleobases in purines (C2: 77%, C4: 69%, C5: 62% and C6: 92%) and pyrimidines (C2: 15% and C4: 15%) were partially assigned. Here, uridine C2 and C4 resonances as well as guanosine C2 and G$_{-1}$, G188, G198 and G208 C6 resonances were assigned at 283 K. Also, non-protonated tertiary nitrogen atoms of purines (N3: 15% (only

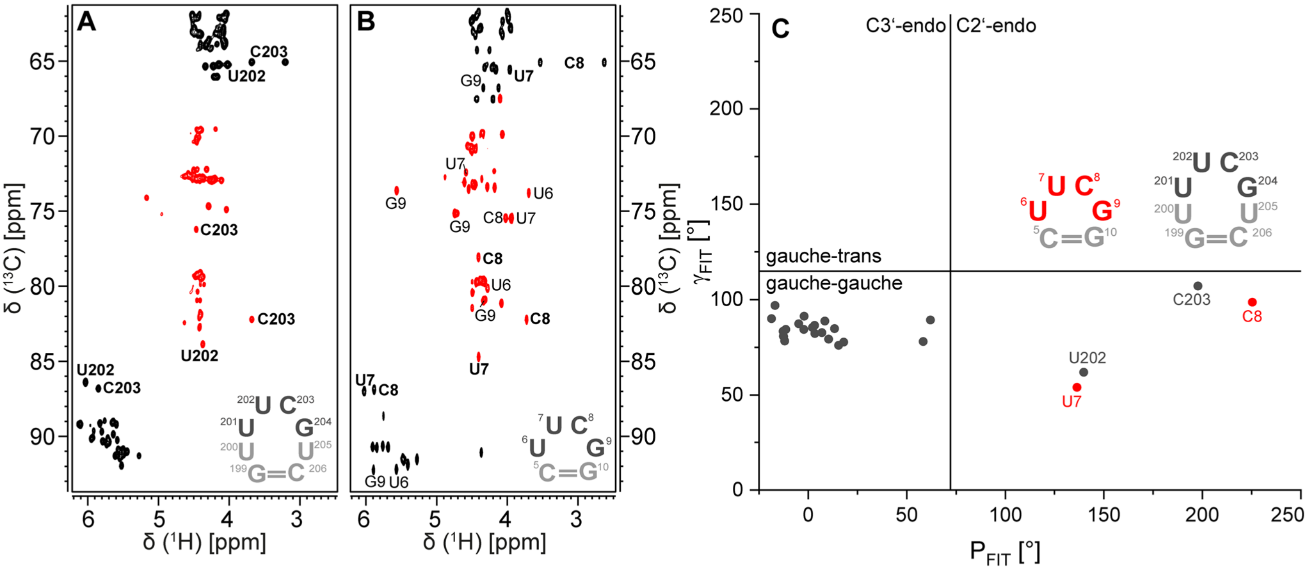


Fig. 2 Comparison of ^{1}H,^{13}C-CT-HSQC spectra of the ribose regions of **a** SL5a and **b** a 14 nt RNA with 5′-cUUCGg-3′ tetraloop (Fürtig et al. 2004; Nozinovic et al. 2010). Positive contours are given in black, negative contours in red. Experimental details are given in Table 1. The loop sequences are displayed and U202/U7 and C203/C8 resonances are highlighted in bold font. **c** Canonical coordinates for all residues of SL5a with a complete carbon ribose assignment. For comparison, the canonical coordinates of residues U7 and C8 of a 14 nt RNA with 5′-UUCG-3′ tetraloop are given in red

adenosines assigned), N7: 100% and N9: 100%) and pyrimidines (N1: 95% and N3: 80% (cytidines)) were successfully assigned to a large extent. Within the ribose moieties, 91% of the H1′ and 91% of the C1′ atoms were assigned. Within the remaining ribose carbon atoms C2′–C5′, 77% were assigned. In summary, we assigned 97% of the ^{1}H (H6/8, H5, H2, H1′) and 92% of the ^{13}C (C6/8, C5(pyr), C1′) atoms, which are considered most important for an in-depth structural characterization. We updated the BMRB deposition with code 50346.

Acknowledgements Work at BMRZ is supported by the state of Hesse. Work in Covid19-nmr was supported by the Goethe Corona Funds, by the IWB-EFRE-programme 20007375 of state of Hesse, the DFG in CRC902: "Molecular Principles of RNA-based regulation." and infrastructure funds (Project Numbers: 277478796, 277479031, 392682309, 452632086, 70653611). R.S. was recipient of a stipend of the Fonds der Chemischen Industrie. H.S. and B.F. are supported by the DFG in graduate school CLIC (GRK 1986). Work at KI is supported by the Karolinska Institute, Department of Medical Biochemistry and Biophysics for the purchase of a 600-MHz Bruker NMR spectrometer.

Funding Open Access funding enabled and organized by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Bodenhausen G, Ruben DJ (1980) Natural abundance Nitrogen-15-NMR by enhanced heteronuclear spectroscopy. Chem Phys Lett 69:185–189

Chen S-C, Olsthoorn RCL (2010) Group-specific structural features of the 5′-proximal sequences of coronavirus genomic RNAs. Virology 401:29–41. https://doi.org/10.1016/j.virol.2010.02.007

Cheong C, Varani G, Tinoco I (1990) Solution structure of an unusually stable RNA hairpin, 5GGAC(UUCG)GUCC. Nature 346:680–682

Cherepanov AV, Glaubitz C, Schwalbe H (2010) High-resolution studies of uniformly 13C,15N-labeled RNA by solid-state NMR spectroscopy. Angew Chem Int Ed 49:4747–4750. https://doi.org/10.1002/anie.200906885

Dingley AJ, Grzesiek S (1998) Direct observation of hydrogen bonds in nucleic acid base pairs by internucleotide 2JNN couplings. J Am Chem Soc 7863:714–718

Dingley AJ, Nisius L, Cordier F, Grzesiek S (2008) Direct detection of N–H[...]N hydrogen bonds in biomolecules by NMR spectroscopy. Nat Protoc 3:242–248. https://doi.org/10.1038/nprot.2007.497

Ebrahimi M, Rossi P, Rogers C, Harbison GS (2001) Dependence of 13 C NMR chemical shifts on conformations of RNA nucleosides and nucleotides. J Magn Reson 150:1–9. https://doi.org/10.1006/jmre.2001.2314

Favier A, Brutscher B (2011) Recovering lost magnetization: polarization enhancement in biomolecular NMR. J Biomol NMR 49:9–15. https://doi.org/10.1007/s10858-010-9461-5

Fürtig B, Richter C, Wöhnert J, Schwalbe H (2003) NMR spectroscopy of RNA. ChemBioChem 4:936–962. https://doi.org/10.1002/cbic.200300700

Fürtig B, Richter C, Bermel W, Schwalbe H (2004) New NMR experiments for RNA nucleobase resonance assignment and chemical shift analysis of an RNA UUCG tetraloop. J Biomol NMR 28:69–79. https://doi.org/10.1023/B:JNMR.0000012863.63522.1f

Guan B-J, Wu H-Y, Brian DA (2011) An optimal cis -replication stem-loop IV in the 5′ untranslated region of the mouse coronavirus genome extends 16 nucleotides into open reading frame 1. J Virol 85:5593–5605. https://doi.org/10.1128/JVI.00263-11

Guillerez J, Lopez PJ, Proux F, Launay H, Dreyfus M (2005) A mutation in T7 RNA polymerase that facilitates promoter clearance. Proc Natl Acad Sci USA 102:5958–5963. https://doi.org/10.1073/pnas.0407141102

Hu B, Guo H, Zhou P, Shi ZL (2020) Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol. https://doi.org/10.1038/s41579-020-00459-7

Kay LE, Xu G-Y, Singer AU, Muhandiram DR, Forman-Kay JD (1993) A gradient-enhanced HCCH-TOCSY experiment for recording side-chain 1H and 13C correlations in H2O samples of proteins. J Magn Reson Ser B 101:333–337

Kelly JA, Olson AN, Neupane K, Munshi S, Emeterio JS, Pollack L, Woodside MT, Dinman JD (2020) Structural and functional conservation of the programmed −1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). J Biol Chem 295:10741–10748. https://doi.org/10.1074/jbc.AC120.013449

Lee W, Tonelli M, Markley JL (2015) NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics 31:1325–1327. https://doi.org/10.1093/bioinformatics/btu830

Madhugiri R, Fricke M, Marz M, Ziebuhr J (2016) Coronavirus cis-acting RNA elements. Adv Virus Res 96:127–163

Manfredonia I, Nithin C, Ponce-Salvatierra A, Ghosh P, Wirecki TK, Marinus T, Ogando NS, Snijder EJ, Van HMJ, Bujnicki JM, Incarnato D (2020) Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. Nucleic Acids Res. https://doi.org/10.1093/nar/gkaa1053

Martin S, Blankenship C, Rausch JW, Sztuba-Solinska J (2019) Using SHAPE-MaP to probe small molecule-RNA interactions. Methods 167:105–116. https://doi.org/10.1016/j.ymeth.2019.04.009

Masters PS (2019) Coronavirus genomic RNA packaging. Virology 537:198–207

Nozinovic S, Fürtig B, Jonker HRA, Richter C, Schwalbe H (2010) High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. Nucleic Acids Res 38:683–694. https://doi.org/10.1093/nar/gkp956

Ohlenschläger O, Wöhnert J, Bucci E, Seitz S, Häfner S, Ramachandran R, Zell R, Görlach M (2004) The structure of the stemloop D subdomain of coxsackievirus B3 cloverleaf RNA and its interaction with the proteinase 3C. Structure 12:237–248. https://doi.org/10.1016/j.str.2004.01.014

Piotto M, Saudek V, Sklen V (1992) Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. J Biomol NMR 2:661–665

Rangan R, Zheludev IN, Das R (2020) RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses. RNA 26:937–959

Richter C, Kovacs H, Buck J, Wacker A, Fürtig B, Bermel W, Schwalbe H (2010) 13C-direct detected NMR experiments for the sequential J-based resonance assignment of RNA oligonucleotides. J Biomol NMR 47:259–269. https://doi.org/10.1007/s10858-010-9429-5

Schürer H, Lang K, Schuster J, Mörl M (2002) A universal method to produce in vitro transcripts with homogeneous 3′ ends. Nucleic Acids Res 30:e56

Shaka AJ, Hwang T-L (1995) Water suppression that works. Excitation sculpting using arbitrary waveforms and pulsed field gradients. J Magn Reson Ser A 112:275–279

Shaka AJ, Lee CJ, Pines A (1988) Iterative schemes for bilinear operators; application to spin decoupling. J Magn Reson 77:274–293

Simon B, Zanier K, Sattler M (2001) A TROSY relayed HCCH-COSY experiment for correlating adenine H2/H8 resonances in uniformly 13C-labeled RNA molecules. J Biomol NMR 20:173–176

Sklenar V, Peterson RD, Rejante MR, Feigon J (1993a) Two- and three-dimensional HCN experiments for correlating base and sugar resonances in lSN,13C-labeled RNA oligonucleotides. J Biomol NMR 3:721–727

Sklenar V, Piotto M, Leppik R, Saudek V (1993b) Gradient-tailored water suppression for 1H-15N HSQC experiments optimized to retain full sensitivity. J Magn Reson Ser A 102:241–245

Sklenár V, Dieckmann T, Butcher SE, Feigon J (1996) Through-bond correlation of imino and aromatic resonances in 13C-, 15N-labeled RNA via heteronuclear TOCSY. J Biomol NMR 7:83–87

Solyom Z, Schwarten M, Geist L, Konrat R, Willbold D, Brutscher B (2013) BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. J Biomol NMR 55:311–321. https://doi.org/10.1007/s10858-013-9715-0

Tidu A, Janvier A, Schaeffer L, Sosnowski P, Kuhn L, Hammann P, Westhof E, Eriani G, Martin F (2020) The viral protein NSP1 acts as a ribosome gatekeeper for shutting down host translation and fostering SARS-CoV-2 translation. RNA. https://doi.org/10.1261/rna.078121.120

V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V (2020) Coronavirus biology and replication: implications for SARS-CoV-2. Nat Rev Microbiol. https://doi.org/10.1038/s41579-020-00468-6

Vuister W, Bax AD (1992) Resolution enhancement and spectral editing of uniformly 13C-enriched proteins by homonuclear broadband 13C decoupling. J Magn Reson 98:428–435

Wacker A, Weigand JE, Akabayov SR, Altincekic N, Bains K, Banijamali E, Binas O, Castillo-martinez J, Cetiner E, Chiu L, Davila-Calderon J, Dhamotharan K, Duchardt-Ferner E, Ferner J, Frydman L, Fürtig B, Hacker C, Haddad C, Haehnke M, Hengesbach M, Hiller F, Hohmann KF, Hymon D, De JV, Jonker H, Luo L, Mertinkus KR, Keller H, Knezic B, Landgraf T, Löhr F, Muhs C, Novakovic M, Oxenfarth A, Palomino-Schätzlein M, Petzold K, Peter SA, Pyper DJ, Qureshi NS, Riad M, Richter C, Saxena K, Schamber T, Scherf T, Schlagnitweit J, Schlundt A, Schnieders R, Schwalbe H, Simba-lahuasi A, Sreeramulu S, Stirnal E, Sudakov A, Tants J-N, Tolbert BS, Vögele J, Weiß L, Wirmer-Bartoschek J, Wirz Martin MA, Wöhnert J, Zetzsche H (2020) Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy. Nucleic Acids Res. https://doi.org/10.1093/nar/gkaa1013

Wishart DS, Bigam CG, Yao J, Abildgaard F, Jane HD, Oldfield E, Markley JL, Sykes BD (1995) Chemical shift referencing in biomolecular NMR. J Biomol NMR 6:135–140

Wöhnert J, Dingley AJ, Stoldt M, Görlach M, Grzesiek S, Brown LR (1999) Direct identification of NH ··· N hydrogen bonds in non-canonical base pairs of RNA by NMR spectroscopy. Nucleic Acids Res 27:3104–3110

Wöhnert J, Görlach M, Schwalbe H (2003) Triple resonance experiments for the simultaneous correlation of H6/H5 and exchangeable protons of pyrimidine nucleotides in 13C,15N-labeled RNA applicable to larger RNA molecules. J Biomol NMR 26:79–83

5.11 **Research article:** ^{1}H, ^{13}C, and ^{15}N backbone chemical shift assignments of coronavirus-2 non-structural protein Nsp10

Nina Kubatova*, Nusrat S. Qureshi*, Nadide Altincekic*, Rupert Abele, Jasleen Kaur Bains, Betül Ceylan, Jan Ferner, Christin Fuks, Bruno Hargittay, Marie T. Hutchison, Vanessa de Jesus, Felicitas Kutz, Maria A. Wirtz Martin, Nathalie Meiser, Verena Linhard, Dennis J. Pyper, Sven Trucks, Boris Fürtig, Martin Hengesbach, Frank Löhr, Christian Richter, Krishna Saxena, Andreas Schlundt, Harald Schwalbe, Sridhar Sreeramulu, Anna Wacker, Julia E. Weigand, Julia Wirmer-Bartoschek and Jens Wöhnert; *Biomolecular NMR Assignments* **15**, 65-71 (2021)

* These authors contributed equally to this work.

In this article, the non-structural protein 10 (Nsp10) involved in the viral replication-transcription complex of the SARS-CoV-2 genome was characterized by ^{1}H, ^{13}C, and ^{15}N backbone chemical shift assignments. Nsp10 was described as a part of the replicase complex that regulates the exonuclease activity of Nsp14 and the methyltransferase activity of Nsp16. Thus, these results provide a basis for further NMR investigations as well as ligand-binding studies with Nsp10.

The project was designed and coordinated by the Covid-19 NMR consortium. The manuscript was prepared by N. Kubatova, N. S. Qureshi and N. Altincekic. The author of this thesis was part of the Covid-19 NMR consortium and the protein production team that prepared and purified various isotope-labeled protein samples for NMR spectroscopy.

https://doi.org/10.1007/s12104-020-09973-4

Biomolecular NMR Assignments
https://doi.org/10.1007/s12104-020-09984-1

ARTICLE

^{1}H, ^{13}C, and ^{15}N backbone chemical shift assignments of coronavirus-2 non-structural protein Nsp10

N. Kubatova[1] · N. S. Qureshi[1] · N. Altincekic[1] · R. Abele[2] · J. K. Bains[1] · B. Ceylan[1] · J. Ferner[1] · C. Fuks[1] · B. Hargittay[1] · M. T. Hutchison[1] · V. de Jesus[1] · F. Kutz[1] · M. A. Wirtz Martin[1] · N. Meiser[1] · V. Linhard[1] · D. J. Pyper[1] · S. Trucks[1] · B. Fürtig[1] · M. Hengesbach[1] · F. Löhr[4] · C. Richter[1] · K. Saxena[1] · A. Schlundt[3] · H. Schwalbe[1] · S. Sreeramulu[1] · A. Wacker[1] · J. E. Weigand[5] · J. Wirmer-Bartoschek[1] · J. Wöhnert[3]

Received: 7 August 2020 / Accepted: 3 November 2020

Abstract

The international Covid19-NMR consortium aims at the comprehensive spectroscopic characterization of SARS-CoV-2 RNA elements and proteins and will provide NMR chemical shift assignments of the molecular components of this virus. The SARS-CoV-2 genome encodes approximately 30 different proteins. Four of these proteins are involved in forming the viral envelope or in the packaging of the RNA genome and are therefore called structural proteins. The other proteins fulfill a variety of functions during the viral life cycle and comprise the so-called non-structural proteins (nsps). Here, we report the near-complete NMR resonance assignment for the backbone chemical shifts of the non-structural protein 10 (nsp10). Nsp10 is part of the viral replication-transcription complex (RTC). It aids in synthesizing and modifying the genomic and subgenomic RNAs. Via its interaction with nsp14, it ensures transcriptional fidelity of the RNA-dependent RNA polymerase, and through its stimulation of the methyltransferase activity of nsp16, it aids in synthesizing the RNA cap structures which protect the viral RNAs from being recognized by the innate immune system. Both of these functions can be potentially targeted by drugs. Our data will aid in performing additional NMR-based characterizations, and provide a basis for the identification of possible small molecule ligands interfering with nsp10 exerting its essential role in viral replication.

Keywords SARS-CoV-2 · Non-structural protein · Solution NMR-spectroscopy · Covid19-NMR

Biological context

The current worldwide COVID-19 pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has severely impacted nearly every area of human life. In the absence of reliable therapeutic options or vaccination, research efforts have been increased in order to understand molecular characteristics of the functional components of the virus. SARS-CoV-2 belongs to the family of *Coronaviridae*, whose members share a distinct pattern in their genomic organization (Lai 1990; Snijder et al. 2016). In contrast to several other viruses, the large (+)-strand RNA

N. Kubatova, N. S. Qureshi and N. Altincekic are joint first authors.

✉ M. Hengesbach
hengesbach@nmr.uni-frankfurt.de

✉ H. Schwalbe
schwalbe@nmr.uni-frankfurt.de;
covid19-nmr@dlist.server.uni-frankfurt.de

[1] Institute for Organic Chemistry and Chemical Biology, Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue-Str. 7, 60438 Frankfurt/M, Germany

[2] Institute for Biochemistry, Biocentre, Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue-Str. 7, 60438 Frankfurt/M, Germany

[3] Institute for Molecular Biosciences, Johann Wolfgang Goethe-University, Max-von-Laue-Str. 7, 60438 Frankfurt/M, Germany

[4] Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance (BMRZ), Johann Wolfgang Goethe-University, Max-von-Laue-Str. 7, 60438 Frankfurt/M, Germany

[5] Department of Biology, Technical University of Darmstadt, Schnittspahnstr 10, 64287 Darmstadt, Germany

Published online: 07 November 2020

genome of ~ 30 kb requires high fidelity genome replication while offering coding space for additional proteins.

The RNA genome of SARS-CoV-2 encodes at least 14 polypeptides, some of which are cleaved by viral proteases into their functional forms. The non-structural protein 10 (nsp10) is generated from the first open reading frame of the virus (ORF1) through cleavage by the main protease nsp5 to yield a 139 amino acid protein of 14.8 kDa.

Previous work has been reported on the homologous proteins of nsp10 from SARS-CoV (Bouvet et al. 2014), middle eastern respiratory syndrome coronavirus (MERS-CoV) (Aouadi et al. 2017) and murine hepatitis virus (MHV) (Matthes et al. 2006).

Nsp10 plays numerous roles in coronavirus replication. As part of the replicase complex, it interacts with nsp14 (Minskaia et al. 2006) to stimulate the exonuclease activity of nsp14 (Ferron et al. 2017) and thus contributes to the enhanced replication fidelity of coronaviruses in comparison to other RNA viruses. It also increases the catalytic activity of nsp16 (Bouvet et al. 2010), which methylates the 2′-hydroxyl group of the +1 adenosine of genomic and subgenomic (+)-strand RNAs. The presence of this modification prevents the recognition and degradation of viral RNAs by the innate immune system of the host (Daffis et al. 2010; Züst et al. 2011). Additional regulatory functions for nsp10, such as interactions with nsp1 or nsp7, have been proposed (Brockway et al. 2004), but these functionalities require additional characterization. In summary, the functional diversity of nsp10 during the viral life cycle renders it a promising drug target (Wang et al. 2015).

Several structures of nsp10 from SARS-CoV have previously been solved both in isolation (Joseph et al. 2006; Su et al. 2006) and in complex with either nsp14 (Bouvet et al. 2012; Ma et al. 2015) or nsp16 (Chen et al. 2011; Decroly et al. 2011). The structures showed the presence of a completely novel fold including two unusual zinc finger motifs. The first zinc finger includes residues C74, C77, H81 and C90 (SARS-CoV numbering) and is loosely characterized as a Zn^{2+}-binding loop. The second zinc-binding motif involving C117, C120, C128, and C130 can be classified as a "gag-knuckle" type motif (Krishna et al. 2003). All of these residues can also be found in the nsp10 sequence of SARS-CoV-2 (Krafcikova et al. 2020; Viswanathan et al. 2020). Here, we provide a near complete assignment of nsp10 backbone NMR resonances. Due to the known interaction sites with other proteins, this assignment will aid further NMR-based structural investigations as well as ligand binding studies.

Methods and experiments

Construct design

The amino acid sequence of SARS-CoV-2 nsp10 was obtained from the NCBI reference genome entry NC_045512.2, identical to GenBank entry MN908947.3 (Wu et al. 2020). The gene was codon-optimized for expression in *E. coli*, commercially synthesized and sub-cloned into a pET21b(+) vector (*Genscript*), carrying an additional sequence at the N-terminus (MGSDKIHHHHHH) including a hexa-histidine (His_6)-tag for purification.

Sample preparation

The pET21b(+) plasmid containing the nsp10 sequence was transformed into T7-Express *E. coli* cells. Nsp10 was heterologously expressed with an N-terminal His_6-tag (Joseph et al. 2006). Uniformly ^{13}C,^{15}N-labeled nsp10 protein was expressed in M9 minimal medium containing 1 g/L $^{15}NH_4Cl$, 2 g/L $^{13}C_6$-D-glucose and 100 mg/L ampicillin. After the OD_{600} reached a value between 0.6 and 0.7, the culture was induced with 0.5 mM isopropyl-β-D-thiogalactoside (IPTG) and supplemented with 50 µM $ZnCl_2$. The final preparative expression was performed at 20 °C, 120 rpm for 12 h. Cells from a 2 L culture were harvested at 4000 g for 15 min using a Beckmann centrifuge with a JLA 8.1000 rotor. The pellet was resuspended in 100 mL of buffer A (25 mM Tris pH 8, 300 mM NaCl, 5 mM imidazole, 10 mM β-mercaptoethanol) and supplemented with 50 µM $ZnCl_2$, and two protease-inhibitor tablets (*cOmplete™, Roche, Germany*). Cells were mechanically lysed using Microfluidics M-110P at 15,000 PSI (pounds per square inch) under continuous ice cooling, followed by centrifugation at 4 °C and 38,400×*g* for 45 min using a Beckmann centrifuge with a JA 20 rotor.

The supernatant was further purified using immobilized metal ion affinity chromatography (IMAC) followed by size exclusion chromatography (SEC). After centrifugation, the supernatant was loaded onto a 5 mL HisTrap HP column (*GE Healthcare, USA*) connected to an FPLC system (*Äktapurifier™, GE Healthcare, USA*). Elution of the bound protein was achieved with a linear gradient of 500 mM imidazole containing buffer A. Protein-containing fractions were combined, concentrated using centrifugal concentrator devices (*VivaSpin20, Sartorius, Germany, MWCO 10000*) and further purified with a 320 mL HiLoad 26/600 Superdex 75 pg gel filtration column (*GE Healthcare, USA*) using 50 mM sodium phosphate buffer (pH 7.5), 50 mM NaCl, 5 mM dithiothreitol (DTT). Purity of the produced protein was verified by SDS-PAGE analysis and confirmed using mass spectrometry (MALDI). Finally, the protein was concentrated using centrifugal concentrator devices (*VivaSpin20*, *Sartorius, Germany, MWCO 10000*).

SEC-MALS analysis was performed using a Superdex 75, 10/300 GL column at a flow rate of 0.5 mL/min, using a Wyatt miniDAWN TREOS with a 658 nm laser. Refractive index was monitored using a Wyatt Optilab rEX.

NMR experiments

The protein samples were measured in NMR buffer containing 50 mM phosphate buffer pH 7.5, 50 mM NaCl, 5 mM DTT, 95% H_2O/5% D_2O. Spectra were recorded at 298 K on Bruker spectrometers ranging from 600 to 950 MHz equipped with z-axis gradient ^{1}H{^{13}C,^{15}N} triple resonance cryogenic probes and on a Bruker 500 MHz spectrometer equipped with a room-temperature triple-resonance probe. The spectrometer was locked on D_2O. ^{1}H chemical shifts were referenced to DSS at 0.00 ppm and ^{13}C,^{15}N chemical shifts were calculated relative to the ^{1}H frequency according to (Wishart 2011). All NMR spectra were processed by using TopSpin version 3.2 (Bruker Biospin) and analyzed and visualized with SPARKY version 3.114 (Lee et al. 2015). Parameters of the NMR experiments and spectrometers used in this study are listed in Table 1.

For the sequential backbone resonances assignment, a set of 3D NMR experiments including BEST-TROSY (Farjon et al. 2009; Solyom et al. 2013) based HNCACB, HN(CO)CACB, HNCO, HN(CA)CO experiments were used.

For the temperature series, 2D ^{1}H,^{15}N BEST-TROSY spectra were measured from 293 to 308 K with 5 K increments on a 600 MHz spectrometer. Temperature coefficients of the amide protons were calculated from a linear fit of a chemical shift perturbation as a function of a temperature (Baxter and Williamson 1997).

Heteronuclear ^{15}N relaxation experiments ({^{1}H}-^{15}N hetNOE, T_1 and T_2) were performed at 600 MHz and 298 K using a 0.15 mM and a 1.6 mM sample. For determining the longitudinal T_1 ^{15}N relaxation time, a series of spectra with the following relaxation delays was used: 20, 60, 100, 200, 400, 600, 800, 1200, 1400, 1600, 1800 ms. The transverse ^{15}N relaxation time T_2 was determined from spectra with the following relaxation delays: 16.96, 33.92, 67.84, 101.76, 135.68, 169.60, 203.52, 271.36 ms. The {^{1}H}-^{15}N hetNOEs were calculated from the signal intensity ratio (I_{on}/I_{off}) obtained from spectra recorded with and without saturation of amide protons with a recovery (d_1) and saturation delay of 10 s.

TRACT experiments (Lee et al. 2006) were carried out at 500 MHz to determine rotational correlation times at four different protein concentrations (1.2 mM, 0.6 mM, 0.3 mM and 0.15 mM). A two-dimensional BEST version (Rennella and Brutscher 2013) was employed to avoid an underestimation of the correlation time by the contribution of intense signals from mobile residues. A total of 21 well-resolved cross peaks from structured regions of the protein were chosen for evaluation.

Assignments and data deposition

The backbone chemical shift assignment of nsp10 protein was conducted manually by using standard double- and triple-resonance NMR experiments on uniformly labeled samples at 298 K. With the heteronuclear 3D experiments listed above we could assign 89% of the ^{1}H,^{15}N amide signals and 94% of the total backbone assignment (C_α—93%, C_β—95%, C′—92%). When excluding the first twelve N-terminal residues including the His_6-tag, the backbone assignment comprises 92% of the ^{1}H,^{15}N pairs and 95.5% of the total resonances (C_α—96%, C_β—95%, C′—96%). Residues M1, G2, F28, C29, T59 H60 and the N-terminal His_6-tag were not assigned. Assignment of the residues K5 and W135 is tentative.

Table 1 List of experiments collected to perform the sequence specific assignment of nsp10. Main parameters used are reported

Experiments	Time domain data size (points)			Spectral width (ppm)			Number of scans	Delay time (s)	NUS %
	T_1	T_2	T_3	F_1	F_2	F_3			
^{1}H,^{15}N BEST-TROSY	1024	1272		50 (^{15}N)	10.0 (^{1}H)		4	0.3	
BEST-TROSY-HN(CO)CACB	208	256	1272	59.7 (^{13}C)	26.0 (^{15}N)	10.0 (^{1}H)	8	0.3	
BEST-TROSY-HNCACB	216	256	1272	59.7 (^{13}C)	26.0 (^{15}N)	10.0 (^{1}H)	8	0.3	
BEST-TROSY-HN(CA)CO	216	256	1272	12.0 (^{13}C)	26.0 (^{15}N)	10.0 (^{1}H)	8	0.3	25
BEST-TROSY-HNCO	216	256	1272	12.0 (^{13}C)	26.0 (^{15}N)	10.0 (^{1}H)	4	0.3	
^{15}N R_1	12	256	2048	–	28.0 (^{15}N)	16.0 (^{1}H)	8	2.0	
^{15}N R_2	12	256	2048	–	28.0 (^{15}N)	16.0 (^{1}H)	8	2.0	
^{15}N-NOE	2	256	2048	–	28.0 (^{15}N)	16.0 (^{1}H)	80	10	
TRACT	28	96	896	–	30.0 (^{15}N)	12.0 (^{1}H)	8–40	0.3	

The number of signals observed in BEST-TROSY experiments is approximately 5% higher than expected for the nsp10 protein. Lower intensities for a subset of the signals are indicative of the presence of a second minor conformation in slow exchange on the NMR time scale. The resonances in question were assigned to the last seven C-terminal amino acids (Fig. 1). The backbone resonance assignment for the main conformation was deposited in the biological magnetic resonance bank (BMRB ID: 50392).

Study of nsp10 dynamics

We also characterized the dynamics of nsp10 by performing {^{1}H}^{15}N heteronuclear NOE experiments for two sample concentrations (1.6 mM and 0.15 mM) and analyzing intramolecular hydrogen bonds using temperature factors (Fig. 2). The last nine C-terminal residues from the major conformation show low heteronuclear NOEs values, indicating the high flexibility for this region (Fig. 2b).

Intramolecular hydrogen bonding was analyzed by monitoring the amide protons chemical shift changes as a function of temperature. Thus, a series of 2D ^{1}H,^{15}N BEST-TROSY spectra were recorded for a temperature range from 293 to 308 K in 5 K increments. Temperature coefficients calculated from a linear fit of the chemical shift changes with values below—4.5 ppb/K are indicative for fast exchangeable, not hydrogen-bonded amide protons, and temperature coefficients higher than—4.5 ppb/K are characteristic for the involvement of this amide resonance in hydrogen bonding (Baxter and Williamson 1997). The temperature coefficients of the C-terminus of nsp10 are consistent with the relaxation data, supporting a dynamic nature of this part of the protein. The amide group of residue A32 located in the N-terminal region shows an extremely high temperature coefficient with a positive value of 6.3 ± 0.7 ppb/K, the reason of which is not clear (Fig. 2e, f). Furthermore, we performed TRACT measurements (Lee et al. 2006) at different concentrations and determined concentration-dependent oligomerization (Fig. 3a). Experimental τ_c for the low concentration sample (0.15 mM) is 10.7 ns, which is agreement with the monomer value predicted by HydroNMR (10.8 ns) (de la Torre et al. 2000). Analysis of the oligomerization status by size exclusion chromatography coupled to multiple angle light scattering (SEC-MALS) (Fig. 3b) shows that a minor fraction of proteins forms higher order structures, whose mass corresponds to a protein dimer.

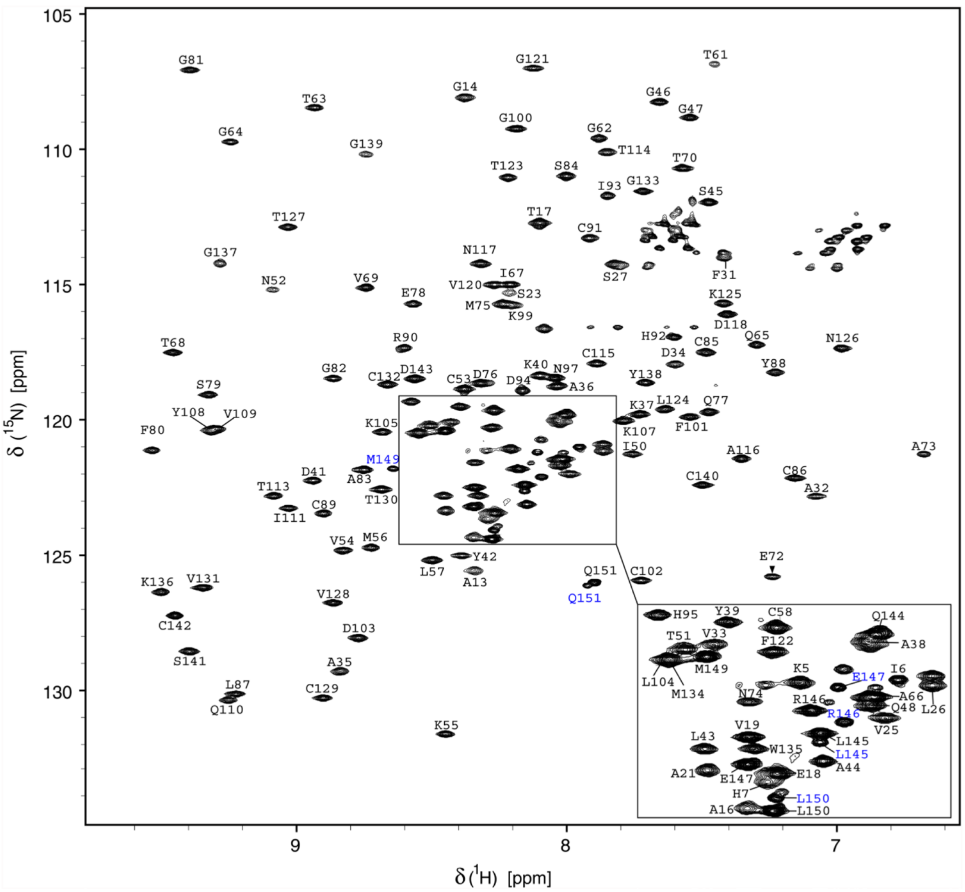


Fig. 1 ^{1}H,^{15}N-BEST-TROSY spectrum of ^{13}C,^{15}N-labeled SARS-CoV-2 nsp10 (0.4 mM) in 50 mM phosphate buffer pH 7.5, 50 mM NaCl, 5 mM DTT, and 5% D_2O measured at 298 K on a 950 MHz spectrometer. Backbone resonance assignment labels are given. The minor conformation adopted by the nine C-terminal residues is highlighted in blue

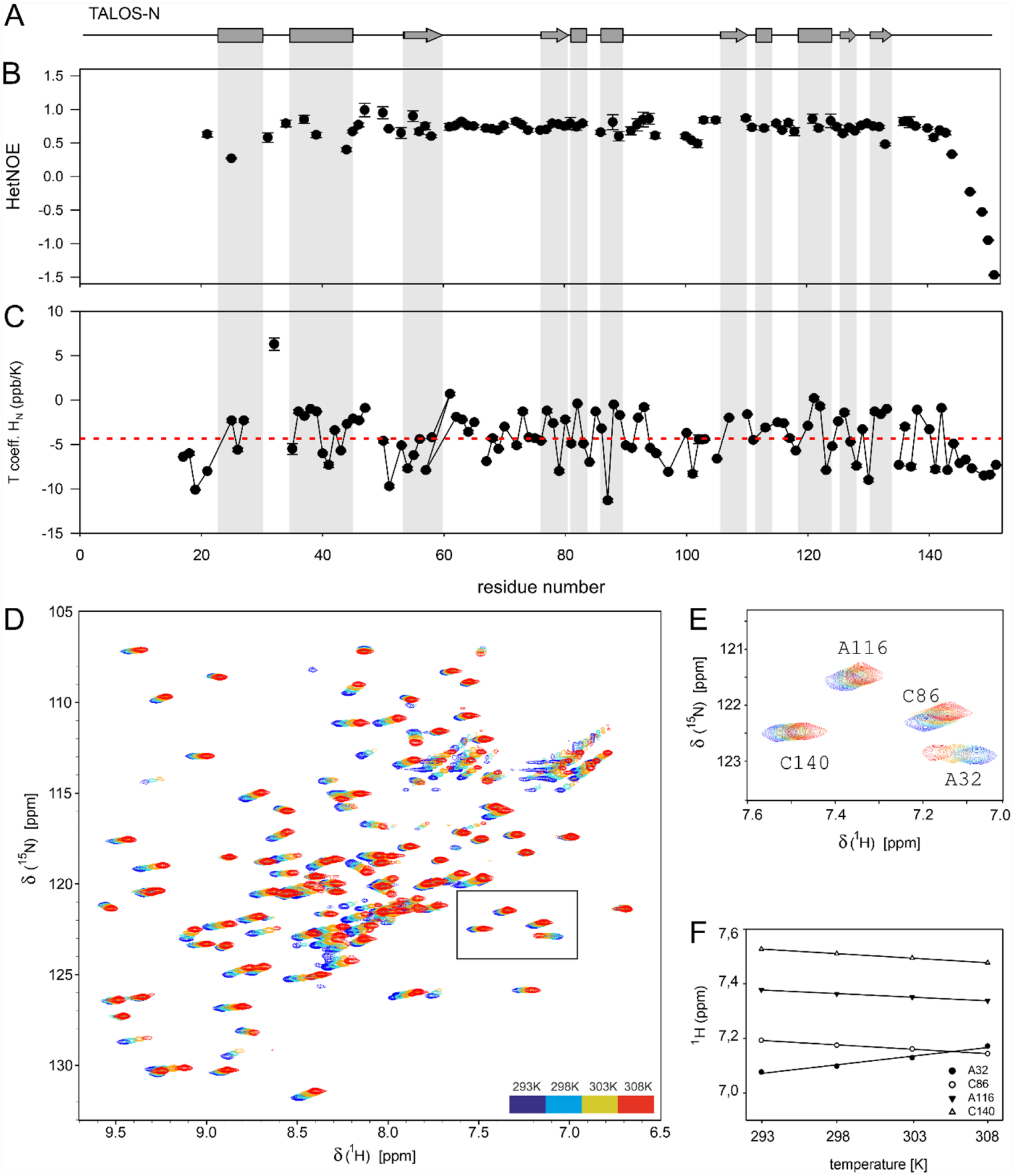


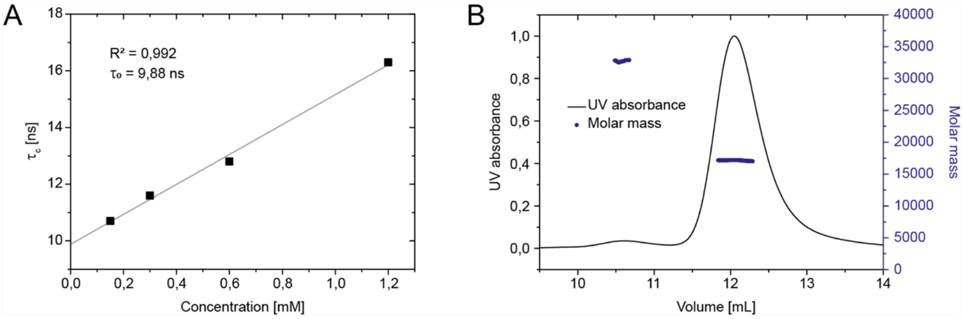
Fig. 2 Backbone dynamics of nsp10. **a** Schematic representation of an NMR chemical-shift-based TALOS-N secondary-structure prediction of nsp10. **b** {1H}-^{15}N heteronuclear NOEs measured at 298 K and 0.15 mM nsp10 are shown as a function of residue number. **c** The temperature coefficients (T_{coeff}) determined from a series of 2D ^{1}H,^{15}N BEST-TROSY spectra measured from 293 to 308 K with 5 K increments at 600 MHz. **d** Overlay of 2D ^{1}H,^{15}N BEST-TROSY spectra recorded at 600 MHz at different temperatures (color code is shown in the figure). **e** Selected zoom-in showing residues A32, C86, A116 and C140, representing different hydrogen bond strengths. **f** Plot of the amide proton chemical shift (in ppm) as a function of temperature (in K)

Structural comparison of nsp10 from SARS-CoV and SARS-CoV-2

The backbone resonance assignment was used to predict secondary structure elements of nsp10 using TALOS-N (Shen and Bax 2013). Six α-helixes and five β-strands were identified. Experimentally determined structural elements were compared with motifs obtained from the X-ray structure of SARS-CoV nsp10 (PDB: 6WQ3) (Fig. 4).

Overall, the structural elements of both proteins are similar. Differences can be detected in the location of a

Fig. 3 Oligomerization analysis. **a** TRACT analysis. Linear fit of τ_c against protein concentration from data points at 0.15, 0.3, 0.6, and 1.2 mM nsp10. **b** SEC-MALS analysis performed at 1.2 mM nsp10 shows a minor amount of oligomerization, which can be assigned to the nsp10 dimer by mass determination



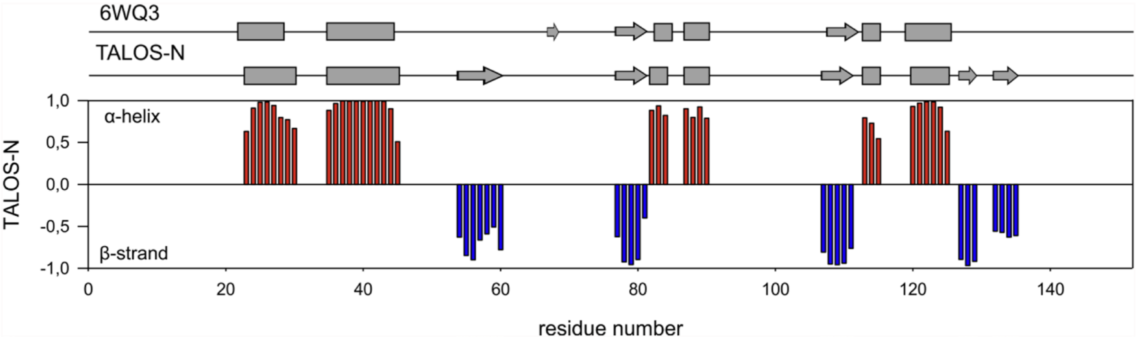


Fig. 4 Schematic representation of an NMR chemical-shift-based TALOS-N secondary-structure prediction of nsp10 and its comparison with the secondary structure elements obtained from the X-ray structure extracted from PDB entry 6WQ3, which shows 100% sequence identity between residues S23 and Q144

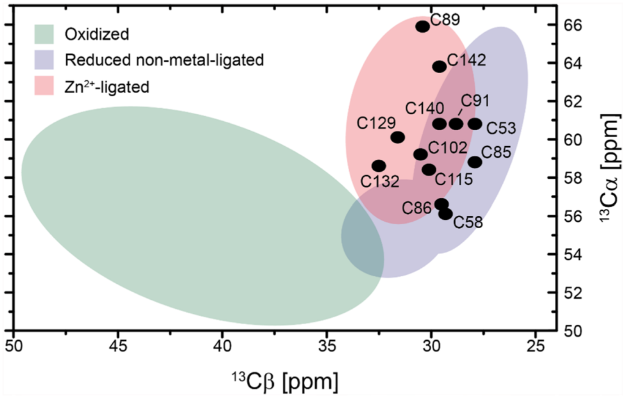


Fig. 5 ^{13}Cα and ^{13}Cβ chemical shifts of the assigned nsp10 cysteines. The regions for different chemical states of the cysteines (colored areas) are derived from (Kornhaber et al. 2006)

stretch of amino acids that adopts a β-strand conformation according to the TALOS-N analysis (aa 54–60), which is directly involved in interactions with the complex partner nsp16 in the crystal structure (PDB 6WQ3). A second difference can be observed for some of the C-terminal amino acids, which adopt a zinc finger fold, but are classified as ß-strands in the TALOS prediction. However, all of the assigned cysteine residues show Cβ chemical shifts that are indicative of a Zn^{2+}-ligated form (Fig. 5) (Kornhaber et al. 2006).

Acknowledgements Work at BMRZ is supported by the state of Hesse. Work in Covid19-NMR was supported by the Goethe Corona Funds, and by the DFG within the SFB902.

Funding Open Access funding enabled and organized by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

Aouadi W, Blanjoie A, Vasseur JJ, Debart F, Canard B, Decroly E (2017) Binding of the methyl donor S-adenosyl-L-methionine to middle east respiratory syndrome coronavirus 2'-O-methyltransferase nsp16 promotes recruitment of the allosteric activator nsp10. J Virol. https://doi.org/10.1128/jvi.02217-16

Baxter NJ, Williamson MP (1997) Temperature dependence of 1H chemical shifts in proteins. J Biomol NMR 9:359–369. https://doi.org/10.1023/a:1018334207887

Bouvet M, Debarnot C, Imbert I, Selisko B, Snijder EJ, Canard B, Decroly E (2010) In vitro reconstitution of sars-coronavirus mRNA cap methylation. PLoS Pathog 6:1–13. https://doi.org/10.1371/journal.ppat.1000863

Bouvet M, Imbert I, Subissi L, Gluais L, Canard B, Decroly E (2012) RNA 3′-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. Proc Natl Acad Sci USA 109:9372–9377. https://doi.org/10.1073/pnas.1201130109

Bouvet M et al (2014) Coronavirus Nsp10, a critical co-factor for activation of multiple replicative enzymes. J Biol Chem 289:25783–25796. https://doi.org/10.1074/jbc.m114.577353

Brockway SM, Lu XT, Peters TR, Dermody TS, Denison MR (2004) Intracellular localization and protein interactions of the gene 1 protein p28 during mouse hepatitis virus replication. J Virol 78:11551–11562. https://doi.org/10.1128/jvi.78.21.11551-11562.2004

Chen Y et al (2011) Biochemical and structural insights into the mechanisms of sars coronavirus RNA ribose 2′-O-methylation by nsp16/nsp10 protein complex. PLoS Pathog. https://doi.org/10.1371/journal.ppat.1002294

Daffis S et al (2010) 2′-O methylation of the viral mRNA cap evades host restriction by IFIT family members. Nature 468:452–456. https://doi.org/10.1038/nature09489

Decroly E et al (2011) Crystal structure and functional analysis of the SARS-coronavirus RNA cap 2′-o-methyltransferase nsp10/nsp16 complex. PLoS Pathog. https://doi.org/10.1371/journal.ppat.1002059

de la Torre GJ, Huertas ML, Carrasco B (2000) HYDRONMR: prediction of NMR relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. J Magn Reson 147(1):138–146. https://doi.org/10.1006/jmre.2000.2170

Farjon J, Boisbouvier JR, Schanda P, Pardi A, Simorre JP, Brutscher B (2009) Longitudinal-relaxation-enhanced NMR experiments for the study of nucleic acids in solution. J Am Chem Soc 131:8571–8577. https://doi.org/10.1021/ja901633y

Ferron F et al (2017) Structural and molecular basis of mismatch correction and ribavirin excision from coronavirus RNA. Proc Natl Acad Sci USA 115:E162–E171. https://doi.org/10.1073/pnas.1718806115

Joseph JS et al (2006) Crystal structure of nonstructural protein 10 from the severe acute respiratory syndrome coronavirus reveals a novel fold with two zinc-binding motifs. J Virol 80:7894–7901. https://doi.org/10.1128/jvi.00467-06

Kornhaber GJ, Snyder D, Moseley HN, Montelione GT (2006) Identification of zinc-ligated cysteine residues based on 13Calpha and 13Cbeta chemical shift data. J Biomol NMR 34:259–269. https://doi.org/10.1007/s10858-006-0027-5

Krafcikova P, Silhan J, Nencka R, Boura E (2020) Structural analysis of the SARS-CoV-2 methyltransferase complex involved in RNA cap creation bound to sinefungin. Nat Commun 11:1–7. https://doi.org/10.1038/s41467-020-17495-9

Krishna SS, Majumdar I, Grishin NV (2003) Structural classification of zinc fingers: survey and summary. Nucleic Acids Res 31:532–550. https://doi.org/10.1093/nar/gkg161

Lai MM (1990) Coronavirus: organization, replication and expression of genome. Annu Rev Microbiol 44:303–333. https://doi.org/10.1146/annurev.mi.44.100190.001511

Lee D, Hilty C, Wider G, Wüthrich K (2006) Effective rotational correlation times of proteins from NMR relaxation interference. J Magn Reson 178:72–76

Lee W, Tonelli M, Markley JL (2015) NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics 31:1325–1327. https://doi.org/10.1093/bioinformatics/btu830

Ma Y et al (2015) Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. Proc Natl Acad Sci USA 112:9436–9441. https://doi.org/10.1073/pnas.1508686112

Matthes N, Mesters JR, Coutard B, Canard B, Snijder EJ, Moll R, Hilgenfeld R (2006) The non-structural protein Nsp10 of mouse hepatitis virus binds zinc ions and nucleic acids. FEBS Lett 580:4143–4149. https://doi.org/10.1016/j.febslet.2006.06.061

Minskaia E, Hertzig T, Gorbalenya AE, Campanacci V, Cambillau C, Canard B, Ziebuhr J (2006) Discovery of an RNA virus 3′→5′ exoribonuclease that is critically involved in coronavirus RNA synthesis. Proc Natl Acad Sci USA 103:5108–5113. https://doi.org/10.1073/pnas.0508200103

Rennella E, Brutscher B (2013) Fast real-time NMR methods for characterizing short-lived molecular states. ChemPhysChem 14:3059–3070. https://doi.org/10.1002/cphc.201300339

Shen Y, Bax AD (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. J Biomol NMR 56(3):227–241. https://doi.org/10.1007/s10858-013-9741-y

Snijder EJ, Decroly E, Ziebuhr J (2016) The nonstructural proteins directing coronavirus RNA synthesis and processing. Adv Virus Res 96:59–126. https://doi.org/10.1016/bs.aivir.2016.08.008

Solyom Z, Schwarten M, Geist L, Konrat R, Willbold D, Brutscher B (2013) BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. J Biomol NMR 55:311–321. https://doi.org/10.1007/s10858-013-9715-0

Su D et al (2006) Dodecamer structure of severe acute respiratory syndrome coronavirus nonstructural protein nsp10. J Virol 80:7902–7908. https://doi.org/10.1128/jvi.00483-06

Viswanathan T et al (2020) Structural basis of RNA cap modification by SARS-CoV-2. Nat Commun 11:4–10. https://doi.org/10.1038/s41467-020-17496-8

Wang Y et al (2015) Coronavirus nsp10/nsp16 methyltransferase can be targeted by nsp10-derived peptide in vitro and in vivo to reduce replication and pathogenesis. J Virol 89:8416–8427. https://doi.org/10.1128/jvi.00948-15

Wishart DS (2011) Interpreting protein chemical shift data. Prog Nucl Magn Reson Spectrosc 58:62–87. https://doi.org/10.1016/j.pnmrs.2010.07.004

Wu F et al (2020) A new coronavirus associated with human respiratory disease in China. Nature 579:265–269. https://doi.org/10.1038/s41586-020-2008-3

Züst R et al (2011) Ribose 2'-O-methylation provides a molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5. Nat Immunol 12:137–143. https://doi.org/10.1038/ni.1979

5.12 **Research article:** ^{1}H, ^{13}C, and ^{15}N backbone chemical shift assignments of the apo and the ADP-ribose bound forms of the macrodomain of SARS-CoV-2 non-structural protein 3b

Francesca Cantini*, L. Banci, Nadide Altincekic, Jasleen Kaur Bains, Karthikeyan Dhamotharan, Christin Fuks, Boris Fürtig, Santosh L. Gande, Bruno Hargittay, Martin Hengesbach, Marie T. Hutchison, Sophie M. Korn, Nina Kubatova, Felicitas Kutz, Verena Linhard, Frank Löhr, Nathalie Meiser, Dennis J. Pyper, Nusrat S. Qureshi, Christian Richter, Krishna Saxena*, Andreas Schlundt, Harald Schwalbe, Sridhar Sreeramulu, Jan-Niklas Tants, Anna Wacker, Julia E. Weigand, Jens Wöhnert, Aikaterini C. Tsika*, Nikolaos K. Fourkiotis and Georgios A. Spyroulias; *Biomolecular NMR Assignments* **14**, 339-346 (2020)

* These authors contributed equally to this work.

In this article, the macrodomain non-structural protein 3b (Nsp3b) of the SARS-CoV-2 genome was characterized by NMR spectroscopy. ^{1}H, ^{13}C, and ^{15}N backbone chemical shift assignments of Nsp3b in its apo-form and in complex with ADP-ribose were presented. In addition, the secondary structure of Nsp3b in solution was solved as a basis for further NMR studies between potential small-molecule inhibitors and the catalytic site of Nsp3b.

The project was designed and coordinated by the Covid-19 NMR consortium. The manuscript was prepared by F. Cantini, K. Saxena and A. C. Tsika. The author of this thesis was part of the Covid-19 NMR consortium and the protein production team that prepared and purified various isotope-labeled protein samples for NMR spectroscopy.

https://doi.org/10.1007/s12104-020-09973-4

Biomolecular NMR Assignments (2020) 14:339–346
https://doi.org/10.1007/s12104-020-09973-4

ARTICLE

^{1}H, ^{13}C, and ^{15}N backbone chemical shift assignments of the apo and the ADP-ribose bound forms of the macrodomain of SARS-CoV-2 non-structural protein 3b

F. Cantini[1,2] · L. Banci[1,2] · N. Altincekic[3] · J. K. Bains[3] · K. Dhamotharan[4] · C. Fuks[3] · B. Fürtig[3] · S. L. Gande · B. Hargittay[3] · M. Hengesbach[3] · M. T. Hutchison[3] · S. M. Korn[4] · N. Kubatova[3] · F. Kutz[3] · V. Linhard[3] · F. Löhr[5] · N. Meiser[3] · D. J. Pyper[3] · N. S. Qureshi[3] · C. Richter[3] · K. Saxena[3] · A. Schlundt[4] · H. Schwalbe[3] · S. Sreeramulu[3] · J.-N. Tants[4] · A. Wacker[3] · J. E. Weigand[6] · J. Wöhnert[4] · A. C. Tsika[7] · N. K. Fourkiotis[7] · G. A. Spyroulias[7]

Received: 13 July 2020 / Accepted: 31 July 2020 / Published online: 14 August 2020

Abstract

The SARS-CoV-2 genome encodes for approximately 30 proteins. Within the international project COVID19-NMR, we distribute the spectroscopic analysis of the viral proteins and RNA. Here, we report NMR chemical shift assignments for the protein Nsp3b, a domain of Nsp3. The 217-kDa large Nsp3 protein contains multiple structurally independent, yet functionally related domains including the viral papain-like protease and Nsp3b, a macrodomain (MD). In general, the MDs of SARS-CoV and MERS-CoV were suggested to play a key role in viral replication by modulating the immune response of the host. The MDs are structurally conserved. They most likely remove ADP-ribose, a common posttranslational modification, from protein side chains. This de-ADP ribosylating function has potentially evolved to protect the virus from the anti-viral ADP-ribosylation catalyzed by poly-ADP-ribose polymerases (PARPs), which in turn are triggered by pathogen-associated sensing of the host immune system. This renders the SARS-CoV-2 Nsp3b a highly relevant drug target in the viral replication process. We here report the near-complete NMR backbone resonance assignment (^{1}H, ^{13}C, ^{15}N) of the putative Nsp3b MD in its apo form and in complex with ADP-ribose. Furthermore, we derive the secondary structure of Nsp3b in solution. In addition, ^{15}N-relaxation data suggest an ordered, rigid core of the MD structure. These data will provide a basis for NMR investigations targeted at obtaining small-molecule inhibitors interfering with the catalytic activity of Nsp3b.

Keywords SARS-CoV-2 · Non-structural protein · Macrodomain · Solution NMR-spectroscopy · Protein drugability · COVID19-NMR

F. Cantini, K. Saxena and A. C. Tsika are joint first authors.

✉ L. Banci
banci@cerm.unifi.it;
covid19-nmr@dlist.server.uni-frankfurt.de

✉ H. Schwalbe
Schwalbe@nmr.uni-frankfurt.de

✉ G. A. Spyroulias
G.A.Spyroulias@upatras.gr

[1] Magnetic Resonance Center – CERM, University of Florence, Via Luigi Sacconi 6, Sesto Fiorentino, 50019 Florence, Italy

[2] Department of Chemistry, University of Florence, Via della Lastruccia 3, Sesto Fiorentino, 50019 Florence, Italy

[3] Institute for Organic Chemistry and Chemical Biology, Center for Biomolecular Magnetic Resonance (BMRZ), Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue-Str. 7, 60438 Frankfurt, Germany

[4] Institute for Molecular Biosciences, Center for Biomolecular Magnetic Resonance (BMRZ), Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue-Str. 7, 60438 Frankfurt, Germany

[5] Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance (BMRZ), Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue-Str. 7, 60438 Frankfurt, Germany

[6] Department of Biology, Technical University of Darmstadt, Schnittspahnstr. 10, 64287 Darmstadt, Germany

[7] Department of Pharmacy, University of Patras, 26504 Patras, Greece

Biological context

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the cause of the pandemic that began in early 2020 and is accompanied by the respiratory disease COVID-19, is the latest member of a *Coronaviridae* clade, which also includes SARS-CoV from 2002 and the Middle east respiratory syndrome (MERS)-CoV. The severe velocity of the virus spread demands rapid action both in the development of a vaccine and in the development of potent virus inhibitors to weaken or eliminate the symptoms, which pose a major threat to the lives of elderly people worldwide.

The ~ 30 kb long positive sense single-stranded RNA genome of SARS-CoV-2 is one of the largest known virus genomes. The SARS-CoV-2 genome contains 14 putative open reading frames (ORFs). The majority of these ORFs was shown to be translated into functional viral proteins (Gordon et al. 2020). Among the highly conserved proteins of Betacoronaviruses (Yoshimoto 2020), the ORF1a/b-coded non-structural proteins (Nsps) 1–16 form the replication/transcription-complex—an incompletely understood network of viral-viral and viral-host protein–protein and RNA–protein interactions. Besides the membrane-bound Spike protein, which is important for the entry of the virus into the cell, a number of Nsps such as the two proteases Nsp5 (Mpro) and Nsp3d (PLpro), the Nsp3b ADP-ribose-phosphatase macrodomain (MD), and the Nsp7/8/12 RNA-dependent RNA polymerase complex are obvious drug targets.

Nsp3, the largest Nsp, is one of the most intriguing SARS-coronavirus proteins, consisting of a multitude of functionally related, but nevertheless independent domains (Snijder et al. 2003). The proteolytic processing of Nsp3 from the full-length ORF1-encoded polypeptide chain yields, a 1945 amino acid long multidomain protein. Starting from the N-terminus, its individual functional domains are named Nsp3a to Nsp3e followed by the ectodomain, which is embedded between two transmembrane regions, and the C-terminal CoV-Y domain. Nsp3b is a conserved ADP-ribose binding MD. In general, the MDs of SARS and MERS are implicated to play a key role in viral replication and modulate the immune response of the host. The MDs are structurally conserved and are thought to enzymatically remove ADP-ribose, a common posttranslational protein modification. The de-ADP ribosylating function of these enzymes has evolved to protect the virus against the anti-viral ADP-ribosylation catalyzed by poly-ADP-ribose polymerases (PARPs), which are activated by the innate immune system of the host upon sensing of pathogens. Therefore, the SARS-CoV-2 Nsp3b is a highly relevant drug target in the viral replication process.

Table 1 List of experiments collected to perform the sequence specific assignment of apo-Nsp3b (A) and ADP-ribose bound Nsp3b (B). Main parameters used are reported

Experiments	Time domain data size (points)			Spectral width (ppm)			ns	Delay time (s)
	t_1	t_2	t_3	F_1	F_2	F_3		
A. apo-Nsp3b								
^{1}H-^{15}N-HSQC	256	2048		36.5 (^{15}N)	16.0 (^{1}H)		4	1.2
^{1}H-^{15}N best-TROSY	256	2048		36.5 (^{15}N)	16.0 (^{1}H)		4	0.2
Best-TROSY-HN(CO)CACB	112	64	3072	75.3 (^{15}N)	32.2 (^{15}N)	13.9 (^{1}H)	40	0.25
Best-TROSY-HNCACB	112	64	3072	75.3 (^{13}C)	32.2 (^{15}N)	13.9 (^{1}H)	40	0.25
Best-TROSY-HN(CA)CO	104	64	3072	14.7 (^{13}C)	36.5 (^{15}N)	13.9 (^{1}H)	40	0.25
Best-TROSY-HNCO	104	64	3072	14.7 (^{13}C)	36.5 (^{15}N)	13.9 (^{1}H)	4	0.25
^{15}N R_1	10	128	2048	10.0 (^{1}H)	35.0 (^{15}N)	14.0 (^{1}H)	16	1.2
^{15}N R_2	10	128	2048	10.0 (^{1}H)	35.0 (^{15}N)	14.0 (^{1}H)	16	1.2
^{15}N-NOE	2	128	2048	10.0 (^{1}H)	35.0 (^{15}N)	14.0 (^{1}H)	16	3
B. Nsp3b-ADP-ribose								
^{1}H-^{15}N-HSQC	256	2048		36.5 (^{15}N)	16.0 (^{1}H)		4	1.2
^{1}H-^{15}N best-TROSY	256	2048		36.5 (^{15}N)	16.0 (^{1}H)		4	0.2
Best-TROSY-HN(CO)CACB	112	64	3072	75.3 (^{15}N)	32.2 (^{15}N)	13.9 (^{1}H)	48	0.25
Best-TROSY-HNCACB	112	64	3072	75.3 (^{13}C)	32.2 (^{15}N)	13.9 (^{1}H)	40	0.25
Best-TROSY-HNCO	104	64	3072	14.7 (^{13}C)	36.5 (^{15}N)	13.9 (^{1}H)	4	0.25
^{15}N R_1	10	128	2048	10.0 (^{1}H)	35.0 (^{15}N)	14.0 (^{1}H)	16	1.2
^{15}N R_2	10	128	2048	10.0 (^{1}H)	35.0 (^{15}N)	14.0 (^{1}H)	16	1.2
^{15}N-NOE	2	128	2048	10.0 (^{1}H)	35.0 (^{15}N)	14.0 (^{1}H)	16	3

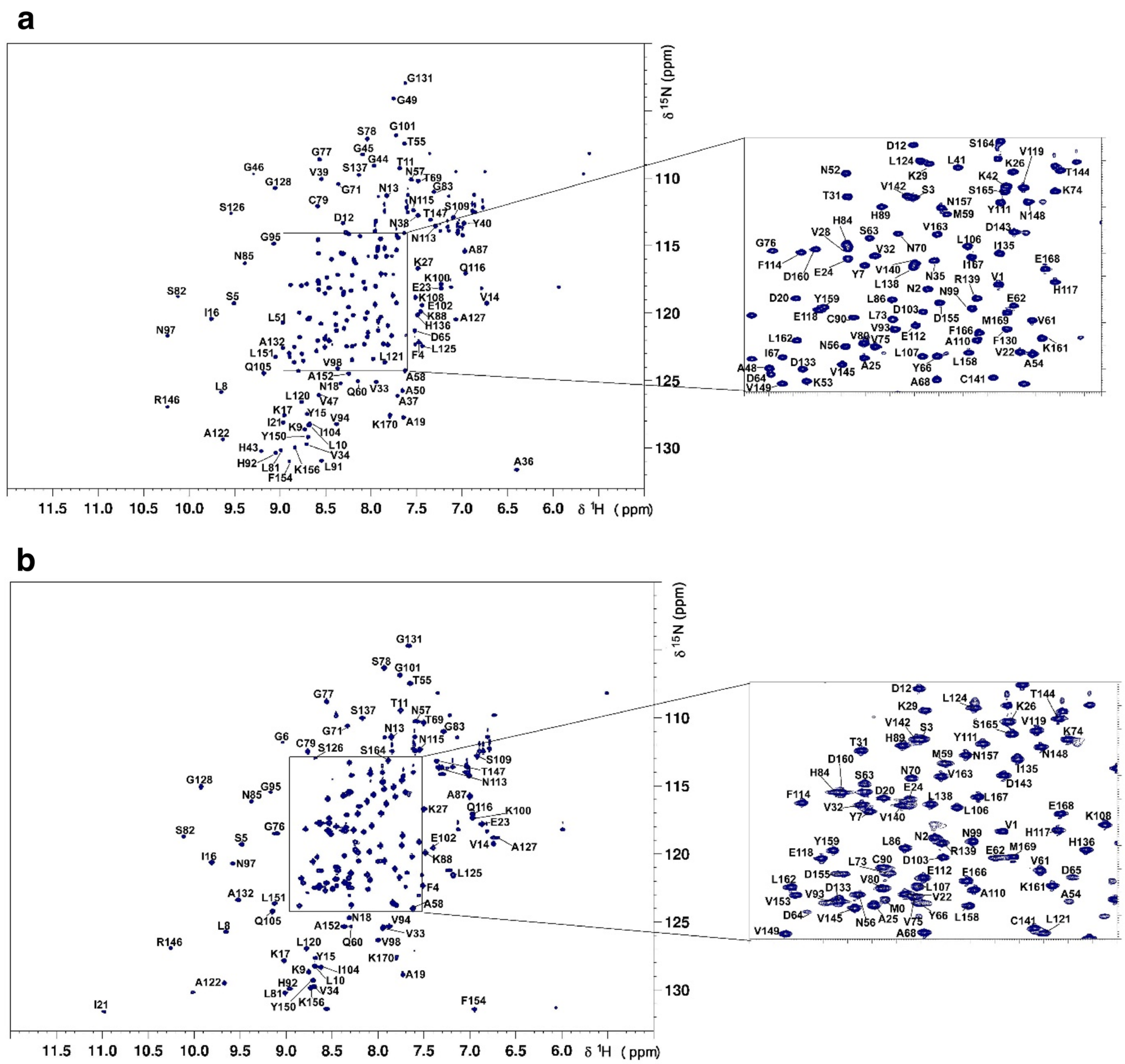


Fig. 1 ^{1}H,^{15}N-HSQC spectrum of the apo (**a**) and ADP-ribose bound (**b**) forms ^{13}C,^{15}N-labelled SARS-CoV-2 Nsp3b at 650 μM concentration in 25 mM Bis–Tris pH 6.5, 150 mM NaCl, 3 mM TCEP and 5% D_2O measured at 298 K on a 1.2 GHz Spectrometer with chemical shift assignment depicted. Backbone NH peaks are labelled with their assignments

The research consortium COVID19-NMR, which was founded at the end of March 2020, rapidly and publicly supports the search for antiviral drugs by enabling an NMR-based screening approach. This requires the large-scale production of all drugable proteins and RNAs of SARS-CoV-2, as well as an extensive assignment of their NMR resonances and the determination of their structures as a prerequisite for rational drug design. We provide here the near-complete backbone assignment of the SARS-CoV-2 Nsp3b MD and thereby enable its exploitation in subsequent applications, such as drug screening and interaction mapping with amino acid resolution.

Methods and experiments

Construct design

This study uses the SARS-CoV-2 NCBI reference genome entry NC_045512.2, identical to GenBank entry MN908947.3 (Wu, 2020, #13). The Nsp3b domain includes amino acids V207 to K376 within the full-length Nsp3 primary sequence, as reported in previous studies 10.2210/pdb6YWM/pdb. This sequence was inserted into a pET28a(+) vector, containing an N-terminal His$_6$-tag and a tobacco etch virus (TEV) cleavage site. Due to the nature of the TEV cleavage site, the produced protein contained three artificial N-terminal residues (G-2, H-1, M0) preceding the native protein sequence.

Fig. 3 S^2 order parameters of the backbone of SARS-CoV-2 Nsp3b ▶ in its apo form (**a**) and in complex with ADP-ribose (**b**). Values close to 1 suggest ordered structure on the ps/ns timescale. Errors were derived through Monte Carlo error analysis embedded in the fitting routine of Bruker software TopSpin3.6 Dynamic Center

Sample preparation

Uniformly ^{13}C,^{15}N-labeled Nsp3b protein was expressed in E. *coli* strain T7express in M9 minimal medium containing 1 g/L $^{15}NH_4Cl$ (Cambridge Isotope Laboratories), 2 g/L $^{13}C_6$-D-glucose (Eurisotop) and 50 µg/mL kanamycin. Protein expression was induced at an OD$_{600}$ of 0.7 with 0.5 mM IPTG and the cells were incubated for 13 h at 18 °C and 120 rpm shaking. The cell pellet was resuspended in buffer A (25 mM Tris–HCl–pH 8.0, 150 mM NaCl, 5 mM imidazole and 10 mM 2-mercaptoethanol), containing one protease inhibitor tablet (cOmplete™, Roche, Germany). The cells were mechanically lysed with Microfluidics M-110P at 15,000 psi (pound per square inch) under cooling with ice using three lysis cycles. The lysate was clarified from cell debris at maximal centrifugation speed for 45 min using a JLA 16.250 rotor. Clarified supernatant was purified via FPLC using two HisTrap HP columns (2 × 5 mL,

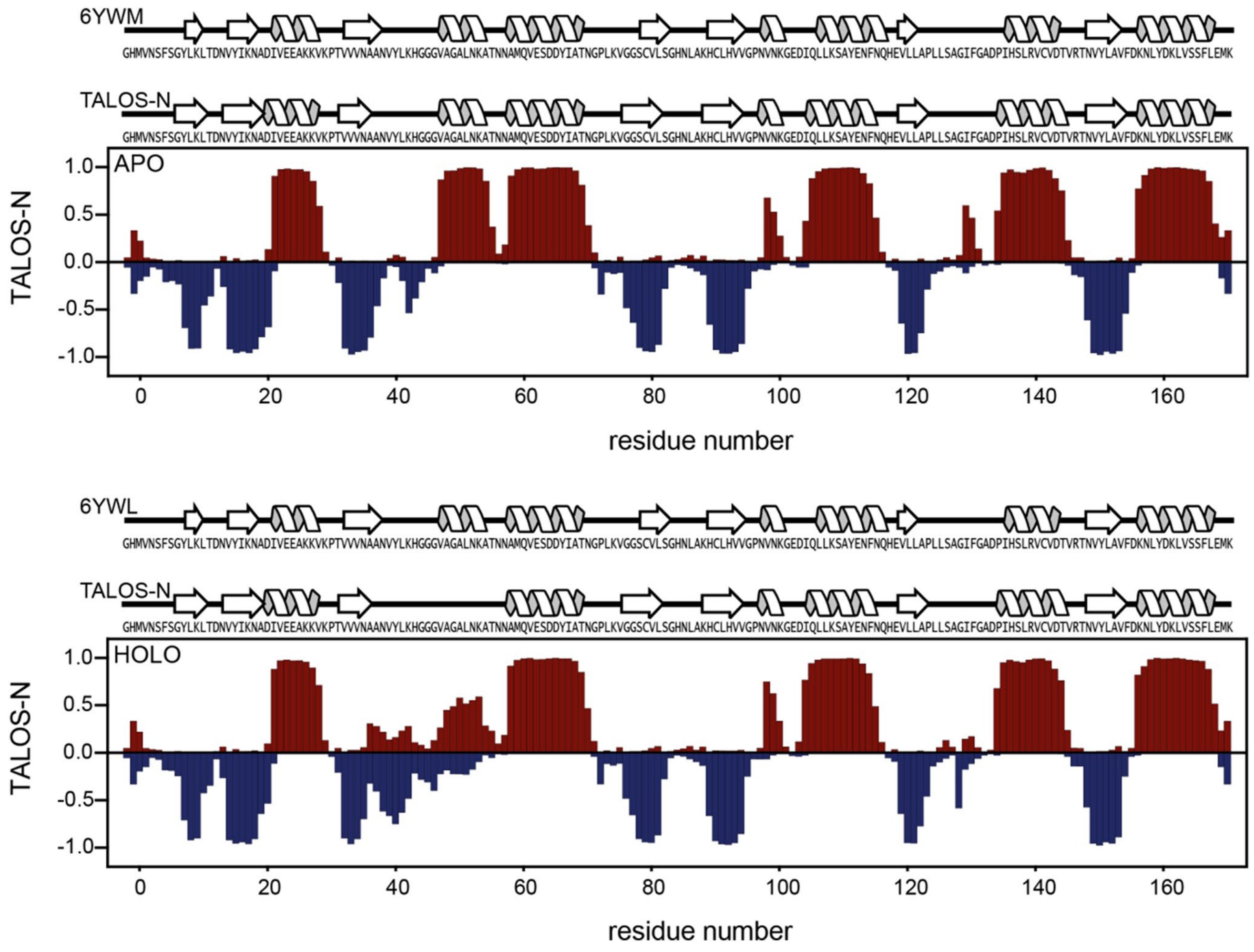
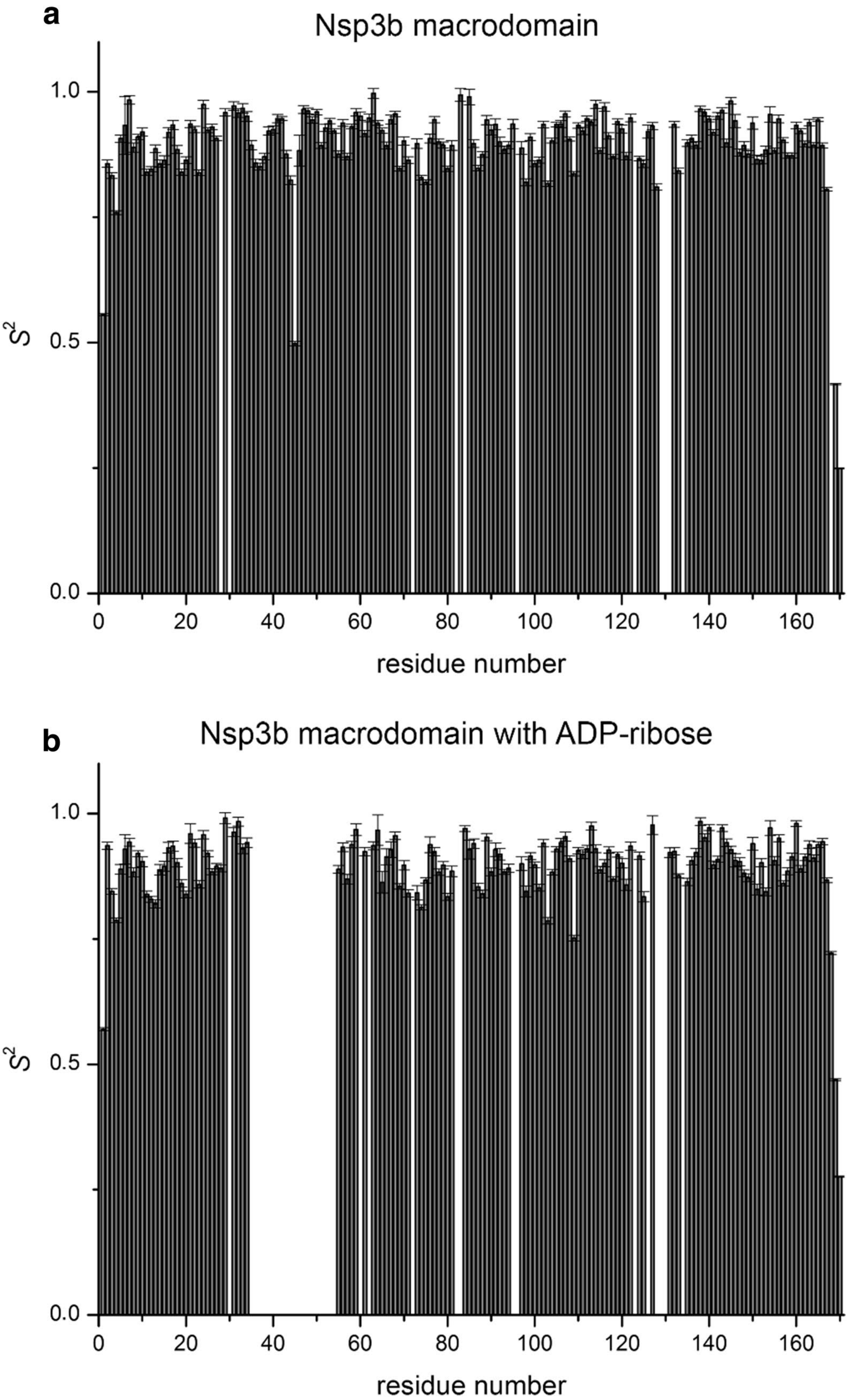


Fig. 2 Display of TALOS predicted secondary structure for the apo (**a**) and holo (**b**) Nsp3b. For comparison secondary structure elements obtained from X-ray structures and TALOS-N (Shen and Bax 2013) are displayed on the top of each plot. For the for residues between 35 and 53 in the ADP-bound Nsp3b, the predictions were sequence based. In case of X-ray structures, the secondary structures were extracted with pdbsum (Laskowski et al. 1997) using the pdb entries 6YWM (apo) and 6YWL (ADP-ribose bound)


a
Nsp3b macrodomain
S^2
1.0
0.5
0.0
0
20
40
60
80
100
120
140
160
residue number
b
Nsp3b macrodomain with ADP-ribose
S^2
1.0
0.5
0.0
0
20
40
60
80
100
120
140
160
residue number

GE Healthcare, USA). Bound protein was washed with 4% buffer B (buffer A + 500 mM imidazole) and eluted with 100% buffer B. Protein containing fractions were pooled and subjected to TEV cleavage over night at 4 °C while dialyzing against 25 mM Tris–HCl pH 8.0, 150 mM NaCl, 10 mM 2-mercaptoethanol. TEV protease and tag were removed via a second IMAC purification. Protein containing fractions were pooled and concentrated using Amicon Ultra-4 filtration devices (regenerated cellulose 10 kDa NMWL) and purified with a Superdex 75 26/600 PG (320 mL GE Healthcare, USA) using a buffer containing 25 mM Bis–Tris pH 6.5, 150 mM NaCl, 3 mM tris-(2-carboxyethyl)-phosphin (TCEP). The holo sample was prepared as follows. A 100 mM stock solution of ADP-ribose sodium (Sigma A0752) was prepared in water. This stock solution was used to prepare the Nsp3b-ADP-ribose complex by adding a tenfold molar excess to the protein Nsp3b (650 µM).

NMR experiments

All experiments for the backbone assignment of both apo and ADP-ribose bound Nsp3b were recorded at 298 K using an ultra-high field Avance NEO 1.2 GHz NMR spectrometer, equipped with a 3 mm TCI H/C/N CryoProbe. All spectra were acquired using standard pulse sequences (Favier and Brutscher 2011; Lescop et al. 2007; Solyom et al. 2013) and processed using the Bruker software TopSpin 4.0.6. For the assignment, a set of double and triple resonance experiments was performed. The set of NMR experiments used for sequence specific assignment is summarized in Table 1.

Relaxation experiments (^{15}N T_1, T_2 and {^{1}H}–^{15}N NOE) were conducted on a Bruker Avance III four-channel 700 MHz NMR spectrometer equipped with a cryogenically cooled 5 mm ^{1}H/^{13}C/^{15}N/D Z-gradient probe (TCI), at 298 K using the TROSY pseudo3D pulse sequences (Zhu et al. 2000). The delays for the ^{15}N $T1$ were 20, 60, 100, 200, 400, 600, 800 and 1200 ms, while delays of 15.68, 31.36, 62.72, 94.08, 125.44, 156.80, 188.16 and 219.52 ms were used in the ^{15}N $T2$ experiments. The model free approach in the Dynamic Center/Topsin3.6 software was used for data analysis and in order to obtain the S^2 values. Data were fitted using a global isotropic model (M1 included in Dynamic Center software, using the equation: $j(\omega) = (2/5)\tau_c\left[S^2/(1 + (\tau_c\omega)^2)\right]$.

Proton resonances were calibrated with respect to the signal of 2,2-dimethylsilapentane-5-sulfonic acid (DSS). Nitrogen and carbon chemical shifts were referenced indirectly to the ^{1}H standard using a conversion factor derived from the ratio of NMR frequencies (Wishart et al. 1995).

Assignments and data deposition

The ^{1}H,^{15}N-HSQC spectra showed well-dispersed amide signals for both apo (Fig. 1a) and ADP-ribose bound Nsp3b (Fig. 1b). Assignments of apo and ADP-ribose bound Nsp3b were performed with the program CARA (https://www.nmr.ch). For apo Nsp3b we assigned 98%

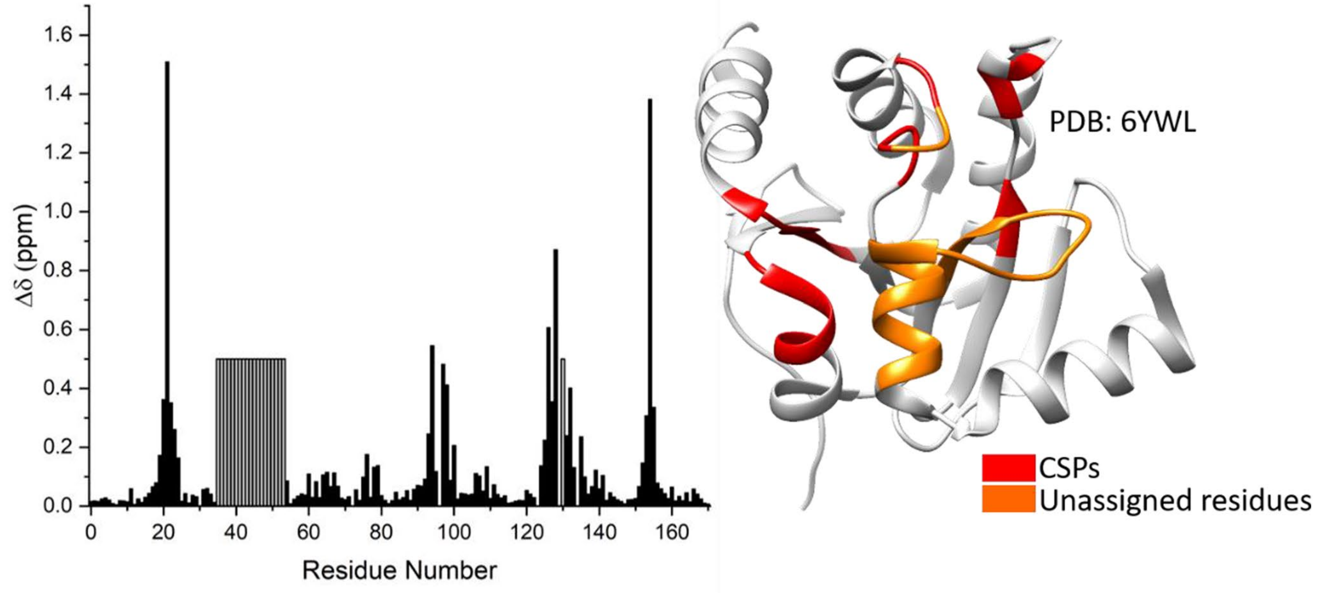


Fig. 4 Chemical shift perturbations (CSPs) between the apo and holo Nsp3b-ADP-ribose complex are plotted as a function of Nsp3b residue number. The observed CSPs are mapped onto the crystal structure (6YWL)

of ^{1}H/^{15}N backbone pairs and 98.8, 99.4 and 99.4% of all CO, Cα and Cβ chemical shifts, respectively. In the case of ADP-ribose Nsp3b we assigned 86% of ^{1}H/^{15}N backbone pairs and 83.2, 87.3 and 89.3% of all CO, Cα and Cβ chemical shifts, respectively. The unassigned residues of the ADP-ribose bound Nsp3b correspond to the stretches Asn35-Lys53 and Ile129-Phe130 (see Fig. 2). Furthermore, none of the unassigned peaks are seen in the HSQC suggesting exchange broadening of these residues in the presence of ADP-ribose.

Secondary structure evaluation was performed using chemical shift assignments of five atoms (H^{N}, Cα, Cβ, CO, N) for a given residue in the sequence with TALOS-N (Shen and Bax 2013). The results for Nsp3b and ADP-ribose bound Nsp3b (Fig. 2) are in good agreement with each other. Furthermore, we observed that the dihedral angles predicted by TALOS-N (Shen and Bax 2013) for both apo and holo Nsp3b are in excellent agreement with the dihedral angles found in the apo (6YWM) and holo (6YWL) crystal structures, indicating that ligand binding does not alter the overall the secondary structure within the MD.

Backbone amide order parameters S^2 are presented in Fig. 3 and reveal an ordered, rigid core of the structure, with slightly flexible termini both, for the apo Nsp3b and its complexed form with ADP-ribose. The correlation time for isotropic tumbling in solution as calculated from the R_2/R_1 ratio is 9.15 ± 0.5 and 9.10 ± 0.5 ns for the apo Nsp3b and the holo Nsp3b-ADP-ribose complex, respectively (theoretical MW 18.65 kDa).

ADP-ribose binds to Nsp3b with a dissociation constant (K_D) of 13 μM (Frick et al. 2020). The position of the ADP-ribose molecule within the SARS-CoV-2 Nsp3b binding site was defined through an NMR characterization in solution (Fig. 4). The mapped binding site is in good agreement with the binding pocket observed in the crystal structure (6YWL). The chemical shift values for the ^{1}H, ^{13}C and ^{15}N resonances of apo and holo forms of SARS-CoV-2 Nsp3b have been deposited at the BioMagResBank (https://www.bmrb.wisc.edu) under accession numbers 50387 and 50,388, respectively.

Acknowledgements Open Access funding provided by Projekt DEAL. Work at BMRZ is supported by the state of Hesse. Work in COVID19-NMR was supported by the Goethe Corona Funds and the DFG in CRC902: "Molecular Principles of RNA-based regulation." Work at CERM is supported by the Italian Ministry for University and Research (FOE funding) to the Italian Center (CERM, University of Florence) of Instruct-ERIC, an European Research Infrastructure, ESFRI Landmark. The work was supported by the INSPIRED (MIS 5002550) which is implemented under the Action 'Reinforcement of the Research and Innovation Infrastructure,' funded by the Operational Program 'Competitiveness, Entrepreneurship and Innovation' (NSRF 2014–2020) and co-financed by Greece and the European Union (European Regional Development Fund). EU FP7 REGPOT CT-2011-285950 – "SEE-DRUG" project is acknowledged for the purchase of UPAT's 700 MHz NMR equipment. We thank Prof. Dr. Stefan Knapp for kindly providing the Nsp3b clone.

Compliance with ethical standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

Favier A, Brutscher B (2011) Recovering lost magnetization: polarization enhancement in biomolecular NMR. J Biomol NMR 49:9–15

Frick DN, Virdi RS, Vuksanovic N, Dahal N, Silvaggi NR (2020) Molecular basis for ADP-ribose binding to the Mac1 domain of SARS-CoV-2 Nsp3. Biochemistry. https://doi.org/10.1021/acs.biochem.0c00309

Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, O'Meara MJ, Rezelj VV, Guo JZ, Swaney DL et al (2020) A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 583(7816):459–468

Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM (1997) PDBsum: a Web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci 22:488–490

Lescop E, Schanda P, Brutscher B (2007) A set of BEST triple-resonance experiments for time-optimized protein resonance assignment. J Magn Reson 187:163–169

Shen Y, Bax A (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. J Biomol NMR 56:227–241

Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, Gorbalenya AE (2003) Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. J Mol Biol 331:991–1004

Solyom Z, Schwarten M, Geist L, Konrat R, Willbold D, Brutscher B (2013) BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. J Biomol NMR 55:311–321

Wishart DS, Bigam CG, Yao J, Abildgaard F, Dyson HJ, Oldfield E, Markley JL, Sykes BD (1995) 1H, 13C and 15N chemical shift referencing in biomolecular NMR. J Biomol NMR 6:135–140

Yoshimoto FK (2020) The proteins of severe acute respiratory syndrome coronavirus-2 (SARS CoV-2 or n-COV19), the cause of COVID-19. Protein J 39:198–216

Zhu G, Xia Y, Nicholson LK, Sze KH (2000) Protein dynamics measurements by TROSY-based NMR experiments. J Magn Reson 143:423–426

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5.13 **Research article:** Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy

Anna Wacker*, Julia E. Weigand*, Sabine R. Akabayov, Nadide Altincekic, Jasleen Kaur Bains, Elnaz Banijamali, Oliver Binas, Jesus Castillo-Martinez, Erhan Cetiner, Betül Ceylan, Liang-Yuan Chiu, Jesse Davila-Calderon, Karthikeyan Dhamotharan, Elke Duchardt-Ferner, Jan Ferner, Lucio Frydman, Boris Fürtig, José Gallego, J. Tassilo Grün, Carolin Hacker, Christina Haddad, Martin Hähnke, Martin Hengesbach, Fabian Hiller, Katharina F. Hohmann, Daniel Hymon, Vanessa de Jesus, Henry Jonker, Heiko Keller, Bozana Knezic, Tom Landgraf, Frank Löhr, Le Luo, Klara R. Mertinkus, Christina Muhs, Mihajlo Novakovic, Andreas Oxenfarth, Martina Palomino-Schätzlein, Katja Petzold, Stephen A. Peter, Dennis J. Pyper, Nusrat S. Qureshi, Magdalena Riad, Christian Richter, Krishna Saxena, Tatjana Schamber, Tali Scherf, Judith Schlagnitweit, Andreas Schlundt, Robbin Schnieders, Harald Schwalbe, Alvaro Simba-Lahuasi, Sridhar Sreeramulu, Elke Stirnal, Alexey Sudakov, Jan-Niklas Tants, Blanton S. Tolbert, Jennifer Vögele, Lena Weiß, Julia Wirmer-Bartoschek, Maria A. Wirtz Martin, Jens Wöhnert and Heidi Zetzsche; *Nucleic Acids Research* **48**, 12415-12435 (2020)

* These authors contributed equally to this work.

In this work, potential RNA drug targets of SARS-CoV-2 (SCoV2) were characterized by NMR spectroscopy within the Covid-19 NMR consortium. The NMR-based secondary structure determination included 15 conserved RNA elements located in the 5'-UTR, the ribosomal frameshift region, and the 3'-UTR of the SCoV2 genome. In addition, 11 of these RNAs were further analyzed by DMS footprinting experiments to detect the highly flexible residues within the unstructured RNA regions.

The project was designed and coordinated by the Covid-19 NMR consortium. The RNA samples were prepared by the RNA production team and the NMR spectra were assigned in smaller groups. The manuscript was written by A. Wacker and J. E. Weigand. The author of this thesis was part of the Covid-19 NMR consortium and the RNA production team that prepared and purified various isotope-labeled RNA samples for NMR spectroscopy.

https://doi.org/10.1093/nar/gkaa1013

Published online 10 November 2020 *Nucleic Acids Research, 2020, Vol. 48, No. 22* ***12415–12435***
https://doi.org/10.1093/nar/gkaa1013

NAR Breakthrough Article

Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy

Anna Wacker[1,†], Julia E. Weigand[2,†], Sabine R. Akabayov[3], Nadide Altincekic[1], Jasleen Kaur Bains[1], Elnaz Banijamali[4], Oliver Binas[1], Jesus Castillo-Martinez[5], Erhan Cetiner[1], Betül Ceylan[1], Liang-Yuan Chiu[6], Jesse Davila-Calderon[6], Karthikeyan Dhamotharan[7], Elke Duchardt-Ferner[7], Jan Ferner[1], Lucio Frydman[3], Boris Fürtig[1], José Gallego[5], J. Tassilo Grün[1], Carolin Hacker[8], Christina Haddad[6], Martin Hähnke[8], Martin Hengesbach[1], Fabian Hiller[8], Katharina F. Hohmann[1], Daniel Hymon[1], Vanessa de Jesus[1], Henry Jonker[1], Heiko Keller[7], Bozana Knezic[1], Tom Landgraf[1], Frank Löhr[9], Le Luo[4], Klara R. Mertinkus[1], Christina Muhs[1], Mihajlo Novakovic[3], Andreas Oxenfarth[1], Martina Palomino-Schätzlein[5], Katja Petzold[4], Stephen A. Peter[2], Dennis J. Pyper[1], Nusrat S. Qureshi[1], Magdalena Riad[4], Christian Richter[1], Krishna Saxena[1], Tatjana Schamber[1], Tali Scherf[3], Judith Schlagnitweit[4], Andreas Schlundt[7], Robbin Schnieders[1], Harald Schwalbe[1,*], Alvaro Simba-Lahuasi[5], Sridhar Sreeramulu[1], Elke Stirnal[1], Alexey Sudakov[1], Jan-Niklas Tants[7], Blanton S. Tolbert[6], Jennifer Vögele[7], Lena Weiß[7], Julia Wirmer-Bartoschek[1], Maria A. Wirtz Martin[1], Jens Wöhnert[7] and Heidi Zetzsche[1]

[1]Institute for Organic Chemistry and Chemical Biology, Max-von-Laue-Strasse 7, 60438 Frankfurt/M., Germany, [2]Department of Biology, Technical University of Darmstadt, Schnittspahnstrasse 10, 64287 Darmstadt, Germany, [3]Faculty of Chemistry, Weizmann Institute of Science, 7610001 Rehovot, Israel, [4]Department of Medical Biochemistry and Biophysics, Karolinska Institute, Biomedicum 9B, Solnavägen 9, 17177 Stockholm, Sweden, [5]School of Medicine, Catholic University of Valencia, C/Quevedo 2, 46001 Valencia, Spain, [6]Department of Chemistry, Case Western Reserve University, Cleveland, OH 44106, USA, [7]Institute for Molecular Biosciences, [8]Signals GmbH & Co. KG, Graf-von-Stauffenberg-Allee 83, 60438 Frankfurt/M, Germany and [9]Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance (BMRZ), Johann Wolfgang Goethe-University Frankfurt, Max-von-Laue-Strasse 7, 60438 Frankfurt/M., Germany

Received June 23, 2020; Revised September 08, 2020; Editorial Decision October 12, 2020; Accepted October 14, 2020

ABSTRACT

The current pandemic situation caused by the Betacoronavirus SARS-CoV-2 (SCoV2) highlights the need for coordinated research to combat COVID-19. A particularly important aspect is the development of medication. In addition to viral proteins, structured RNA elements represent a potent alternative as drug targets. The search for drugs that target RNA requires their high-resolution structural characterization. Using nuclear magnetic resonance (NMR) spectroscopy, a worldwide consortium of NMR researchers aims to characterize potential RNA drug targets of SCoV2. Here, we report the characterization of 15 conserved RNA elements located at the 5′ end, the ribosomal frameshift segment and the 3′-untranslated region (3′-UTR) of the SCoV2 genome, their large-scale production and NMR-based secondary structure determination. The NMR data are corroborated with secondary structure probing by

[*]To whom correspondence should be addressed. Tel: +49 69798 29737; Fax: +49 69798 29515; Email: schwalbe@nmr.uni-frankfurt.de

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

DMS footprinting experiments. The close agreement of NMR secondary structure determination of isolated RNA elements with DMS footprinting and NMR performed on larger RNA regions shows that the secondary structure elements fold independently. The NMR data reported here provide the basis for NMR investigations of RNA function, RNA interactions with viral and host proteins and screening campaigns to identify potential RNA binders for pharmaceutical intervention.

INTRODUCTION

Betacoronaviruses contain a large single-stranded RNA genome of ∼30 000 nucleotides (nts). SCoV2 causing the COVID-19 disease contains multiple regions with very high sequence or secondary structure conservation relative to the 2002 SARS-Coronavirus (SCoV) and other Betacoronaviruses (1,2). The function of these putative *cis*-acting RNA elements have been characterized in different Betacoronaviruses and are associated with regulation of replication, subgenomic mRNA (sg mRNA) production and translation (3–5). Structural models for many of these RNA elements conserved in SCoV2 have been provided by *in silico* methods using RNA structure prediction, mutational covariance analysis or homology models (1,6,7).

Here, we systematically characterize the secondary structures of all conserved *cis*-acting RNA elements of SCoV2 by nuclear magnetic resonance (NMR) spectroscopy, providing high-resolution experimental secondary structure models. We analyzed nine RNA constructs representing the eight stem-loop (SL) domains present at the genomic 5′-end, two RNA constructs corresponding to *cis*-acting elements from the ORF1a/b frameshifting region and four RNA constructs representing functional SLs within the viral 3′-UTR (Figure 1). These 15 RNA elements were chosen based on their conservation within the Betacoronavirus family and their importance for viral propagation. A detailed comparison regarding their structural conservation between SCoV2 and SCoV is shown in Supplementary Figure S1.

We provide a streamlined pipeline for the fast and unambiguous NMR-based assignment of 2D structures in high-throughput. In a coordinated approach involving laboratories at the Goethe-University Frankfurt, the Technical University of Darmstadt, Case Western Reserve University (CWRU), and the Catholic University of Valencia, we produced 20 RNAs in isotopically labeled form representing 15 *cis*-acting RNA elements of the SCoV2 genome. Here, we report the NMR spectroscopic investigation to experimentally determine the secondary structure for SL1, SL2+3, SL4, three substructures of SL5 (SL5stem, SL5a, SL5b+c), SL6, SL7 and SL8, of the 5′-genomic region (subsequently denoted ‘5_’), the attenuator hairpin (att HP) and the three-stemmed pseudoknot (PK) linked to programmed ribosomal frameshifting (PRF), and SL1, SL2, SL3 and the isolated s2m motif from the 3′-UTR (subsequently denoted ‘3_’)

In addition, we prepared four larger constructs: two constructs representing SLs SL1 to SL8 (5′-geRNA, the first 472 nt of the SCoV2 genome) and SL1 to SL4 (5_SL1-4, nts 6–125) from the 5′-genomic end and two constructs from the 3′-end, the hypervariable region (3_HVR, nts 29 697 to 29 805) and the full-length 337 nt 3′-UTR (3′-UTR, nts 29 534 to 29 870) in ^{15}N-labeled form for NMR investigations.

The NMR structural analyses were conducted at Weizmann Institute Rehovot, Karolinska Institute Stockholm, Catholic University of Valencia, CWRU and Goethe University Frankfurt (BMRZ). In addition, a DMS footprinting analysis was performed for the 5′- and 3′-genomic ends of the SCoV2 genomic RNA at CWRU and compared to the NMR results. Our data will facilitate research efforts aiming to map interactions with viral and host proteins or characterize the binding mode of small-molecule ligands targeting the viral RNA.

We continuously update and report results on the webpage http://covid19-nmr.de and NMR chemical shift assignment in the BMRB (8) (http://www.bmrb.wisc.edu/).

MATERIALS AND METHODS

DMS footprinting

A total of 5 μl of 200 ng/μl purified SCoV2 5′-end and 3′-UTR RNA were heated to 95°C for 15 s and flash cooled on ice for 2 min. A total of 95 μl DMS modification buffer (100 mM sodium cacodylate, 140 mM KCl, 3 mM $MgCl_2$, pH 7.5) was added to the RNA sample and incubated for 30 min at room temperature. Two microliter of DMS was added, and the mixture was incubated at 37°C with 500 rpm shaking for 10 min. The methylation reaction was terminated by adding 60 μl of neat β-mercaptoethanol (Sigma-Aldrich). The modified RNA samples were purified and desalted using the RNA-cleanup-and-concentrator-5 column kit (Zymo Research) to recover RNAs containing more than 200 nts.

Methylated RNAs were used for reverse transcription to generate DNA with thermostable group II intron reverse transcriptase, third generation (TGIRT-III, InGex), following the manufacturer’s protocol. The reverse primers 5R2 and 3R2 (Supplementary Table S1) for SCoV2 5′-end and 3′-UTR, respectively, were commercially synthesized (IDT). The RNA templates were digested using RNase H (NEB) for 20 min at 37°C following the reverse transcription. The synthesized DNA was sequentially polymerase chain reaction (PCR) amplified using Phusion DNA polymerase (NEB) with the specific primer sets (5F1/5R1,5F2/5R2;3F1/3R1,3F2/3R2, see Supplementary Table S1). The following PCR protocol was applied: denaturing for 30 s at 98°C, followed by 30 PCR cycles, including denaturing for 5 s at 98°C, annealing for 10 s at optimal temperature and extension for 15 s at 72°C; final extension continued for 5 min at 72°C. PCR products were desalted using the DNA-cleanup-and-concentrator-5 column kit (Zymo Research). The homogeneity of the amplified samples was verified using agarose gel electrophoresis prior to sending out for sequencing.

The sequencing was performed on an Illumina *HiSeq 2000* system, which used cluster generation and sequencing by synthesis (SBS) chemistry. The sequence results of SCoV2 5′-genomic region and 3′-UTR RNA were aligned against index files and then used to generate the population

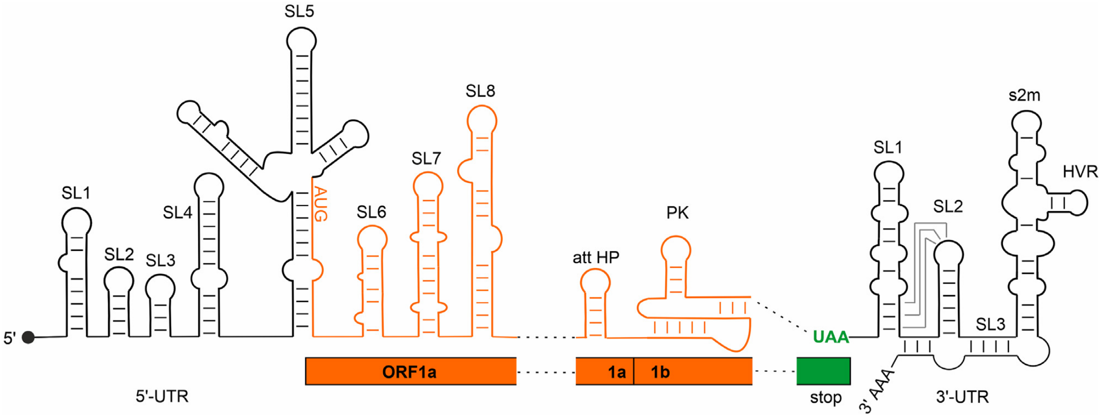


Figure 1. Overview of *cis*-acting RNA elements of the SCoV2 genome. Black: untranslated regions; orange: coding regions. The SL structures investigated in this study and their relative positions within the SCoV2 genome are shown schematically. 'stop' represents the end of ORF9 coding for the Nucleocapsid protein.

average and read coverage quality control files (9). The mutational signal of 5′ and 3′ primer overlapping regions and signal from T and G nucleotides in the sequence were determined to be below background. Raw data showing the signal intensities for DMS treated samples and untreated controls are shown in Supplementary Figure S2. The signals from A and C nucleotides were normalized to the highest reactivity following 95% Winsorization. The DMS restraint files for each RNA were used as input to guide folding of full-length SCoV2 5′-end and 3′-UTR with the fold algorithm from the RNAStructure (10) webserver and visualized by VARNA (11).

RNA synthesis for NMR experiments

Fast RNA production was achieved by distribution and parallelization of individual synthesis steps. In general, all RNAs were produced by T7 polymerase-based *in-vitro* transcription (12). In the DNA template production step, the sequences of 5_SL1, 5_SL2+3, 5_SL4, 5_SL5stem, 5_SL5a, 5_SL5b+c, 5_SL6, 5_SL7, 5_SL8, 5_SL8loop, PK, 3_SL1, 3_SL2, 3_SL3base and 3_s2m together with the T7 promoter were generated by hybridization of complementary oligonucleotides and introduced into the EcoRI and NcoI sites of an HDV ribozyme encoding plasmid based on the pSP64 vector (Promega). The DNA template for 5_SL1-4 was produced by PCR amplification from the full-length 5′-genomic region. RNAs were transcribed as HDV ribozyme fusions to obtain 3′ homogeneity (13). The DNA sequences of the full-length 5'-genomic region, the 3′-UTR and the hypervariable region (3_HVR) were purchased from Eurofins Genomics and introduced into the EcoRI and HindIII sites of the pSP64 vector (Promega). The DNA template for att HP together with the T7 promoter was generated by hybridization of complementary oligonucleotides and directly used for run-off transcription. Isolated RNA hairpins 5_SL5b and 5_SL5c were purchased from Dharmacon and purified according to the manufacturer's instructions. All RNA sequences are summarized in Supplementary Table S2. The complete vector sequences are available upon request.

The recombinant vectors were transformed and amplified in *Escherichia coli* strain DH5α. Plasmid-DNA (2–10 mg plasmid per liter SB medium) was purified by Gigaprep (Qiagen) according to the manufacturer's instructions and linearized with *Hind*III (or *Sma*I in case of 3_HVR) prior to *in-vitro* transcription by T7 RNA polymerase (P266L mutant, prepared as described in (14)). Four preparative-scale transcriptions (10–15 ml each) were routinely performed in parallel. The RNAs were purified as follows: preparative transcription reactions (6 h at 37°C) were terminated by addition of ethylenediaminetetraacetic acid (EDTA) and RNAs were precipitated with 2-propanol. RNA fragments were separated on 10–12% denaturing polyacrylamide (PAA) gels and visualized by UV shadowing at 254 nm. Desired RNA fragments were excised from the gel and RNA was eluted by passive diffusion into 0.3 M NaOAc, precipitated with EtOH or EtOH/acetone and desalted via PD10 columns (GE Healthcare). Except for the att HP RNA, residual PAA was removed by reversed-phase HPLC using a Kromasil RP 18 column and a gradient of 0–40% 0.1 M acetonitrile/triethylammonium acetate. After freeze-drying of RNA-containing fractions and cation exchange by $LiClO_4$ precipitation (2% in acetone), each RNA was folded in water by heating to 80°C followed by rapid cooling on ice except for the pseudoknot RNA, which was kept below 40°C to optimize folding into monomeric form. Buffer exchange to NMR buffer (25 mM potassium phosphate buffer, pH 6.2, 50 mM potassium chloride) was performed using centrifugal concentrators with a suitable molecular weight cut-off membrane. RNA purity was verified by denaturing PAA gel electrophoresis and homogenous folding was monitored by native PAA gel electrophoresis, loading the same RNA concentration as used for NMR experiments (Supplementary Figures S3 and 4).

The RNA constructs representing the entire 5′-genomic end and the entire 3′-UTR were purified without denaturing steps in order to maintain the native fold. Thus, transcription reactions for 5′-geRNA and 3′-UTR were terminated by addition of EDTA and RNA was extracted by aqueous phenolic extraction at pH 4. Phenol was removed by three cycles of chloroform extraction and subsequent buffer exchange via PD10 columns equilibrated with NMR buffer. The RNAs were purified by size exclusion chromatography using a HiLoad 26/60 Superdex 200 pg column. Fractions containing RNAs of the proper size were identified by denaturing and native gel electrophoresis, pooled and concentrated by centrifugal concentrators (MWCO: 50 kDa).

One production round of four NMR samples in parallel typically took 10 days from plasmid transformation to NMR tube filling. Table 1 provides a summary of the produced RNAs with their isotope labeling scheme, concentrations and site of conducting NMR assignment experiments.

At CWRU, transcription reactions were optimized in individual trials (15). The RNA constructs were purified by 8–10% denaturing polyacrylamide gel electrophoresis depending on the size of the RNA and eluted in Tris-borate-EDTA (TBE) buffer. The RNA samples were desalted and adjusted to <20 μM concentration using Nanodrop (ThermoFisher Scientific). All desalted RNA samples were annealed in RNase-free water by heating the sample to 95°C for 2 min followed by snap cooling on ice for a minimum 15 min. The annealed samples were concentrated using a centrifugation filtration system (Amicon) and exchanged into NMR buffer. For isotope labeled samples, uniformly ^{13}C,^{15}N-labeled uridine (rUTP) and guanosine (rGTP) (Cambridge Isotope Laboratories), and unlabeled adenosine (rATP) and cytidine (rCTP) (Sigma-Aldrich) were used in the preparative transcriptions. Comparison of RNAs prepared in different laboratories yielded close-to-identical NMR spectra, demonstrating the value of the coordinated research approach within the NMR consortium Covid19-NMR.

NMR experiments

At BMRZ, NMR measurements were carried out on Bruker 600, 800, 900 and 950 MHz AVIIIHD or AV NEO spectrometers equipped with cryogenic triple-resonance HCN probes, on a Bruker 700 MHz AVIIIHD spectrometer equipped with a cryogenic quadruple-resonance QCI-P probe and a Bruker 800 MHz AVIII spectrometer equipped with a carbon optimized TXO cryogenic probe. All probes at BMRZ have z-axis pulsed-field gradient accessory. At CWRU, NMR experiments were carried out on Bruker AVIII 900/800 MHz NMR spectrometers equipped with cryogenically cooled HCN triple resonance probes and a *z*-axis pulsed-field gradient accessory. At Valencia, NMR spectra were acquired using a Bruker AVII 600 MHz NMR spectrometer equipped with a triple resonance TCI cryogenic probe. NMR experiments at the Weizmann Institute, were conducted on Bruker AVANCE NEO 1000 and 600 MHz NMR spectrometers equipped with 5-mm cryogenic triple-resonance HCN TCI probes and *x*, *y* and *z*-axis pulsed-field gradient accessory. NMR experiments in Stockholm were carried out on a Bruker 600 MHz AVIII spectrometer equipped with a cryogenic quadruple-resonance QCI probe and z-axis pulsed-field gradient accessory. Data were processed and analyzed using TOPSPIN 4.0 (Bruker BioSpin, Germany) or NMRpipe/NMRDraw. NMRFAM-SPARKY (16), CcpNmr (17) or CARA (http://www.cara.nmr-software.org/downloads/3-85600-112-3.pdf) were used for chemical shift assignment. Table 2 summarizes the NMR experiments conducted for all RNAs.

The conducted NMR experiments yielded the following information (18): ^{1}H,^{1}H-NOESY experiments (Table 2, #1, #7) correlate signals from RNA protons through space within up to 5.0 Å spatial proximity. In A-form RNA helices, sequential imino protons are within this distance and NOESY spectra thus provide correlation between consecutive base pairs within A-helical regions of the RNA, both to the next nucleotide within the sequence as well as across strands. Uridine and guanosine imino protons can be distinguished by their characteristic ^{15}N chemical shifts in the ^{1}H,^{15}N-TROSY experiment (Table 2, #2), where guanosine imino protons in canonical G-C base pairs have a chemical shift between 145 and 150 ppm (δ ^{15}N) and uridine imino protons in canonical A-U base pairs between 160 and 164 ppm (δ ^{15}N). Non-canonical U-U or G-U base-pairs result in upfield shifts of the corresponding imino proton resonances to 10–11.5 ppm/10.5–12.5 ppm (δ ^{1}H) and 141–145 ppm/155–160 ppm (δ ^{15}N) for guanosine and uridine residues, respectively. Further evidence for canonical base-pairing is provided by the HNN-COSY experiment (Table 2, #5), which correlates the donor nitrogen of the guanosine or uridine imino group with the acceptor nitrogen atom of the corresponding cytidine or adenosine residue via through-space scalar ^{2h}J-NN coupling. Non-canonical U-U or G-U base-pairs do not give rise to cross peaks in HNN-COSY spectra and can thus be distinguished from involvement into canonical base pairing. For unstable A-U base pairs, for which imino resonances are broadened beyond detection due to solvent exchange, we recorded long-range (lr) HNN-COSY experiments correlating the adenosine H2 proton to the uridine imino group across the hydrogen bond (Table 2, #10). The cytidine amino groups involved in base pairing can be detected in the ^{1}H,^{1}H-NOESY experiment due to their strong cross peaks to the partner guanosine imino protons and can be correlated to their respective N4 nitrogen via an exchange-optimized ^{1}H,^{15}N-HSQC (Table 2, #3). The ^{1}H,^{15}N-CPMG-NOESY experiment (Table 2, #4) complements imino-to-amino correlations in A-form helical structural regions of RNAs. With this standard set of experiments, secondary structure predictions for the above-mentioned SCoV2 RNAs were probed. Experimental secondary structures were thus determined for all 15 individual RNA elements and are described in the 'Results' section.

For sufficiently small RNAs (up to ∼35 nt), natural abundance ^{1}H,^{13}C HMQC experiments (Table 2, #11) were measured and analyzed to further assign aromatic H6/H8 and H5 protons and anomeric H1′ ribose protons. Assignment of these NMR signals is essential for the NMR sequential assignment procedure via ^{15}N-filtered NOESY experiments in regular A-form conformations (Table 2, #8). The ^{1}H,^{1}H-TOCSY experiment (Table 2, #9) provides a quick identification of intra-nucleobase H5-H6 cross peaks for pyrimidines in this region, reducing ambiguities resulting from poor signal dispersion observed for RNAs larger than ∼30 nts in general. Line shapes of these TOCSY cross peaks are also a good indication for dynamics exhibited by the respective nucleotide. The strongest signals are usually observed for highly flexible pyrimidine residues in loops. In addition, ribose H1′-H2′ cross peaks can be observed for the non-A-

Table 1. RNA constructs corresponding to individual SCoV2 *cis*-acting elements, designated according to their location position in the viral genome, together with the isotope labeling scheme and concentration of the samples, and site where the NMR experiments were conducted

RNA	BMRB ID*	Isotope labeling	Concentration [μM]	NMR experiments
5_SL1234		^{15}N	446	BMRZ
119 nts				
5_SL1	50 349	^{15}N	155	BMRZ
29 nts		$^{13}C,^{15}N$	650	
5_SL2+3	50 344	^{15}N	400	BMRZ, Catholic University of Valencia
32 nts		$^{13}C,^{15}N$	652	
5_SL4	50 347	^{15}N	775	BMRZ
44 nts		($^{13}C,^{15}N$ G, U)	400	BMRZ
		($^{13}C,^{15}N$ A, C)	500	BMRZ
			250	CWRU
5_SL5stem	50 340	^{15}N	700	BMRZ
69 nts				
5_SL5a	50 346	^{15}N	807	BMRZ, Karolinska Institute
33 nts		$^{13}C,^{15}N$	680	
5_SL5b+c	50 339	^{15}N	351	BMRZ, Weizmann Institute
37 nts		$^{13}C,^{15}N$	728	
5_SL5b		unlabeled	1200	BMRZ
25 nts				
5_SL5c		unlabeled	1500	BMRZ
12 nts				
5_SL6	50 351	(^{15}N A, C) ($^{13}C,^{15}N$ G, U)	600	BMRZ
46 nts		($^{13}C,^{15}N$ G, U)	250	CWRU
5_SL7	**	(^{15}N A, C, U)	550	BMRZ
50 nts		($^{13}C,^{15}N$ G)		
5_SL8	50 352	^{15}N	830	BMRZ, Weizmann Institute
63 nts				
5_SL8loop		$^{13}C,^{15}N$ G, U	411	BMRZ
31 nts				
att HP	**	$^{15}N,^{13}C$	90	Catholic University of Valencia
26 nts				
PK	50 348	^{15}N	700	BMRZ
69 nts				
3_SL1	50 342	^{15}N	766	BMRZ
72 nts				
3_SL2	50 343	^{15}N	80	BMRZ
31 nts		$^{13}C,^{15}N$	405	
3_SL3base	50 350	^{15}N	230	BMRZ
90 nts				
3_HVR		$^{15}N,^{13}C$ G, U	841	BMRZ
115 nts				
3_s2m	50 341	^{15}N	388	BMRZ
45 nts		$^{13}C,^{15}N$	596	
5′-geRNA		^{15}N	130	BMRZ
472 nts				
3′-UTR		^{15}N	180	BMRZ
337 nts				

*Chemical shift assignment will be continuously updated in BMRB (8).
**Deposition in progress/additional experiments required.

form 2′-endo-puckered ribose conformations in the spectral region of 4–6 ppm (δ ^{1}H).

Guanosine H8 protons and adenosine H8 and H2 protons are correlated with N7/N9 or N1/N3 nitrogens by a long-range (lr) $^{1}H,^{15}N$-HSQC experiment (Table 2, #6). H2 protons of base-paired adenosines identified in the $^{1}H,^{1}H$-NOESY experiment due to their strong cross peak to the partner uridine imino proton can thus be used to assign adenosine N1 and N3 atoms.

NMR experimental data were stored centrally and managed using the scientific data management system LOGS (19) (https://logs-repository.com). All experimental data can be downloaded at www.covid19-nmr.de.

RESULTS

All 20 RNA samples for NMR experiments exhibited high purity and adopted one homogeneous conformation without addition of Mg^{2+}, with the exception of the pseudoknot RNA, which formed dimeric species to a significant extent, as described previously for the PK of SCoV (3). However, in the absence of Mg^{2+} and at 283 K, the monomeric species dominated (Supplementary Figure S3) and NMR spectra were sufficiently resolved to analyze the imino protons (Figure 13). Thus, secondary structure could be characterized for the entire set of *cis*-acting RNAs chosen for this study. The NMR spectra of the isolated RNA elements were compared to TROSY spectra for the large 5′-geRNA

Table 2. A set of NMR experiments to delineate secondary structure

#	NMR experiments*	Sample utilized Solvents Groups detected	Experiment-specific parameter settings	MT	References
1	^{1}H,^{1}H-NOESY with jump-return water suppression	unlabeled or ^{15}N labeled Homonuclear NOEs between iminos and all other sites	NOE mixing time: 150 ms	30 h	(56)
2	^{1}H,^{15}N-BEST-TROSY	^{15}N labeled J-based heteronuclear H,N correlation for imino groups	Transfer delay $\Delta = 5.4$ ms to match ^{1}J(H,N) $\sim$ 90 Hz	1 h	(57,58)
3	^{1}H,^{15}N-HSQC	^{15}N labeled H,N correlation for amino groups	Relaxation and exchange optimized transfer delay $\Delta =$ 4.6 ms	1 h	(59)
4	^{1}H,^{15}N-CPMG-NOESY	^{15}N labeled ^{15}N-edited NOESY correlation to detect NOEs to fast exchanging protons	NOE mixing time: 150 ms	22 h	(60)
5	2D-BEST-TROSY HNN-COSY experiments	^{15}N labeled Through space J-based correlation of nitrogen atoms acting hydrogen bond donor and acceptor nitrogen	NN-Transfer of 30 ms	3 h	(58,61–63); in-house optimization
6	Long range ^{1}H,^{15}N-sfHMQC*	^{15}N labeled J-based correlation of purine N7/N9 nitrogens with H8 and adenine N1/N3 nitrogens with H2	Relaxation optimized transfer delay $\Delta = 20$ ms	2 h	(64)
7	Hadamard-encoded NOESY	unlabeled or ^{15}N labeled Homonuclear NOEs between imino sites and all other sites Emphasis on fast exchanging sites	NOE mixing time: 200 ms	2 h	(65)
8	^{13}C,^{15}N-filter NOESY with WATERGATE water suppression	^{13}C, ^{15}N labeled X-filter in ω_1 and ω_2 to selecting ^{12}C and ^{14}N bound protons, e.g. H2, H6, H8	NOE mixing time: 250 ms	24 h	(66–68)
9	^{1}H,^{1}H-TOCSY with Excitation sculpting water suppression	unlabeled or ^{15}N labeled J-based correlation of H5 and H6 in pyrimidine nucleobases	TOCSY mixing time: 30 ms	4 h	(69,70)
10	BEST-long range HNN-COSY*	^{15}N labeled	Long range correlation adenine H2 to base-paired uridine	8 h	in-house improvements**
11	^{1}H,^{13}C sfHMQC	unlabeled or ^{15}N labeled	For aromatic C's at natural abundance	5 h-24 h	(64)

MT is measurement time.

*marks experiments that can be conducted in D_2O, but were conducted in 95% H_2O, 5% (v/v) D_2O to minimize the need for the production of additional samples.

**pulse sequence and parameter sets for in-house optimized experiments can be obtained upon request and data sets can be downloaded at covid19-nmr.de. The first eight experiments have been conducted for all RNA constructs, additional experiments collected for a subset of RNAs are listed.

and 3′-UTR and to DMS footprinting data obtained from both regions, respectively. DMS footprinting confirmed the presence of well-defined SLs in both regions, corroborating the validity of our approach to analyze the structure of *cis*-regulatory elements in isolation (Figure 2A and B). NMR spectra of both large RNA constructs show extensive stable secondary structure formation in the imino proton region, with differential dynamic behavior of several structural motifs (Supplementary Figure S24A and B), similarly demonstrating independent folding of single SLs. In the following, the secondary structure determination by NMR of individual SLs is described and compared to the obtained DMS data. Where possible, SL constructs were stabilized by terminal G-C base pairs to facilitate *in-vitro* transcription and structure determination and designed to give one defined structure in prior *in-silico* predictions in accordance with phylogenetic conservation. Structure predictions using RNAstructure for the individual SLs (20) and pKiss for the PK (21) are shown in Supplementary Figure S5.

5_SL1 5_SL1 spans nucleotides (nts) 7–33 in the 5′-UTR of the SCoV2 genome. In SCoV, interaction of nsp1 with SL1 has been shown to be crucial to protect viral RNA from degradation (22). NMR studies on SL1 derived from MHV (mouse hepatitis virus) identified a bulged-out adenine corresponding to adenine 27 in SCoV2. Furthermore, the lower-part of SL1 has been shown to be structurally labile, which seems to be crucial for virus propagation in MHV, where SL1 might mediate the cross-talk between the 5′- and the 3′-UTR during viral replication (23). Computational predictions of the secondary structure of 5_SL1 yielded a single lowest-energy conformation with an asymmetrical bulge formed by nucleotides 12, 27 and 28 (Supplementary Figure S6). By assignment of imino resonances, the structural predictions could be confirmed and all observable slow exchanging imino resonances could be assigned from analysis of a ^{1}H,^{1}H-NOESY at 283 K (Figure 3A). Generally, we observe less intense NOE diagonal peaks for residues neighboring the bulge region on the

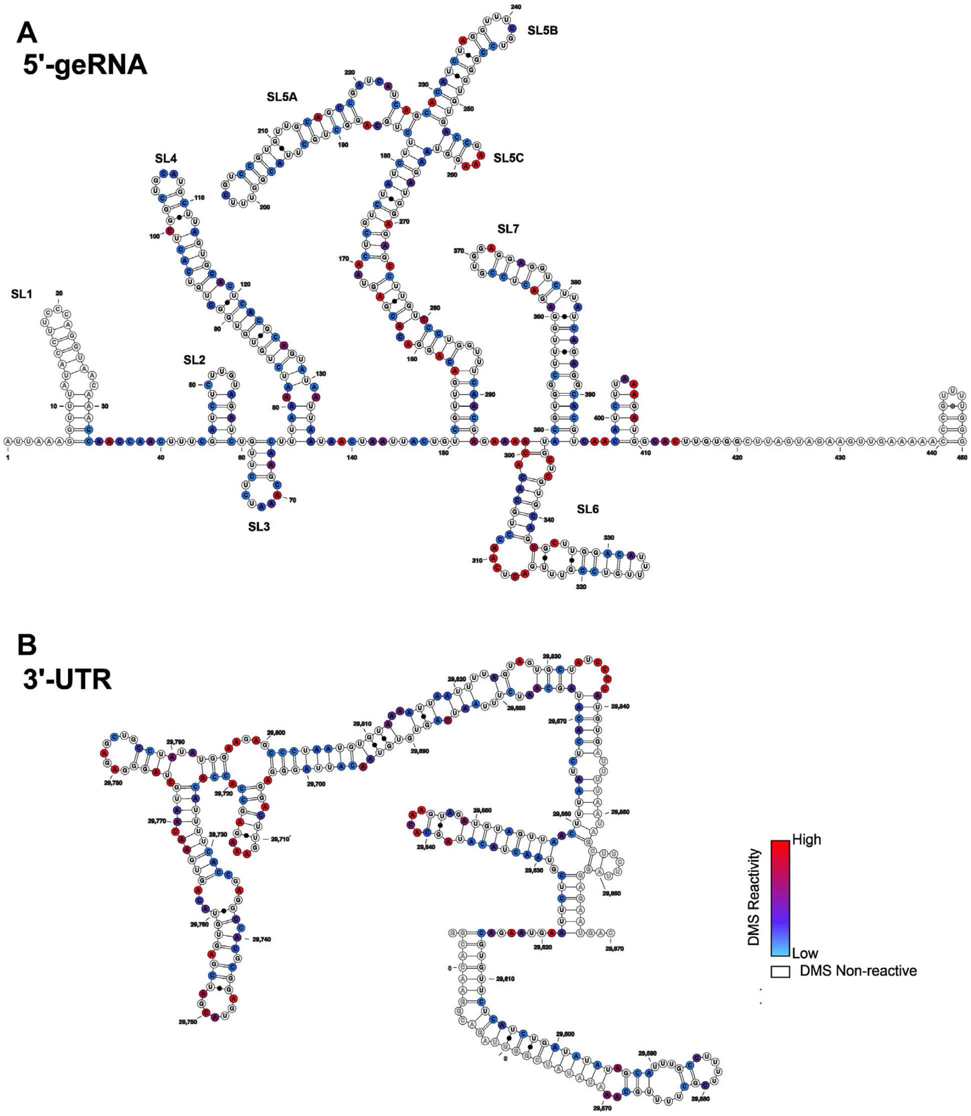


Figure 2. Secondary structures of the SCoV2 5′-genomic end and 3′-UTR as determined by DMS-MaPseq. Normalized DMS reactivity indices used to fold the secondary structures of (**A**) SCoV2 5′-genomic end and (**B**) SCoV2 3′-UTR are superimposed with color codes scaled from the highest reactivity set to 1 (red) to the lowest set to 0 (blue). G and U nucleotides are rendered as white and sites that overlap with sequencing primers are rendered transparent.

loop helix (U13, U25) compared to those neighboring it on the terminal helix (U10, U11). The terminal helix was stabilized by one additional G-C base pair to facilitate *in-vitro* transcription. Thus, for instance, the imino resonance of U11 is assignable, while the U13 imino resonance cannot be detected at all. The imino proton-based secondary structure determination was consistent with results from ^{15}N-correlated experiments (Figure 3B and C). Additionally, aromatic and amino resonances of base-paired A and C residues (H2 and H5/H6 resonances) could be assigned (Supplementary Figure S6). Most of the nucleotides forming 5_SL1 are within the primer binding region in DMS footprinting experiments, rendering this SL invisible for DMS analysis (Supplementary Figure S6E).

5_SL2+3 5_SL2 is the most conserved structure in the 5′-UTR (24). Notably, SL2s from different coronaviruses can functionally replace each other, even across different genera (25). SL2 (nts 45–59) of the 5′-UTR is identical to SCoV SL2. It is thus very likely that it forms the same CUYG-looped short hairpin structure (26). SL3 contains the transcription regulatory sequence (TRS) essential for the synthesis of sg mRNAs by discontinuous transcription (25). Among Betacoronaviruses, SL3 (nts 61–65 in SCoV2) is predicted to be stably formed only for a subset of species, SCoV among them (4). Invariably, whether as part of a hairpin or not, the core sequence (CS-L, CUAAAC) is single-stranded (4). Unwinding of SL3 by the N protein might be necessary for TRS function (27). As SL2, SL3 is identical in sequence between SCoV and SCoV2 (Supplementary Figure S1). In the present study, we investigated a sequence spanning nts 45–75 that contained SLs 2 and 3, the secondary structure of which is shown in Figure 4. The presence of the CUYGU pentaloop in SL2, including a C50:G53 pair and an extrahelical U54 was confirmed by the

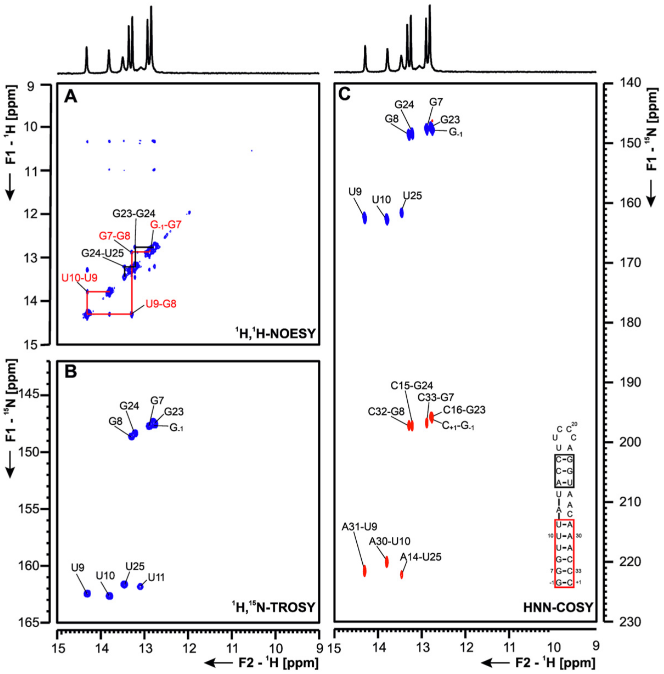


Figure 3. (**A**) ^{1}H,^{1}H-NOESY, (**B**) ^{1}H,^{15}N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlation of the 5′-genomic end construct 5_SL1 encompassing nts 7–33. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using the genomic numbering. Imino-proton correlations in (A) between consecutive base pairs are shown in different colors. Included in (C) is the experimentally observed secondary structures of 5_SL1 with genomic numbering. Additional closing base pairs are annotated with '±x'. Colors of boxes are according to the correlations in (A).

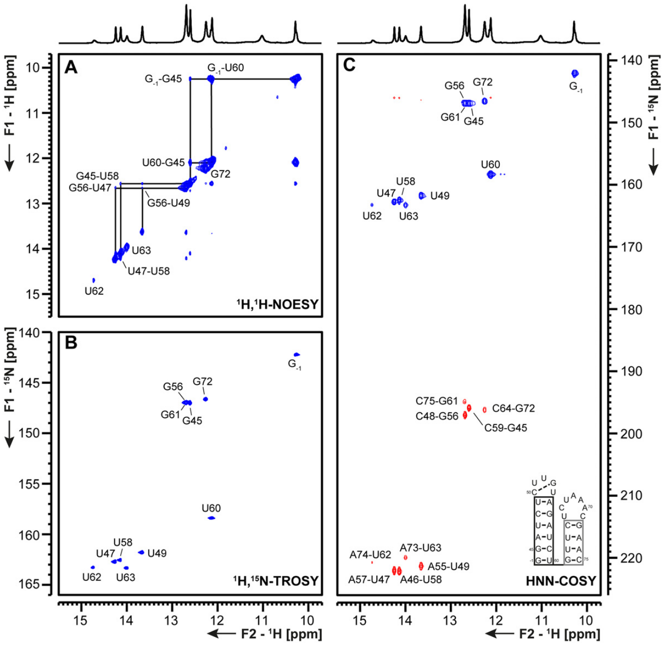


Figure 4. (**A**) ^{1}H,^{1}H-NOESY, (**B**) ^{1}H,^{15}N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlation of the 5′-genomic end construct 5_SL2+3 encompassing nts 45–75. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using genomic numbering. Imino-proton correlations in (A) between consecutive base pairs are shown in black. Included in (C) is the experimentally observed secondary structure of 5_SL2+3 with genomic numbering. Additional closing base pairs are annotated with '±x'. The black box represents the correlations in (A). The gray boxes marks base pairs which were not assigned based on imino-to-imino correlations.

NMR data, which were also consistent with the presence of a UCUAAAC heptaloop closing the SL3 hairpin and encompassing the CS-L (Figure 4 and Supplementary Figure S7). Interestingly, the two stems in this construct, 5_SL2 and 5_SL3 show different stabilities at 283 K. This difference in stability can be deduced from the difference in imino resonance linewidths for the base pairs of SL2 versus SL3, in line with the requirement of the viral transcription machinery to unfold the leader TRS located in SL3. This differential stability is even more pronounced at 298 K, where the U62 and U63 imino proton resonances of SL3 are broadened beyond detection (Supplementary Figure S8). In this regard, the leader TRS of the Alphacoronavirus TGEV (transmissible gastroenteritis coronavirus) was also found to occupy the apical loop of a pseudo-stable hairpin by NMR spectroscopy (25,28). In agreement with the NMR data, DMS footprinting showed stable base pairs in the stem of SL2 and a loop G-C base pair indicated by the very low reactivity of C50. However, the lower stability of SL3, evident from NMR, could not be detected by DMS footprinting (Supplementary Figure S7). Thus, we tested if Mg^{2+} ions, present during DMS treatment, but absent in the NMR buffer, could modulate the stability of SL3. Strikingly, in the presence of 3 mM Mg^{2+}, the SL3 is significantly stabilized, illustrated by an increase of imino proton intensities for U62 and U63 (Supplementary Figure S8). This stabilizing effect of Mg^{2+} is even evident at 298 K, where the U62 and U63 imino proton resonances are broadened beyond detection in the absence of Mg^{2+}, but become clearly observable in the presence of Mg^{2+} (Supplementary Figure S8). Thus, ionic conditions strongly influence the stability of SL3 and might thereby affect TRS function *in vivo*.

5_SL4 5_SL4 (nts 86–125) is structurally conserved among Betacoronaviruses and might function as a spacer element required for sg mRNA synthesis, mediating proper relative orientation of SLs 1–3 (29). SL4 carries a short upstream ORF (uORF), which is also conserved among Betacoronaviruses. Its function is still subject to speculation, since genetic pressure to select for the presence of a uORF could be observed, yet manipulations of this uORF that retain the SL4 RNA structure all yielded viable viruses (30). Interestingly, an additional short downstream hairpin structure predicted for MERS-CoV2 might also be present in SCoV2 (4,6), and is currently under investigation at BMRZ. Here, we present our results for 5_SL4 depicted in Figure 5. The NMR data showed the secondary structure of an SL interrupted by mismatches and capped by a non-canonical five nucleotide loop (Figure 5). All predicted base pairs could be detected, including a G-U base pair formed by residues G91 and U120 by the presence of imino signals in the non-canonical region (Figure 5B). The imino NOESY walk is in agreement with a continuous stem for SL4 with a looped-out U95 and a so far structurally incompletely characterized, but A-form helix compatible arrangement of the opposing residues C92 and C119 as well as

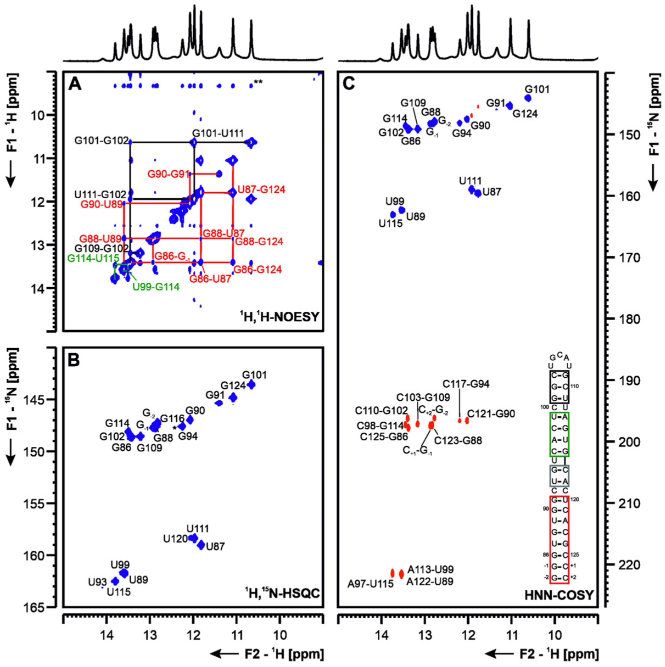


Figure 5. (**A**) ^{1}H,^{1}H-NOESY, (**B**) ^{1}H,^{15}N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlation of the 5′-genomic end construct 5_SL4 encompassing nts 86–125. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using the genomic numbering. Imino-proton correlations in (A) between consecutive base pairs are shown in different colors. Included in (C) is the experimentally observed secondary structure of 5_SL4 with genomic numbering. Additional closing base pairs are annotated with '±x'. Colors of boxes are according to the correlations in (A). Gray boxes mark base pairs, which were not assigned based on imino-to-imino correlations. The asterisk in (B) marks a signal which is visible at lower contour levels.

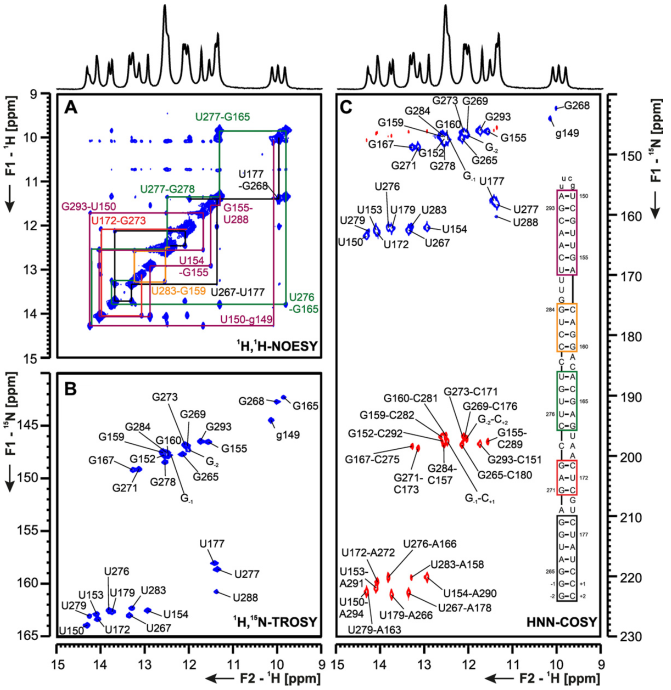


Figure 6. (**A**) ^{1}H,^{1}H-NOESY, (**B**) ^{1}H,^{15}N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlations of the 5′-genomic end construct 5_SL5stem encompassing nts 150–180 and 265–294. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using the genomic numbering. Sequential imino-proton correlations in (A) between neighboring base pairs are shown in colors according to the experimentally determined 5_SL5stem RNA secondary structure schematized as an inset in panel (C). NMR resonances arising from the tetraloop nucleotides are annotated with small letters and the two additional GC closing base pairs at the 5′- and 3′-termini are numbered with ±x, respectively. Colors of boxes are according to the correlations in (A). Note that the stem is depicted upside down with respect to its orientation as presented in the overall scheme of Figure 1.

C100 and U112. So far, the absence of imino resonances for residues of the apical loop, which contains two of the three nucleotides of the start codon of the uORF, precludes any conclusions about the structure of the loop. Starting from the assigned imino protons (Figure 5B), 12 of 14 cytidine amino groups, three out of five adenine amino groups and eight of 14 guanine amino groups were assigned (Supplementary Figure S9D). The cytidine amino groups served as starting point to assign the H5 resonances in a ^{15}N-filtered NOESY spectrum together with a ^{1}H,^{1}H-TOCSY experiment (Supplementary Figure S9B), which also allowed the connection of the H5 to the H6 of the same pyrimidine residue. Starting from the H5-H6 cross peaks and the adenine H2 resonances identified in the ^{1}H,^{1}H-NOESY and the lr-^{15}N-HSQC spectrum (Supplementary Figure S9A), the non-exchangeable aromatic C-H groups of all residues but the two uridine residues in the loop could be assigned in the aromatic-aliphatic walk in the ^{1}H,^{1}H-NOESY spectrum (Supplementary Figure S9C). DMS footprinting data suggest an elongated SL for SL4, ranging from nt 77 to 136.

SL5 (nts 149–294) displays more pronounced structural divergence among the Betacoronaviruses in comparison to the upstream elements, but some features are recurring: a four-helix junction connects the sub-structures SL5a, b, and c to a long-range base-pairing region forming an extended helical stem (31). The AUG start codon for ORF1a is integrated into a stable SL5 helical section, resulting in the 3′-part of SL5 being the N-terminal coding sequence of nsp1 (Figure 1). Sub-structures of SL5 have been associated with viral replication in MHV and BCoV (Bovine coronavirus) (32), but no conclusive model for a conserved functional role of SL5 in Betacoronaviruses is available. We divided SCoV2 SL5 into the three sub-structures 5_SL5stem (nts 150–180 and 265–294), 5_SL5a (nts 188–218) and 5_SL5b+c (nts 227–263), shown in Figures 6–8, respectively.

5_SL5stem For the SL5 core stem RNA element, we used the following construct design: after removal of all predicted SL sub-parts in the SL5 tree (i.e. a, b and c) the remaining core stem encompasses nts 150–180 on its 5′-end and 265–294 on its 3′-end, respectively. This design is fully in line with the suggested arrangements of the SL5 arms and, in particular, the stem structured detected in the DMS footprinting experiments (Figure 2A). Notably, the subgenomic AUG start codon is positioned directly downstream of SL5c, starting from nucleotide A266 (Figures 2 and 6). We decided to stabilize the resulting apical end of 5_SL5stem with two additional G-C base pairs in order to mimick the expected structural rigidity of the nearby four-way-junction and facilitate assignments of the initial base pairs. The 5_SL5stem basal end was stabilized with a *bona*

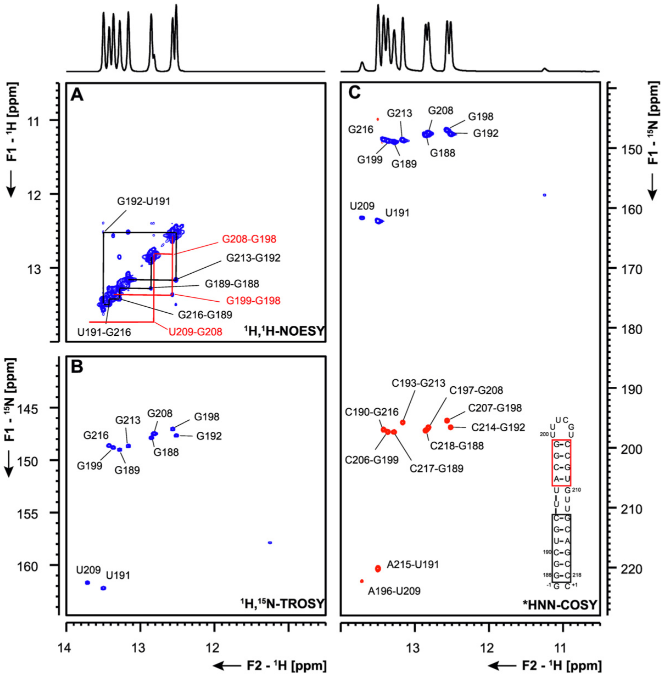


Figure 7. (**A**) ¹H,¹H-NOESY, (**B**) ¹H,¹⁵N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlations of the 5′-genomic end construct 5_SL5a encompassing nts 188–218. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using the genomic numbering. Imino-proton correlations in (A) between consecutive base pairs are shown in different colors. Included in (C) is the experimentally observed secondary structure of 5_SL5a with genomic numbering. The HNN-COSY spectrum in (C) was recorded at 298 K. Additional closing base pairs are annotated with '±x'. Colors of boxes are according to the correlations in (A).

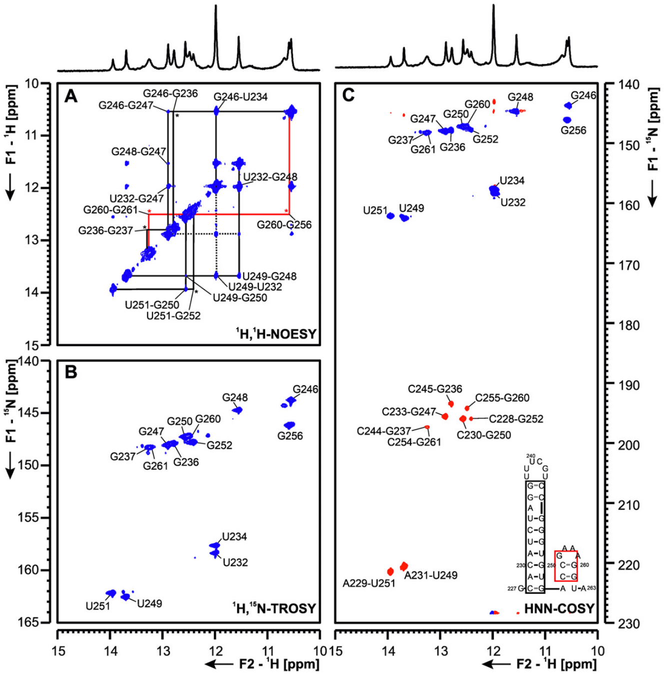


Figure 8. (**A**) ¹H,¹H-NOESY, (**B**) ¹H,¹⁵N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlation of the 5′-genomic end construct 5_SL5b+c encompassing nucleotides 227 to 263 at 283K. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using the genomic numbering. Imino-proton correlations in (A) between consecutive base pairs are shown in different colors, * denote crosspeaks visible at 275K, but not at 283K. Included in (C) is the experimentally observed secondary structure of 5_SL5b+c with genomic numbering. Additional closing base pairs are annotated with '±x'. The boxes are according to the correlations in (A).

fide UUCG tetraloop (33), which allowed transcription of the RNA from one DNA template and resulted in a sufficiently compact and homogeneously folded RNA, despite its relatively large size of 69 nt. For convenience, we display the 5_SL5stem in an upside-down arrangement with respect to the scheme in Figure 1 (Figure 6 and Supplementary Figure S10). Using NMR experiments #1 to #6 and #9 on a uniformly ^{15}N-labeled sample (Tables 1 and 2), we obtained assignments for all, except for one (U175), H-bonded imino ^{1}H,^{15}N pairs within the native sequence in agreement with the predicted secondary structure (Figure 6A and B; Supplementary Figures S10 and 11). We found NOEs that show the tight interaction of the tetraloop-stabilized g149 with the closing base pair A294-U150. However, we were only able to see the stabilizing effect of the UUCG cap at 283 K, but not at higher temperatures.

Apart from some spectral overlap in the central guanosine region (around 12.5 ppm), the 5_SL5stem RNA yielded spectra of excellent quality for initial assignments beyond the G/U iminos. We were also able to assign imino nitrogens of all base-pairing adenines (Figure 6C), supporting the underlying base pair pattern and RNA secondary structure. The well-resolved H2 proton chemical shift dispersion allowed us to obtain assignments for the majority of the adenine N3 resonances (Supplementary Figure S10). We could also assign a significant fraction of H-bonded adenine and cytidine amino groups, although a number of amino NMR signals overlap, which suggests that the secondary structure primarily adopts A-form conformation within consecutive stretches, typical for double-stranded RNAs.

The NMR-derived base-pairing pattern is confirmed by the reactivities obtained for the respective RNA stem in the DMS footprinting experiment (Figure 2 and Supplementary Figure S10). The latter data reflect the increase in reactivity along with the bulges compared to the H-bonding within duplexed stretches (see direct comparison of data from both methods in Supplementary Figure S10D). Notably, while an increase in DMS reactivity at A-U closing base pairs adjacent to bulges is explained by a lower average stability of the H-bonds, we also found the inner-helical adenosines 157 and 166 to exhibit relatively high DMS reactivity, which is in contrast to their apparently stable duplex character shown by the NMR data. To examine the role of different experimental conditions between NMR (283 K and no Mg^{2+}) and DMS footprinting (310 K and 3 mM Mg^{2+}), we undertook both an NMR-based titration of 5_SL5stem with Mg^{2+} as well as a temperature comparison of nitrogen correlation spectra between 283 K and 298 K (Supplementary Figure S11). Our data show that the locally confined increase in DMS reactivity can well be explained by the higher temperature. Of note, the overall comparison between the two buffer conditions reveals identical spectral quality and thus underlines the complementary strength of the two methods to fully probe RNA secondary structure. Consequently, we found the 5_SL5stem part of the SCoV2 RNA to adopt the global secondary structure as suggested in Figures 1 and 2.

5_SL5a based on secondary structure prediction, the SL5a construct is composed of two helical stems that are separated by a three nucleotide U-rich bulge (Figure 7). By NMR spectroscopy, we were able to confirm the presence of both stem regions, while the resonances of the two U nucleotides participating in canonical A-U base pairs (U191 and U209) served as starting points for the sequential imino-walk indicated in Figure 7A. Additionally, we detected a weak U-imino resonance (CS ^{1}H: 11.3 ppm and ^{15}N: 157.8 ppm, Figure 7A and B), which is involved in a non-canonical nucleobase interaction. This U shows correlations to the C193-G213 base pair and is thus either part of the U-rich bulge or represents U195 if the other uridines of the bulge are flipped out and solvent exposed. However, at this point an unambiguous assignment is not possible and also in line with the available SHAPE data (34,35), no imino proton signal for G210 is detected at room temperature. Furthermore, the stem is closed by a UUUCGU loop. While imino resonances remain elusive for this loop, the ^{1}H,^{15}N-HSQC spectrum of the amino groups show characteristic resonances for a UUCG-tetraloop (Supplementary Figure S12) (33). This suggests that the UUUCGU loop might adopt a conformation similar to the UUCG tetraloop. All observations by NMR spectroscopy are in good agreement with the DMS footprinting data.

5_SL5b+c key to the assignment of 5_SL5b+c (Figure 8 and Supplementary Figures S13-14) was the use of low temperature (275 K) and the comparison of the construct to the single hairpins 5_SL5b and 5_SL5c. In the double hairpin construct 5_SL5b+c, only the lower part of the SL5b stem (nts 229–234, 246–251) gave rise to imino-imino cross peaks in the NOESY at 283 K, while imino-imino correlations for the upper two base pairs closing the UUUCGU-loop, the closing base pair for the stem of 5_SL5b and for the base pairs of 5_SL5c (nts 253-263) only appear at 275 K (Supplementary Figure S14). Stem SL5c is closed by a GAAA-tetraloop (36). In these tetraloops, the H1 of the G is protected by hydrogen bonding to the phosphate of the last adenosine in the loop. In line with this, we observe an additional cross peak in the TROSY (Figure 8B) in the non-canonical region, which arises from G256, however, due to overlap with G246 and ambiguities in the amino region, an assignment was only feasible by comparison of the 5_SL5b+c construct to the single hairpins of 5_SL5b and 5_SL5c (Supplementary Figure S14). Data from DMS footprinting are in line with our observations at 275 K, while we observe more dynamics at higher temperature, suggesting stabilization of SL5b and c within the full length SL5.

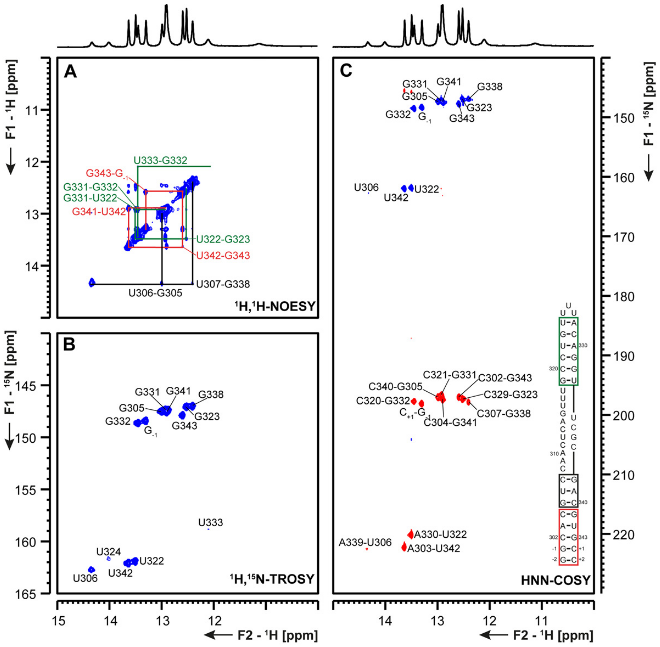


Figure 9. (**A**) ^{1}H,^{1}H-NOESY, (**B**) ^{1}H,^{15}N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlation of the 5′-genomic end construct 5_SL6 encompassing nts 302–343. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using the genomic numbering. Imino-proton correlations in (A) between consecutive base pairs are shown in different colors. Included in (C) is the experimentally observed secondary structure of 5_SL6 with genomic numbering. Additional closing base pairs are annotated with '±x'. Colors of boxes are according to the correlations in (A).

5_SL6 5_SL6 comprises nts 302–343 of SCoV2 and is located in the 5′-region of ORF1a coding for nsp1. For a number of members of the Betacoronavirus class, including BCoV, MHV and several human coronaviruses (HCoVs), SL6 was shown to be conserved, but its precise function in viral replication remains unclear (37). Mutational studies leading to destabilization of SL6 resulted in reduced virus viability in MHV. However, this effect was attributed to a reduction of nsp1 protein levels (4,38). A recent study identified the asymmetrically bulged region of SCoV2 SL6 as one major binding site for the N protein (34). Here, we study SCoV2 5_SL6 that comprises nts 302–344 of SCoV2 extended by two G-C base pairs. (Figure 9). We conducted resonance assignment on a sample containing ^{15}N-labeled A and C nucleotides and ^{13}C,^{15}N- labeled G and U nucleotides. We could assign the observable resonances of all exchangeable imino- and C amino-protons in this RNA (Figure 9 and Supplementary Figure S15). Based on these assignments, we can delineate two base-paired regions in 5_SL6. The first paired region is comprised of a stem of eight consecutive canonical Watson–Crick base pairs including the 5′- and 3′-termini. The second base-paired region consists of six base pairs including a terminal G-U base pair, spanning from G319-U324 and A328-U333. This second paired region is capped by a triple-U loop. Both base-paired regions are separated by an asymmetric bulge of 11 and 4 nts. These results were in good agreement with DMS footprinting data. Here, the previously observed asymmetric A-rich bulge showed a high reactivity, whereas both stem regions were well protected from DMS (Supplementary Figure S15).

5_SL7 and 5_SL8 The presence of SL7 (nts 349–394) was confirmed by RNA structure probing for MHV, and similar motifs were found by RNA structure prediction in BCoV and SCoV (37). For SCoV and SCoV2, SL8 (nts 413–471) is predicted as a strikingly large hairpin-structure compared

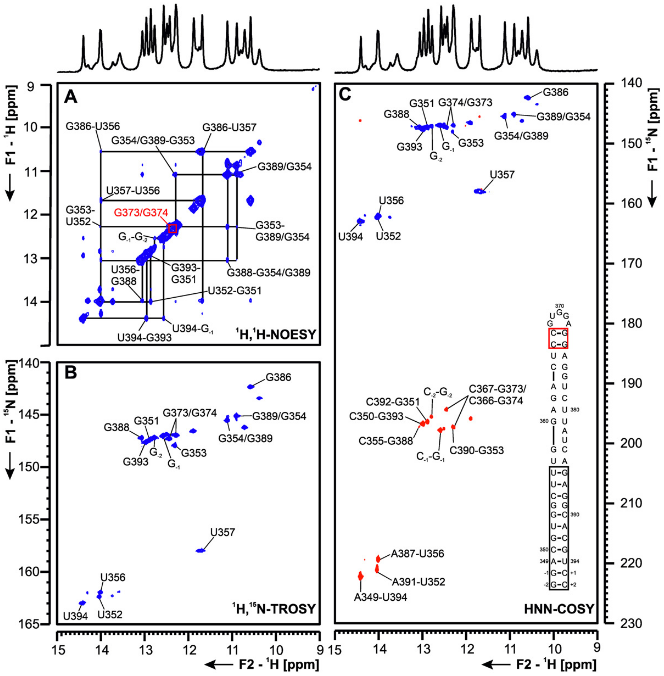


Figure 10. (**A**) ^{1}H,^{1}H-NOESY, (**B**) ^{1}H,^{15}N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlation of the 5′-genomic end construct 5_SL7 encompassing nts 349–394. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using the genomic numbering. Imino-proton correlations in (A) between consecutive base pairs are shown in different colors. Included in (C) is the experimentally observed secondary structure of 5_SL7 with genomic numbering. Additional closing base pairs are annotated with '±x'. Colors of boxes are according to the correlations in (A).

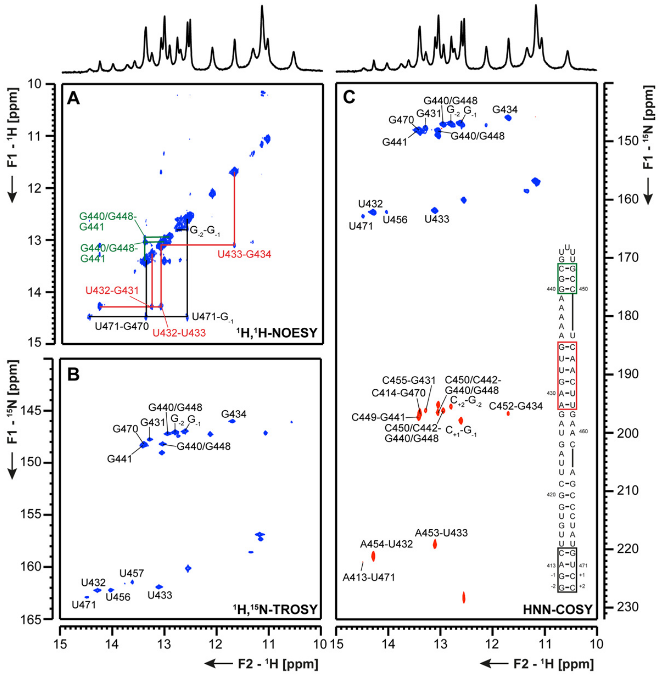


Figure 11. (**A**) ^{1}H,^{1}H-NOESY, (**B**) ^{1}H,^{15}N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlation of the 5′-genomic end construct 5_SL8 encompassing nts 413–471. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using the genomic numbering. Imino-proton correlations in (A) between consecutive base pairs are shown in different colors. Included in (C) is the experimentally observed secondary structures of 5_SL8 with genomic numbering. Additional closing base pairs are annotated with '±x'. Colors of boxes are according to the correlations in (A).

to other Coronaviruses, where it is mostly absent (4). We investigated 5_SL7 and 5_SL8 separately as shown in Figures 10 and 11.

5_SL7 The 50 nt RNA 5_SL7 is subdivided into four different helical sections, bridged by a G-G mismatch and two bulges of one (G377) or two (A382, U383) nucleotides. This general secondary structure, as predicted by RNAstructure and mfold [Supplementary Figure S5 and (39)], could be confirmed by assignment of imino proton resonances in ^{1}H,^{1}H-NOESY spectra (Figure 10). The terminal stem between G-2 and the characteristic G386-U357 wobble base pair at ^{1}H chemical shifts of 11.68 and 10.57 ppm (δ ^{1}H) was unambiguously assigned. For all assigned U imino protons, their corresponding A H2-aromatic protons were assigned (Supplementary Figure S16). The terminal helical stretch comprising two G-C base pairs following one A-U and an additional G-C base-pair was assigned by comparison to the same element present in 5_SL8 (Figure 11). The central unusual G-G-motif gives rise to two non-canonical imino proton resonances, which cannot be unambiguously assigned to one of the two Gs and may result either from stacking or formation of an unusual base pair. In the middle part of 5_SL7, no imino protons could be assigned in the ^{1}H,^{1}H-NOESY spectra, most likely due to an increased instability of this section compared to the remainder of the base pairs. However, two additional, yet unassigned G-C base pairs are detected in the HNN-COSY (Figure 10). The two loop-closing G-C-base pairs were assigned due to their uniqueness. Its unreactive nature was proposed in DMS footprinting. According to DMS footprinting data these base pairs are stable. Further, DMS footprinting data did not detect a higher flexibility of the central region of the SL. Additional insights into the conformation and dynamics of this part of the RNA will include the investigation of shorter RNA constructs.

5_SL8 For 5_SL8, overall 11 imino resonances were assigned unambiguously (Figure 11 and Supplementary Figure S17). With a shortened construct of the upper part of the RNA (5_SL8loop, Table 1) the three consecutive G-C base pairs were assigned (Figure 11, green), while the assignment of the two flanking guanosines, G440 and G448 remains ambiguous. Furthermore, it can be concluded from the HNN-COSY spectrum that a Hoogsteen A-U base pair might be present as the hydrogen bond acceptor nitrogen chemical shift is at ∼230 ppm, which is characteristic for an adenine N7 resonance. Moreover, we detected at least six additional imino protons that are involved in non-canonical structural elements. This NMR experimental fingerprint is not consistent with predictions of the predicted lowest energy structure (Supplementary Figure S5) and will be subject to further studies. As nearly the complete sequence of 5_SL8 constitutes the primer site for DMS footprinting no further DMS analysis was possible.

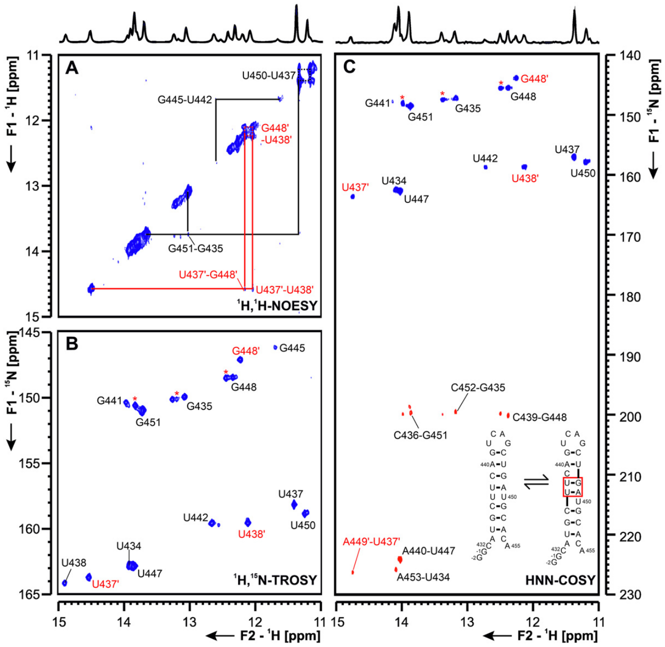


Figure 12. (**A**) ^{1}H,^{1}H-NOESY, (**B**) ^{1}H,^{15}N-TROSY and (**C**) HNN-COSY spectrum of the SARS-CoV-2 attenuator hairpin encompassing nts 13 432 to 13 455. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using genomic numbering, shifted for convenience by 13k nts from 5′. Imino-proton correlations in (A) between consecutive base pairs are shown in black and red for the two present exchanging conformations. Included in (A) is the experimentally observed equilibrium between secondary structures of the attenuator. Additional closing basepairs are annotated with '±x'. Colors boxes are according to the correlations in (A). Asterisks indicate secondary shifts due to conformational exchange.

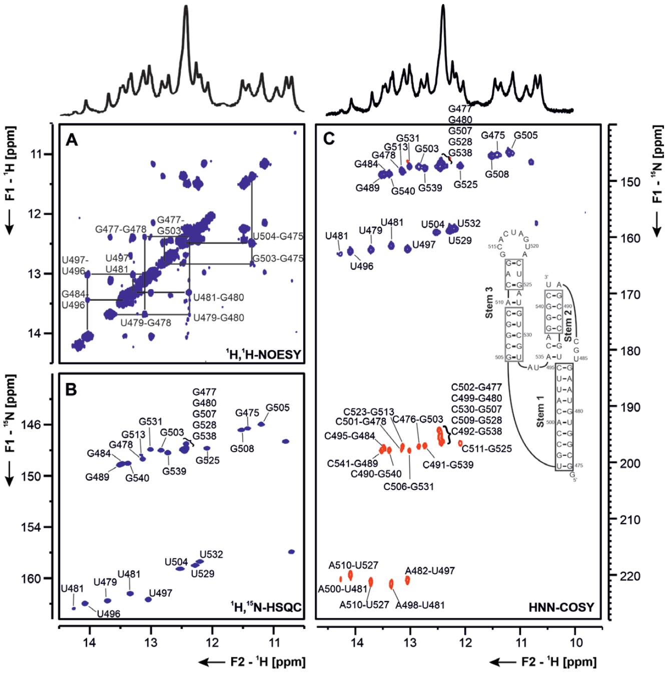


Figure 13. (**A**) ^{1}H,^{1}H-NOESY, (**B**) ^{1}H,^{15}N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlation of the genomic-central located PK construct encompassing nts 13 475–13 542. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using the genomic numbering, shifted for convenience by 13 000 nts from 5′. Imino-proton correlations in (A) between consecutive base pairs are shown in different colors. Included in (C) is the experimentally observed secondary structures of PK with genomic numbering, shortened by 13k nts from 5′. The secondary structure according to (C) is shown with genomic numbering from SCov2. Colors of boxes are according to the correlations in (A). Gray boxes mark base pairs, which were not assigned based on imino-to-imino correlations.

ORF1a/b frameshifting region

Att HP An attenuator hairpin (att HP) located immediately upstream of the slippery site and the PK responsible for programmed ribosomal frameshifting (PRF) has been reported to regulate PRF function by reducing the activity of the PRF PK (40). Similar to the attenuator hairpin of SCoV, the att HP of SCoV2 contains a 10-nt palindrome of unknown function comprising nts 13 441–13 450 (Figure 12 and Supplementary Figure S18). Indeed, we detected homodimerization for this construct in native polyacrylamide gels containing Mg^{2+}. The SCoV2 att HP exhibits five nt variations relative to SCoV that predict it to be significantly less stable (41). The NMR data indicate that the folding of the att HP of SCoV2 is substantially different to that proposed for the SCoV attenuator, as it contains an eight base pair stem including an intrahelical U-U mismatch, instead of the 10 bp stem predicted for the SCoV hairpin. In addition, the SCoV2 att HP does not expose the palindrome sequence in its apical region (Figure 12). The relative instability of the SCoV2 attenuator is reflected by the detection of possible alternative conformations in the central part of the SL in the NMR spectra, containing either the predicted, adjacent U13 437:U13 450 and U13 438-A13 449 bp, or adjacent U13 437-A13 449 and U13 438-G13 448 bp, with likely extrahelical U13 450 and C13 439 (Figure 12). At the current state of investigation, the assignments of these alternative conformation remains tentative and further constructs will be investigated to clarify the topology of these interesting RNA dynamics.

PK The three-stemmed pseudoknot (PK, nts 13 475–13 542, Figure 13) structure controlling the ribosomal frameshifting during ORF1a/b translation (42) has a unique fold in SCoV and SCoV2 (43,44). While in other Coronaviruses, and in many ssRNA viruses in general, this RNA element forms a canonical H-type pseudoknot fold, SCoV and SCoV2 PKs include a third SL. This SL harbors a 6-nts palindromic sequence allowing for homodimerization. Mutation of this sequence resulted in lower frameshifting efficiency and altered SCoV growth kinetics *in vivo* (3). The structure of SCoV PK has been thoroughly investigated before (3,45) and with only one nucleotide changed from SCoV to SCoV2 (C13 533A), assignment of the monomeric form of the PK, once prepared in NMR-suitable purity grade and quantity (∼70% monomer, 8 mg), was straightforward. The presence of the three predicted stem regions could be confirmed in ^{1}H,^{1}H-NOESY spectra combined with the ^{1}H,^{15}N-HSQC spectrum of the imino region and the HNN-COSY spectrum (Figure 13).

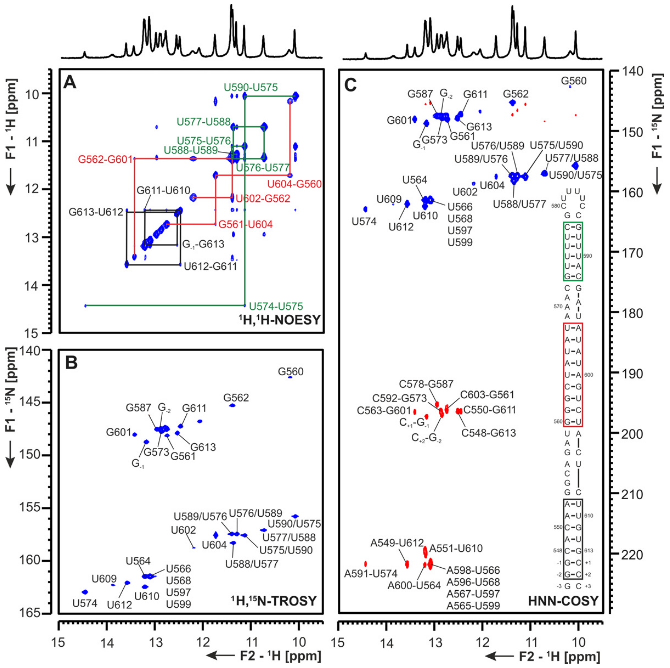


Figure 14. (**A**) ^{1}H,^{1}H-NOESY, (**B**) ^{1}H,^{15}N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlation of the 3′-genomic end construct 3_SL1 encompassing nts 29 548–29 613. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using the genomic numbering, shifted for convenience by 29 000 nts from 5′. Imino-proton correlations in (A) between consecutive base pairs are shown in different colors. Included in (C) is the experimentally observed secondary structure of 3_SL1 with genomic numbering, shortened by 29k nts from 5′. Additional closing base pairs are annotated with '±x'. Colors of boxes are according to the correlations in (A).

3′-UTR

The conserved regulatory RNA sequences necessary for virus replication at the 3′-genomic end of Coronaviruses are exclusively found in the untranslated region. The 5′-most *cis*-regulatory element is a large, bulged SL (3_SL1, Figure 1), containing a 3′-sequence that forms either the lower part of the SL or base-pairs with a single-stranded sequence within the loop of a downstream, smaller SL (3_SL2, Figure 1). Thus, the SL1-SL2 element is a putative RNA switch, associated with regulation of replication (46,47). Furthermore, single-stranded regions flanking 3_SL2 form long-range interactions with the 3′-most part of the genome in all Betacoronaviruses, associated with nsp8 binding in replication initiation (46).

The downstream part of the 3′-UTR is called hypervariable region (HVR) and affects viral pathogenicity (48). The HVR contains the highly conserved s2m motif in Betacoronaviruses, including SCoV, and also the non-related Astroviruses (Rfam ID: RF00164; (49)). Precisely within this RNA element, a single-nucleotide substitution from SCoV to SCoV2 results in dramatic changes of the predicted secondary structure (Supplementary Figure S5), rendering s2m a very interesting subject for high-resolution structural characterization by NMR spectroscopy.

Thus, we divided the 3′-UTR into the single elements 3_SL1 (nts 29, 548-29, 614), 3_SL2 (nts 29, 630-29, 656), 3_SL3base (nts 29, 620–29, 671 fused to 29 849–29 870) and 3_s2m (nts 29, 728–29, 768), reflecting the most important structural elements in the 3′-UTR (Figure 1).

3_SL1 For 3_SL1, secondary structure prediction tools consistently proposed a long helical stem, interrupted by several small bulges and a larger symmetrical bulge composed of six uridine residues in the upper part of the stem. For the 3_SL1 RNA construct, we could assign the imino protons by analyzing cross peaks in a ^{1}H,^{1}H-NOESY experiment conducted on an unlabeled RNA and ^{1}H,^{15}N-TROSY-and HNN-COSY experiments (Figure 14). In addition to the predicted A-helical parts, we could observe three consecutive non-canonical U-U base pairs in the upper part of 3_SL1. These three base pairs give rise to six well-resolved imino proton resonances in the spectral area typical for U imino protons involved in G-U and U-U base pairs (δ ^{1}H: 10–12.5 ppm, δ ^{15}N: 155–160 ppm) (50). Base pairs U29 564-A29 600 to A29 569-U29 695 are symmetrical, resulting in nearly complete chemical shift degeneracy for the middle four base pairs of this helical stretch. DMS footprinting analysis confirms the presence of these base pairs, showing low reactivity of the respective adenosine residues (Supplementary Figure S20).

3_SL2 3_SL2 is the second hairpin element in the 3′-UTR with a rather large loop sequence (11 nts) that is complementary to the upstream 3′-most sequence of the bulged SL (3_SL1). Imino proton spectra show that the A-helical part of 3_SL2 is composed of nine base pairs (Figure 15 and Supplementary Figure S20). The loop-closing U-A base pair is invisible, likely due to an enhanced life-time of the open state that promotes fast solvent exchange of the imino proton. In contrast to computational secondary structure predictions, no stable base pairs between loop nucleotides are observed (Supplementary Figure S20). In line with NMR data, DMS footprinting showed a high reactivity of the adenosine of the loop-closing U-A base pair. In contrast, DMS data showed only partial reactivity in the loop nucleotides suggestive of a confined conformation (Supplementary Figure S20).

3_SL3base 3_SL3base represents the long-range RNA interactions between the 3′-end of the genome and the single-stranded regions flanking 3_SL2, with the complete HVR deleted and replaced by a stable UUCG tetraloop. In the NMR spectra of this construct we find three distinct base paired regions arranged in a three-way junction. The first paired region corresponds to a stretch of 4 bp closing the 5′- and 3′-end of the construct. At both sides five additional nucleotides are found that do not form persistent interactions that would lead to exchange protected imino proton signals (Figure 16). The second paired region is found in the SL part of the molecule that is identical to 3_SL2, except for the stabilizing G-C base pairs at the end of the construct introduced for *in-vitro* transcription. In 3_SL3base a stretch of three Watson–Crick base pairs from A29 634-A29 636 and U29 650-U29 652 plus the additional base pair between C29 632 and G29 654 is formed, the other interactions that were mapped in the smaller and more stabilized 3_SL2 could not be detected (compare Figure 15 and Figure 16). In line with 3_SL2, in the 3_SL3base construct neither the loop nucleotides nor the A-U closing base pair exhibited detectable imino-proton resonances and can therefore

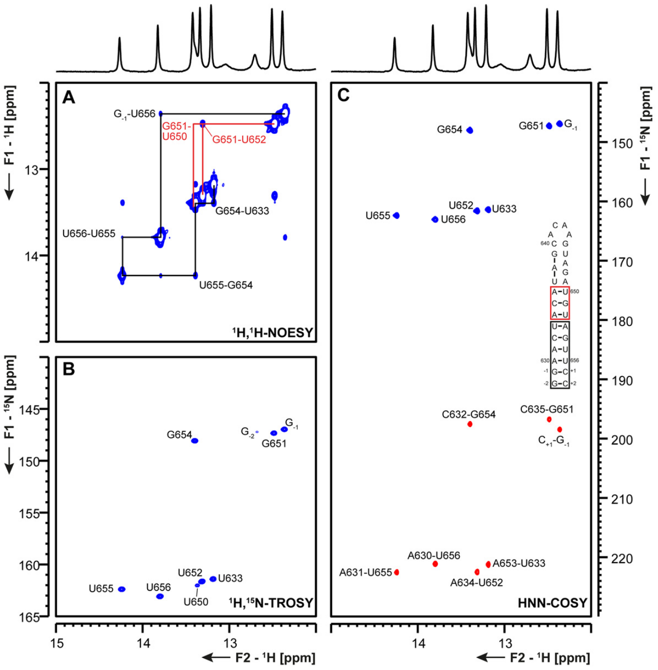


Figure 15. (**A**) ^{1}H,^{1}H-NOESY, (**B**) ^{1}H,^{15}N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlation of the 3′-genomic end construct 3_SL2 encompassing nts 29 630–29 656. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using the genomic numbering, shifted for convenience by 29 000 nts from 5′. Imino-proton correlations in (A) between consecutive base pairs are shown in different colors. Included in (C) is the experimentally observed secondary structure of 3_SL2. Additional closing base pairs are annotated with '±x'. Colors of boxes are according to the correlations in (A).

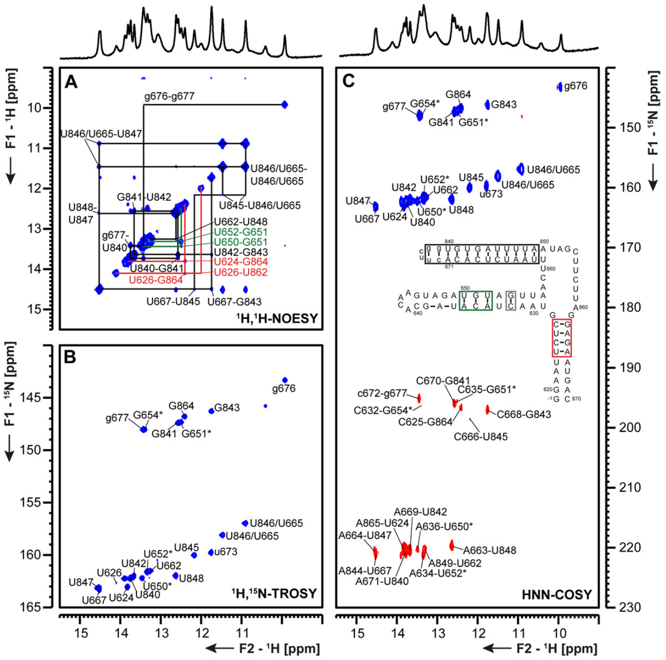


Figure 16. (**A**) ^{1}H,^{1}H-NOESY, (**B**) ^{1}H,^{15}N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlation of the 3′-genomic end construct 3_SL3base encompassing nts 29 620–29 671 and 29 840–29 870. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using the genomic numbering, shifted for convenience by 29 000 nts from 5′. Assignments that are marked with an asterisk are tentative and might be subject to change. Imino-proton correlations in (A) between consecutive base pairs are shown in different colors. Included in (C) is the experimentally observed secondary structure of 3_SL3base. Additional closing base pairs are annotated with '±x'. The colors of boxes are according to the correlations in (A). Gray boxes mark base pairs, which were not assigned based on imino-to-imino correlations.

most probably be regarded as unpaired. The third paired region in 3_SL3base shows a network of 10 consecutive base pairs including a pair between U29 845 and C29 666 as well as between U29 846 and U29 665, in the middle of the stem (Figure 16). That the stability is an intrinsic feature of the sequence and not induced by the high stability cUUCGg-loop that caps this helical region, is confirmed by DMS footprinting. Here the entire stem shows a low DMS reactivity validating the introduced mutations (Supplementary Figure S21). Although part of the construct constitutes the primer binding site, the DMS data are in good agreement with the entire NMR secondary structure, except for the loop of SL2. To investigate if magnesium ions present during DMS footprinting are able to induce the predicted base pair formation (Supplementary Figure S21), which could explain the lower reactivity of C29 640, we performed Mg^{2+} titrations with ^{15}N-labeled 3_SL3base. No additional imino proton resonances could be detected at 3 mM Mg^{2+} (Supplementary Figure S22). Thus, the observed differences between NMR and DMS reactivity is not due to altered base pairing patterns in different ionic conditions, but more likely the result of a restricted conformational flexibility within the loop of 3_SL2 (see 'Discussion' section).

3_s2m For 3_s2m, we performed assignment of the imino protons by analyzing cross peaks in a ^{1}H,^{1}H-NOESY, ^{1}H,^{15}N-TROSY, HNN-COSY, ^{1}H,^{15}N-CPMG-NOESY and long range ^{1}H,^{15}N-sfHMQC experiments (Figure 17 and Supplementary Figure S23). The assigned base pair pattern unambiguously reveals a secondary structure that consists of two stem regions separated by an internal asymmetric loop. This secondary structure proposal is in line with the protection pattern we observed in DMS-footprinting (Supplementary Figure S23C). In addition, the imino proton resonances of 3_s2m almost perfectly overlay with those of 3_HVR in ^{1}H,^{15}N-TROSY spectra (Figure 18), with 3_HVR representing the native sequence context of 3_s2m.

Within the SCoV2 genome, a G29,758U mutation is observed for the s2m element, which seems to cause a register-shifted base-pairing in the upper hairpin stem. Key for the assignment of the symmetric motif of G-C base pairs neighboring A29 740-U29 758 and the two loop closing C-G base pairs was the identification of an imino H2 cross peak to A29 756. The loop-closing G-C base pairs C29 743-G29 755 and G29 744-C29 754 show very weak signals in ^{1}H,^{15}N-HNN-COSY experiments and in ^{1}H,^{1}H-NOESY experiments no imino-imino cross peaks are detected for these residues at 298 K, while they could be observed at 283 K (data not shown). We therefore assume only weak or respectively transient base-pairing interactions at room temperature, despite the fact that C29 743 and C29 754 remain unmodified in DMS footprinting. To understand how this

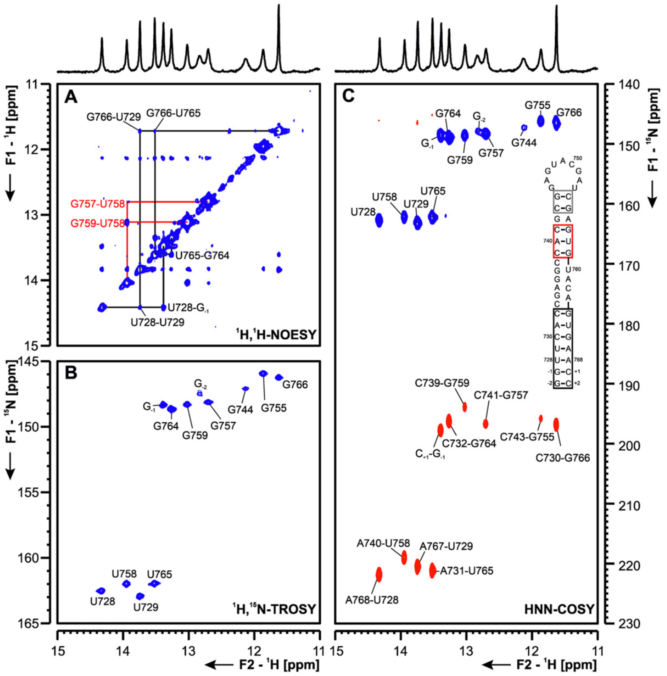


Figure 17. (**A**) ^{1}H,^{1}H-NOESY, (**B**) ^{1}H,^{15}N-TROSY and (**C**) HNN-COSY spectra for imino-proton correlation of the 3′-genomic end construct 3_s2m encompassing nts 29 728–29 768. Positive contours are given in blue, negative contours in red. The imino-proton correlations are annotated using the genomic numbering, shifted for convenience by 29 000 nts from 5′. Imino proton correlations in (A) between consecutive base pairs are shown in different colors. Included in (C) is the experimentally observed secondary structure of 3_s2m. Assignments that are marked with an asterisk are tentative and might be subject to change. Additional closing base pairs are annotated with '±x'. Colors of boxes are according to the correlations in (A).

affects the flexibility of the relatively large loop of 9 nts will require further investigation, in particular as—in contrast to SCoV-2—the SCoV 3_s2m element features a rigid loop-geometry (49). The observed secondary structure for SCoV2, thus strongly deviates from the structure of SCoV.

In order to further substantiate the validity of our approach to investigate the *cis*-acting RNA elements in isolation, we analyzed the SLs SL1 to SL4 from the 5′-UTR and the 3_s2m motif within their native sequence context. We recorded ^{1}H,^{15}N-TROSY spectra of 5_SL1-4 (nts 7–125) and 3_HVR (nts 29 698-29 806) and observed very good agreement with the TROSY spectra of the isolated SLs (Figure 18). Major differences for imino proton resonances are limited to the signals of the additional G/C nucleotides introduced to the isolated RNA elements to enable T7 transcription and stem stabilization. Importantly, these additional nucleotides did not introduce artificial structural changes other than stabilizing the corresponding helical part. The 3_s2m motif forms a stable structural unit within the 3_HVR RNA, demonstrating that its unique fold adopted in SCoV2 is maintained in the longer sequence context. Moreover, it appears to exist as an independent structural unit, since no indications for any additional long-range interactions are observed. This important observation also holds true when comparing the TROSY spectra of SL1-4 with those of the isolated RNAs: no additional resonances appear in the non-canonical spectral regions, which would be indicative for tertiary structure formation or long-range interactions between the individual SLs, exclusively possible within the longer sequence context. Rather, the almost perfect matching of imino proton resonances of 5_SL1, 5_SL2+3 and 5_SL4 onto those of 5_SL1-4, clearly demonstrates that these hairpins form independent structural units within the longer sequence context. Thus, they most likely represent the respective predicted functional units also in the context of the entire genome as well as in the context of the sg mRNA leader sequences (see paragraphs 5_SL1, 5_SL2+3 and 5_SL4).

In summary, by combining reactivity profiles by DMS footprinting and extensive NMR-spectroscopic analysis of a total of 22 RNA constructs representing regulatory relevant regions of the SCoV2 genome, we present conclusive experimental validation of 15 RNA secondary structure models: SLs 1 to 8 of the 5′-genomic end; the attenuator hairpin and the pseudoknot of the frameshift regulating region, and SLs 1 to 3 and s2m of the 3′ genomic end.

DISCUSSION

We analyzed the secondary structures of 15 individual *cis*-acting RNA elements from SCoV2 by NMR spectroscopy, and complemented these data with DMS footprinting analyses of 11 of these RNAs as well as NMR analyses of four 119/472- and 115/337-nt long sequences representing (parts of) the 5′-and 3′-ends of the SCoV2 RNA genome, respectively. All NMR data have been deposited in the BMRB (8). While secondary structure characterization by NMR relies on detecting protons that are stably involved in base-pairing interactions (first of all G and U imino protons), DMS detects highly flexible A and C residues in non-structured regions of the RNA. By combining these two complementary methods, we were able to structurally characterize all 15 RNA elements at single base pair resolution. We find an excellent overall agreement of RNA secondary structures obtained from DMS footprinting within their native, full-length sequence context (top-down approach) with RNA secondary structures determined by NMR spectroscopy for isolated RNA elements (bottom-up approach). Notably, no conflicting secondary structures are suggested by the two methods, except for the low reactivity of some of the loop nucleotides from 3_SL2 observed by DMS probing (see 'Discussion' section below). Recently, preliminary SHAPE and DMS footprinting data have been made available in bioRxiv for the 5′-genomic end, the frameshifting region and even the full-length SCoV2 genome, which overall confirm our secondary structural models (34–35,51–53). Figure 19 shows an overview of the structure of the 15 analyzed *cis*-elements. Regions showing differences in between the available NMR and structural probing data are depicted in gray and discussed below.

5_SL1 NMR analyses indicate open A-U base pairs flanking the upper helix of SL1. Since SL1 is located at the extreme 5′-end of the genome, it is not resolved in most structural studies. In the datasets where SL1 is probed, either the upper or the lower A-U base pair is non-reactive (34,35). This flexibility might affect recognition of SL1 by

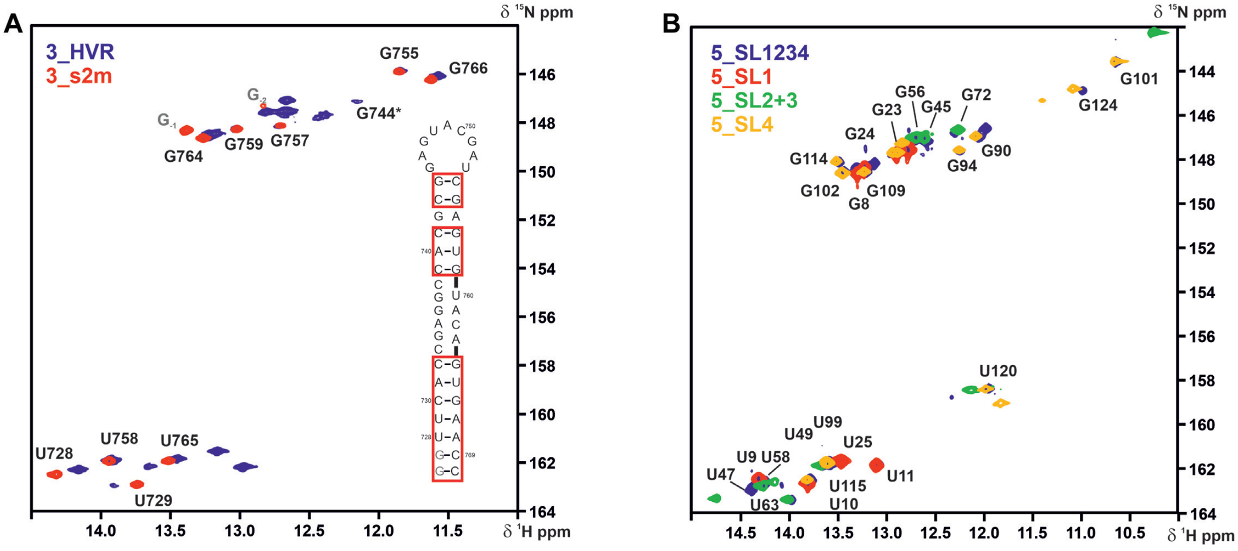


Figure 18. Overlay of ^{1}H,^{15}N-TROSY spectra for the imino-proton correlations of (**A**) 3_HVR (blue) and 3_s2m (red) and (**B**). 5_SL1-4 (blue), 5_SL1 (red), 5_SL2+3 (green) and 5_SL4 (yellow). 5_SL1-4 encompasses nts 7–125 and 3_HVR consists of nts 29 698–29 806. Given assignments were derived from the spectra of the single SLs (see Figures 4, 5 and 6 for (B) and Figure 18 for (A) with genomic numbering shifted for convenience by 29,000 nts from 5′. Spectra in (A) were recorded at 298 K and spectra in (B) were recorded at 283 K. In (A), the isolated 3_s2m is depicted for illustration of assignments. The additional nucleotides G-1 and G-2 are colored in gray. The asterisk denotes a signal which shows up at lower contour levels for 3_s2m. In (B), only resonances which are present both in 5_SL1234 and the respective isolated RNA are annotated. Secondary structures are omitted for clarity.

Nsp1 and should be addressed in future studies of this RNA–protein complex.

5_SL5 Small deviations between DMS- and NMR-based secondary structure models were observed in 5_SL5, where some residues detected to be base-paired in NMR experiments showed high reactivity in DMS footprinting (see Supplementary Figure S10). As DMS experiments were performed at 3 mM Mg^{2+} and higher temperature, we repeated the NMR experiments for 5_SL5stem at 1.5 and 3 mM Mg^{2+} as well as at 298 K. While Mg^{2+} addition mainly induces small chemical shift perturbations (CSPs), the increase in temperature results in dramatic line broadening for imino protons U288, U283, U279 and to a lesser extent U276. These uridine residues are precisely the ones base-pairing with the adenosine residues showing high reactivity in DMS probing, thus demonstrating temperature to be a key effector for stability of this long-range interaction *in-vitro*. Accordingly, other structural probing data likewise showed enhanced reactivity within the stem region of SL5 (35,51,52). A possible reason might be an enhanced conformational heterogeneity within this larger substructure. Further, division of 5_SL5 into three subparts along with terminal stabilization for NMR spectroscopy abolishes the possibility for tertiary contacts between the four helices that may be expected to form within the entire SL5 RNA. More detailed high-resolution analysis of full-length 5_SL5 should be performed in the future, in order to clarify its exact folding details and intrinsic dynamics.

5_SL8 NMR spectra show several non-canonical interactions within the middle part of SL8, which could not be unambiguously assigned so far. A comparison of the available probing data likewise shows differing reactivities across the various datasets, necessitating further detailed high-resolution analyses (34–35,51).

3_SL2 DMS footprinting data showed only intermediate reactivity for residues in the large loop of 3_SL2, while NMR spectroscopy revealed an entirely open structure, without any stable base-pairing interactions. Addition of up to 3 mM Mg^{2+} did not show any influence on the loop stability of SL2 in the 3_SL3base construct. Conversely to DMS probing, SHAPE data suggest that the majority of loop residues within SL2 are flexible (35,51,54), suggesting that the lower reactivity found in DMS footprinting analysis reports on a restricted conformational flexibility of the loop of SL2, rather than base pairing interactions. These variations in 3_SL2 loop conformations are highly interesting, as SL2 is supposed to be part of a molecular switch crucial in regulating genome synthesis (40). The loop of SL2 is either free or forms a pseudoknot interaction with the basal stem of 3_SL1. Strikingly, both the two-hairpin and the pseudoknot structures are essential for viral replication (47). While an open loop structure will facilitate pseudoknot formation, a defined folding of the loop might be required for protein interaction in the two-hairpin conformation. Thus, the loop conformation of 3_SL2 needs to be further investigated.

3_s2m Compared to SCoV, the s2m motif from SCoV2 harbors two mutations (Supplementary Figure S1). One compensatory mutation in the lower helix and one mutation that leads to destabilization of the base pairing of the upper helix. This overall instability of the SCoV2 s2m also detected by NMR, is reflected by high reactivity of this motif in several of the available probing datasets (35,51,54).

In addition to the analysis of individual *cis*-elements, experimental data on the full-length RNA will guide further detailed structural investigations. For example, DMS footprinting suggested an additional helical part for 5_SL4, which is conserved in SCoV. Interestingly, an alternative model for SCoV2 is proposed by computational predic-

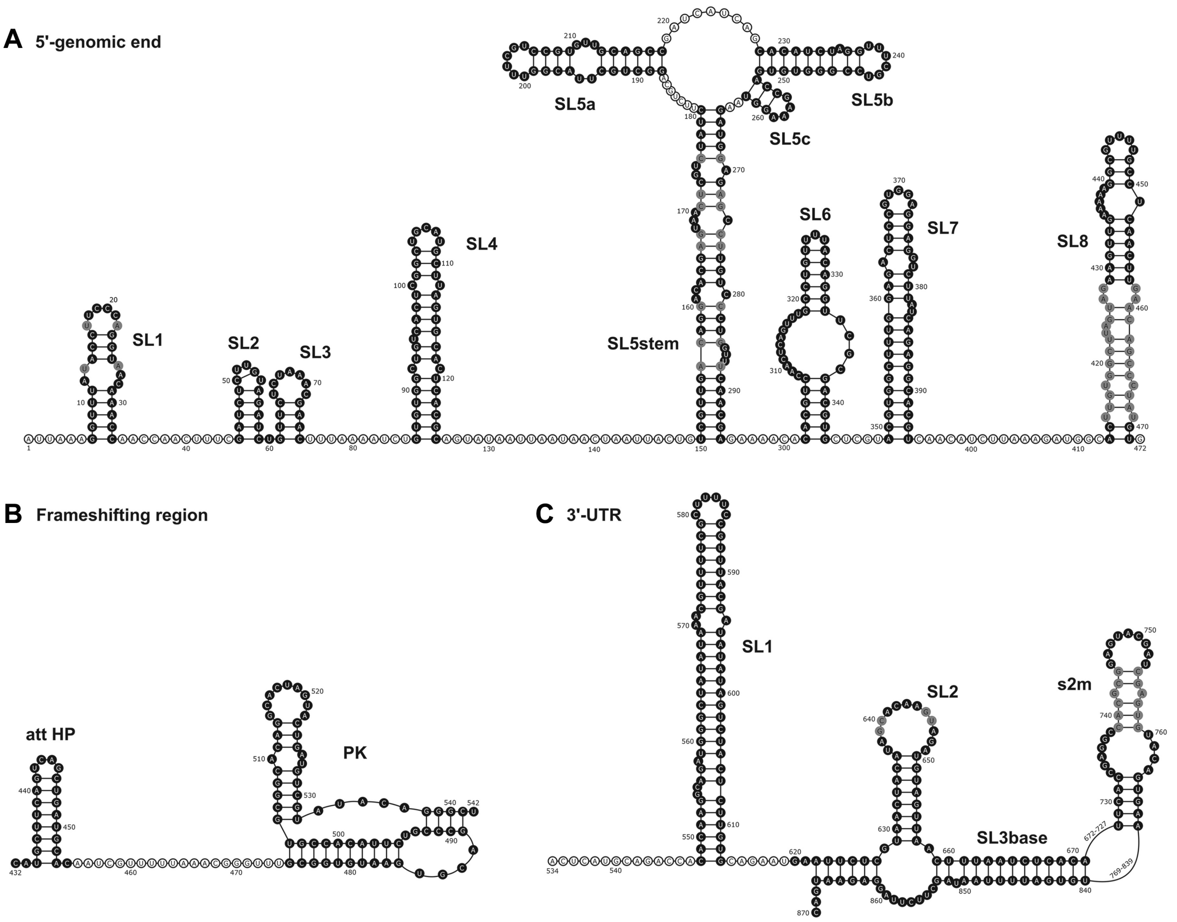


Figure 19. Overview of the experimentally derived structures for the *cis*-elements of the 5′-genomic end (**A**), the frameshifting region (**B**) and the 3′-UTR (**C**). Genomic numbering is shifted for convenience by 13 000 from 5′ for the frameshifting region and 29 000 for the 3′-UTR. *Cis*-elements analyzed by NMR spectroscopy are highlighted in black. Regions with unclear base pairing patterns or high reactivity in structural probing data are shown in gray and discussed in the main text.

tions, where a small SL forms directly downstream of SL4 (45). NMR will be the method of choice to delineate which of the two secondary structures actually form in an RNA construct covering the complete sequence in question. Further, a recent study investigated RNA–RNA long-range interactions within the SCoV2 genome (55). In line with our initial analysis of larger regions of the 5′-genomic end and the 3′-UTR, no long-range interactions between the *cis*-elements within the two regions were identified other than the SL5stem and SL3base. Notably, long-range interactions of the 5′-genomic end and the 3′-UTR with each other as well as coding regions of the genome were identified, which provide interesting new starting points for the investigation of RNA structures important for virus propagation.

In general, we provide a thorough experimental validation of phylogenetic-based *in silico* models for the RNA elements characterized in this study. Differences found between prediction and experimental data show a trend toward fewer stable base pairs in the experimentally determined structures, i.e. A-U base pairs next to bulges or two base pair helices including one G-U predicted to be paired are found to be open in the experimental data (both NMR and DMS). In addition, several non-canonical base pairs could be identified that are not considered in the *in silico* predictions (Supplementary Figure S5). Comparing with the sequence to SCoV, the overall sequence identity of the analyzed RNA elements is 91%, with 5_SL2, 5_SL3 and 5_SL5c showing complete identity, and att HP exhibiting the largest divergence with only 79% identity. In terms of structure, all RNAs are highly conserved although several quite notable exceptions have been detected in this regard. This includes the 5_SL1 element, the s2m element and the att HP. The 24 nt att HP, located immediately upstream of the slippery site and the PRF PK, exhibits five nt variations relative to SCoV that predict it to be significantly less stable (36). In line with this observation, the NMR data indicate that the folding of this motif is substantially different to that proposed for the att HP of SCoV, and involves at least two main conformations (Figure 12). In this last respect, a distinctive advantage of our NMR approach relative to other methodologies is that it allows specific detection of RNA dynamics. Conformational exchange processes have been detected in particular for the att HP and potentially also for other SCoV2 *cis*-acting elements, including 5_SL7 and 5_SL8. These dynamic processes may be relevant for small-molecule or protein recognition.

Overall, however, it is reasonable to expect very similar functional properties for *cis*-acting RNAs in SCoV and SCoV2, which will certainly accelerate progress in elucidating virus biology. Further, 3D modeling efforts such as FARFAR, aiming at providing structures of potential drug targets will benefit from independent, high-resolution experimental validation (6). We provide here an extensive set of chemical shift data covering the entire 5′- and large parts of the 3′-genomic ends as well as two elements from the frameshifting region as a conclusive and reliable basis for further structure-based investigations of the SCoV2 RNA genome. This will greatly facilitate research progress toward defeating the virus and fighting Covid-19.

DATA AVAILABILITY

Chemical shifts are deposited at BMRB. All experimental data are deposited and can be downloaded at http://covid19-nmr.de.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Dr L. Frydman holds the Bertha and Isadore Gudelsky Professorial Chair and heads the Helen and Martin Kimmel Institute for Magnetic Resonance Research, the Clore Institute for High-Field Magnetic Resonance Imaging and Spectroscopy and the Fritz Haber Center for Physical Chemistry—whose collective support is also acknowledged. Dr T. Scherf is the incumbent of the Monroy-Marks Research Fellow Chair. We thank Prof. Eric Westhof for helpful discussions. Special thanks goes to Klara R. Mertinkus and Robbin Schnieders for the figure design.

FUNDING

Goethe University; Deutsche Forschungsgemeinschaft (DFG) [CRC902]; National Institutes of Health [U54AI50470, R01GM126833 B.S.T.]; EU Horizon 2020 [828946]; Weizmann′s Internal Coronavirus Fund; German-Israel Foundation [G-1501-302]; Spanish Ministerio de Economía y Competitividad [RTI2018-093935-B-I00]; la Caixa Banking Foundation; Catholic University of Valencia; Hessisches Ministerium für Wissenschaft und Kunst [BMRZ]; iNEXT-Discovery [871037]. Funding for open access charge: DFG.
Conflict of interest statement. None declared.

This paper is linked to: doi:10.1093/nar/gkaa1053.

REFERENCES

1. Rangan,R., Zheludev,I.N., Hagey,R.J., Pham,E.A., Wayment-Steele,H.K., Glenn,J.S. and Das,R. (2020) RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. *RNA*, **26**, 937–959.
2. Chan,J.F.W., Kok,K.H., Zhu,Z., Chu,H., To,K.K.W., Yuan,S. and Yuen,K.Y. (2020) Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg. Microbes Infect.*, **9**, 221–236.
3. Ishimaru,D., Plant,E.P., Sims,A.C., Yount,B.L., Roth,B.M., Eldho,N. V., Pérez-Alvarado,G.C., Armbruster,D.W., Baric,R.S., Dinman,J.D. *et al.* (2013) RNA dimerization plays a role in ribosomal frameshifting of the SARS coronavirus. *Nucleic Acids Res.*, **41**, 2594–2608.
4. Yang,D. and Leibowitz,J.L. (2015) The structure and functions of coronavirus genomic 3′ and 5′ ends. *Virus Res.*, **206**, 120–133.
5. Madhugiri,R., Fricke,M., Marz,M. and Ziebuhr,J. (2016) Coronavirus cis-acting RNA elements. *Adv. Virus Res.*, **96**, 127–163.
6. Rangan,R., Watkins,A., Kladwang,W. and Das,R. (2020) De novo 3D models of SARS-CoV-2 RNA elements and small-molecule-binding RNAs to guide drug discovery. bioRxiv doi: https://doi.org/10.1101/2020.04.14.041962, 15 April 2020, preprint: not peer reviewed.
7. Andrews,R., Peterson,J., Haniff,H., Chen,J., Williams,C., Grefe,M., Disney,M. and Moss,W. (2020) An in silico map of the SARS-CoV-2 RNA Structurome. bioRxiv doi: https://doi.org/10.1101/2020.04.17.045161, 18 April 2020, preprint: not peer reviewed.
8. Ulrich,E.L., Akutsu,H., Doreleijers,J.F., Harano,Y., Ioannidis,Y.E., Lin,J., Livny,M., Mading,S., Maziuk,D., Miller,Z. *et al.* (2008) BioMagResBank. *Nucleic Acids Res.*, **36**, D402.
9. Tomezsko,P., Swaminathan,H. and Rouskin,S. (2020) Viral RNA structure analysis using DMS-MaPseq. *Methods*, 10.1016/j.ymeth.2020.04.001.
10. Bellaousov,S., Reuter,J.S., Seetin,M.G. and Mathews,D.H. (2013) RNAstructure: web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Res.*, **41**, W471.
11. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
12. Milligan,J.F. and Uhlenbeck,O.C. (1989) Synthesis of small RNAs using T7 RNA polymerase. *Methods Enzymol.*, **180**, 51–62.
13. Ferré-D¢Amaré,A.R. and Doudna,J.A. (1996) Use of cis- and trans-ribozymes to remove 5′ and 3′ heterogeneities from milligrams of in vitro transcribed RNA. *Nucleic Acids Res.*, **24**, 977–978.
14. Guilleres,J., Lopez,P.J., Proux,F., Launay,H. and Dreyfus,M. (2005) A mutation in T7 RNA polymerase that facilitates promoter clearance. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 5958–5963.
15. Chillón,I., Marcia,M., Legiewicz,M., Liu,F., Somarowthu,S. and Pyle,A.M. (2015) Native purification and analysis of long RNAs. In: woodson,S.A. and Allain,F.H.T. (eds). *Methods in Enzymology*. Academic Press Inc., Vol. **558**, pp. 3–37.
16. Lee,W., Tonelli,M. and Markley,J.L. (2015) NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics*, **31**, 1325–1327.
17. Vranken,W.F., Boucher,W., Stevens,T.J., Fogh,R.H., Pajon,A., Llinas,M., Ulrich,E.L., Markley,J.L., Ionides,J. and Laue,E.D. (2005) The CCPN data model for NMR spectroscopy: Development of a software pipeline. *Proteins Struct. Funct. Genet.*, **59**, 687–696.
18. Fürtig,B., Richter,C., Wöhnert,J. and Schwalbe,H.(2003) NMR spectroscopy of RNA. *Chembiochem*, **4**, 936–962.
19. Lopez,J.J. (2020) Forschungsdaten: Alle Zahlen mit wenigen Klicks. *Nachr. Chem.*, **68**, 27–28.
20. Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.
21. Janssen,S. and Giegerich,R. (2015) The RNA shapes studio. *Bioinformatics*, **31**, 423–425.
22. Tanaka,T., Kamitani,W., DeDiego,M.L., Enjuanes,L. and Matsuura,Y. (2012) Severe acute respiratory syndrome coronavirus nsp1 facilitates efficient propagation in cells through a specific translational shutoff of host mRNA. *J. Virol.*, **86**, 11128–11137.
23. Li,L., Kang,H., Liu,P., Makkinje,N., Williamson,S.T., Leibowitz,J.L. and Giedroc,D.P. (2008) Structural lability in SL 1 drives a 5′ UTR-3′ UTR interaction in coronavirus replication. *J. Mol. Biol.*, **377**, 790–803.
24. Chen,S.C. and Olsthoorn,R.C.L. (2010) Group-specific structural features of the 5′-proximal sequences of coronavirus genomic RNAs. *Virology*, **401**, 29–41.
25. Madhugiri,R., Karl,N., Petersen,D., Lamkiewicz,K., Fricke,M., Wend,U., Scheuer,R., Marz,M. and Ziebuhr,J. (2018) Structural and functional conservation of cis-acting RNA elements in coronavirus 5′-terminal genome regions. *Virology*, **517**, 44–55.

26. Lee,C.W., Li,L. and Giedroc,D.P. (2011) The solution structure of coronaviral SL 2 (SL2) reveals a canonical CUYG tetraloop fold. *FEBS Lett.*, **585**, 1049–1053.
27. Grossoehme,N.E., Li,L., Keane,S.C., Liu,P., Dann,C.E., Leibowitz,J.L. and Giedroc,D.P. (2009) Coronavirus N protein N-Terminal Domain (NTD) specifically binds the transcriptional regulatory sequence (TRS) and melts TRS-cTRS RNA duplexes. *J. Mol. Biol.*, **394**, 544–557.
28. Dufour,D., Mateos-Gomez,P.A., Enjuanes,L., Gallego,J. and Sola,I. (2011) Structure and functional relevance of a transcription-regulating sequence involved in coronavirus discontinuous RNA synthesis. *J. Virol.*, **85**, 4963–4973.
29. Yang,D., Liu,P., Giedroc,D.P. and Leibowitz,J. (2011) Mouse hepatitis virus SL 4 functions as a spacer element required to drive subgenomic RNA synthesis. *J. Virol.*, **85**, 9199–9209.
30. Wu,H.-Y., Guan,B.-J., Su,Y.-P., Fan,Y.-H. and Brian,D.A. (2014) Reselection of a genomic upstream open reading frame in mouse hepatitis coronavirus 5′-Untranslated-Region mutants. *J. Virol.*, **88**, 846–858.
31. Guan,B.-J., Su,Y.-P., Wu,H.-Y. and Brian,D.A. (2012) Genetic evidence of a long-range RNA-RNA interaction between the genomic 5 = untranslated region and the nonstructural protein 1 coding region in murine and bovine coronaviruses. *J. Virol.*, **86**, 4631–4643.
32. Guan,B.-J., Wu,H.-Y. and Brian,D.A. (2011) An optimal cis-Replication SL IV in the 5′ untranslated region of the mouse coronavirus genome extends 16 nucleotides into open reading frame 1. *J. Virol.*, **85**, 5593–5605.
33. Nozinovic,S., Fürtig,B., Jonker,H.R.A., Richter,C. and Schwalbe,H. (2010) High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. *Nucleic Acids Res.*, **38**, 683–694.
34. Iserman,C., Roden,C., Boerneke,M., Sealfon,R., McLaughlin,G., Jungreis,I., Park,C., Boppana,A., Fritch,E., Hou,Y.J. *et al.* (2020) Specific viral RNA drives the SARS CoV-2 nucleocapsid to phase separate. bioRxiv doi: https://doi.org/10.1101/2020.06.11.147199, 12 June 2020, preprint: not peer reviewed.
35. Lan,T.C.T.T., Allan,M.F., Malsick,L.E., Khandwala,S., Nyeo,S.S.Y.Y., Bathe,M., Griffiths,A. and Rouskin,S. (2020) Structure of the full SARS-CoV-2 RNA genome in infected cells. bioRxiv doi: https://doi.org/10.1101/2020.06.29.178343, 26 October 2020, preprint: not peer reviewed.
36. Jucker,F.M., Heus,H.A., Yip,P.F., Moors,E.H.M.M. and Pardi,A. (1996) A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.*, **264**, 968–980.
37. Yang,D., Liu,P., Wudeck,E. V., Giedroc,D.P. and Leibowitz,J.L. (2015) Shape analysis of the rna secondary structure of the mouse hepatitis virus 5′ untranslated region and n-terminal nsp1 coding sequences. *Virology*, **475**, 15–27.
38. Brockway,S.M. and Denison,M.R. (2005) Mutagenesis of the murine hepatitis virus nsp1-coding region identifies residues important for protein processing, viral RNA synthesis, and viral replication. *Virology*, **340**, 209–223.
39. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
40. Cho,C.P., Lin,S.C., Chou,M.Y., Hsu,H.T. and Chang,K.Y. (2013) Regulation of programmed ribosomal frameshifting by Co-Translational refolding RNA hairpins. *PLoS One*, **8**, e62283.
41. Kelly,J.A., Olson,A.N., Neupane,K., Munshi,S., San Emeterio,J., Pollack,L., Woodside,M.T. and Dinman,J.D. (2020) Structural and functional conservation of the programmed -1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). *J. Biol. Chem.*, **295**, 10741–10748.
42. Baranov,P. V., Henderson,C.M., Anderson,C.B., Gesteland,R.F., Atkins,J.F. and Howard,M.T. (2005) Programmed ribosomal frameshifting in decoding the SARS-CoV genome. *Virology*, **332**, 498–510.
43. Su,M.C., Chang,C. Te, Chu,C.H., Tsai,C.H. and Chang,K.Y. (2005) An atypical RNA pseudoknot stimulator and an upstream attenuation signal for -1 ribosomal frameshifting of SARS coronavirus. *Nucleic Acids Res.*, **33**, 4265–4275.
44. Plant,E.P., Pérez-Alvarado,G.C., Jacobs,J.L., Mukhopadhyay,B., Hennig,M. and Dinman,J.D. (2005) A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol.*, **3**, 1012–1023.
45. Plant,E.P., Perez-Alvarado,G.C., Jacobs,J.L., Mukhopadhyay,B., Hennig,M., Dinman,J.D., Pérez-Alvarado,G.C., Jacobs,J.L., Mukhopadhyay,B., Hennig,M. *et al.* (2005) A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol.*, **3**, e172.
46. Zust,R., Miller,T.B., Goebel,S.J., Thiel,V. and Masters,P.S. (2008) Genetic interactions between an essential 3′ cis-acting RNA pseudoknot, replicase gene products, and the extreme 3′ end of the mouse coronavirus genome. *J. Virol.*, **82**, 1214–1228.
47. Goebel,S.J., Hsue,B., Dombrowski,T.F. and Masters,P.S. (2004) Characterization of the RNA components of a putative molecular switch in the 3′ untranslated region of the murine coronavirus genome. *J. Virol.*, **78**, 669–682.
48. Goebel,S.J., Miller,T.B., Bennett,C.J., Bernard,K.A. and Masters,P.S. (2007) A hypervariable region within the 3′ cis-acting element of the murine coronavirus genome is nonessential for RNA synthesis but affects pathogenesis. *J. Virol.*, **81**, 1274–1287.
49. Robertson,M.P., Igel,H., Baertsch,R., Haussler,D., Ares,M. and Scott,W.G. (2005) The structure of a rigorously conserved RNA element within the SARS virus genome. *PLoS Biol.*, **3**, e5.
50. Ohlenschläger,O., Wöhnert,J., Bucci,E., Seitz,S., Häfner,S., Ramachandran,R., Zell,R. and Görlach,M. (2004) The structure of the stemloop D subdomain of coxsackievirus B3 cloverleaf RNA and its interaction with the proteinase 3C. *Structure*, **12**, 237–248.
51. Manfredonia,I., Nithin,C., Ponce-Salvatierra,A., Ghosh,P., Wirecki,T.K., Marinus,T., Ogando,N.S., Snijder,E.J., Hemert,M.J. van, Bujnicki,J.M. *et al.* (2020) Genome-wide mapping of therapeutically-relevant SARS-CoV-2 RNA structures. bioRxiv doi: https://doi.org/10.1101/2020.06.15.151647, 15 June 2020, preprint: not peer reviewed.
52. Sanders,W., Fritch,E.J., Madden,E.A., Graham,R.L., Vincent,H.A., Heise,M.T., Baric,R.S. and Moorman,N.J. (2020) Comparative analysis of coronavirus genomic RNA structure reveals conservation in SARS-like coronaviruses. bioRxiv doi: https://doi.org/10.1101/2020.06.15.153197, 16 June 2020, preprint: not peer reviewed.
53. Tavares,R.C.A., Mahadeshwar,G. and Pyle,A.M. (2020) The global and local distribution of RNA structure throughout the SARS-CoV-2 genome. bioRxiv doi: https://doi.org/10.1101/2020.07.06.190660, 07 July 2020, preprint:not peer reviewed.
54. Huston,N., Wan,H., Araujo Tavares,R., de,C., Wilen,C. and Pyle,A.M. (2020) Comprehensive in-vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. bioRxiv doi: https://doi.org/10.1101/2020.07.10.197079, 10 July 2020, preprint: not peer reviewed.
55. Ziv,O., Price,J., Shalamova,L., Kamenova,T., Goodfellow,I., Weber,F. and Miska,E.A. (2020) The short-and long-range RNA-RNA Interactome of SARS-CoV-2. bioRxiv doi: https://doi.org/10.1101/2020.07.19.211110, 20 July 2020, preprint: not peer reviewed.
56. Sklenar,V. (1995) Suppression of radiation damping in multidimensional NMR experiments using magnetic field gradients. *J. Magn. Reson. Ser. A*, **114**, 132–135.
57. Solyom,Z., Schwarten,M., Geist,L., Konrat,R., Willbold,D. and Brutscher,B. (2013) BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. *J. Biomol. NMR*, **55**, 311–321.
58. Favier,A. and Brutscher,B. (2011) Recovering lost magnetization: polarization enhancement in biomolecular NMR. *J. Biomol. NMR*, **49**, 9–15.
59. Mori,S., Abeygunawardana,C., Johnson,M.O. and Van Zljl,P.C.M. (1995) Improved sensitivity of HSQC spectra of exchanging protons at short interscan delays using a new fast HSQC (FHSQC) detection scheme that avoids water saturation. *J. Magn. Reson. Ser. B*, **108**, 94–98.
60. Mueller,L., Legault,P. and Pardi,A. (1995) Improved RNA structure determination by detection of NOE contacts to Exchange-Broadened amino protons. *J. Am. Chem. Soc.*, **117**, 11043–11048.
61. Dingley,A.J. and Grzesiek,S. (1998) Direct observation of hydrogen bonds in nucleic acid base pairs by internucleotide 2J(NN) couplings. *J. Am. Chem. Soc.*, **120**, 8293–8297.
62. Lescop,E., Kern,T. and Brutscher,B. (2010) Guidelines for the use of band-selective radiofrequency pulses in hetero-nuclear NMR:

example of longitudinal-relaxation-enhanced BEST-type 1H-15N correlation experiments. *J. Magn. Reson.*, **203**, 190–198.

63. Schulte-Herbrüggen,T. and Sørensen,O.W. (2000) Clean TROSY: compensation for relaxation-induced artifacts. *J. Magn. Reson.*, **144**, 123–128.
64. Schanda,P. and Brutscher,B. (2005) Very fast two-dimensional NMR spectroscopy for real-time investigation of dynamic events in proteins on the time scale of seconds. *J. Am. Chem. Soc.*, **127**, 8014–8015.
65. Novakovic,M., Olsen,G.L., Pintér,G., Hymon,D., Fürtig,B., Schwalbe,H. and Frydman,L. (2020) A 300-fold enhancement of imino nucleic acid resonances by hyperpolarized water provides a new window for probing RNA refolding by 1D and 2D NMR. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 2449–2455.
66. Ikura,M. and Bax,A. (1992) Isotope-filtered 2D NMR of a protein-peptide complex: study of a skeletal muscle myosin light chain kinase fragment bound to calmodulin. *J. Am. Chem. Soc.*, **114**, 2433–2440.
67. Piotto,M., Saudek,V. and Sklenár,V. (1992) Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR*, **2**, 661–665.
68. Sklenář,V., Piotto,M., Leppik,R. and Saudek,V. (1993) Gradient-tailored water suppression for 1H-15N HSQC experiments optimized to retain full sensitivity. *Journal of Magnetic Resonance - Series A*, **102**, 241–245.
69. Shaka,A., Lee,C. and Pines,A. (1988) Iterative schemes for bilinear operators; application to spin decoupling. *J. Magn. Reson.*, **77**, 274–293.
70. Hwang,T.L. and Shaka,A.J. (1995) Water suppression that works. Excitation sculpting using arbitrary wave-forms and pulsed-field gradients. *J. Magn. Reson. Ser. A*, **112**, 275–279.

5.14 **Research article:** Exploring the Druggability of Conserved RNA Regulatory Elements in the SARS-CoV-2 Genome

Sridhar Sreeramulu*, Christian Richter*, Hannes Berg, Maria A. Wirtz Martin, Betül Ceylan, Tobias Matzel, Jennifer Adam, Nadide Altincekic, Kamal Azzaoui, Jasleen Kaur Bains, Marcel J. J. Blommers, Jan Ferner, Boris Fürtig, Michael Göbel, J. Tassilo Grün, Martin Hengesbach, Katharina F. Hohmann, Daniel Hymon, Bozana Knezic, Jason N. Martins, Klara R. Mertinkus, Anna Niesteruk, Stephen A. Peter, Dennis J. Pyper, Nusrat S. Qureshi, Ute Scheffer, Andreas Schlundt, Robbin Schnieders, Elke Stirnal, Alexey Sudakov, Alix Tröster, Jennifer Vögele, Anna Wacker, Julia E. Weigand, Julia Wirmer-Bartoschek, Jens Wöhnert and Harald Schwalbe; *Angewandte Chemie International Edition* **60**, 19191-19200 (2021)

* These authors contributed equally to this work.

In this work, the druggability of the conserved RNA elements from the SARS-CoV-2 genome was experimentally validated. An NMR-based fragment screening against 15 RNA elements was performed with 768 compounds using ^{1}H-1D NMR binding assays. The pseudoknot and stem-loop 3_SL3base RNA elements were selected as the most promising drug targets with binding affinities in the µM range for further ligand development.

S. Sreeramulu and C. Richter performed the NMR-based fragment screening experiments, analyzed the data, and wrote the manuscript. The author of this thesis was part of the Covid-19 NMR consortium and the RNA production team that prepared and purified various isotope-labeled RNA samples for NMR spectroscopy.

https://doi.org/10.1002/anie.202103693

Covid Virus

How to cite: Angew. Chem. Int. Ed. 2021, 60, 19191–19200
International Edition: doi.org/10.1002/anie.202103693
German Edition: doi.org/10.1002/ange.202103693

Exploring the Druggability of Conserved RNA Regulatory Elements in the SARS-CoV-2 Genome

*Sridhar Sreeramulu+, Christian Richter+, Hannes Berg, Maria A. Wirtz Martin, Betül Ceylan, Tobias Matzel, Jennifer Adam, Nadide Altincekic, Kamal Azzaoui, Jasleen Kaur Bains, Marcel J. J. Blommers, Jan Ferner, Boris Fürtig, Michael Göbel, J. Tassilo Grün, Martin Hengesbach, Katharina F. Hohmann, Daniel Hymon, Bozana Knezic, Jason N. Martins, Klara R. Mertinkus, Anna Niesteruk, Stephen A. Peter, Dennis J. Pyper, Nusrat S. Qureshi, Ute Scheffer, Andreas Schlundt, Robbin Schnieders, Elke Stirnal, Alexey Sudakov, Alix Tröster, Jennifer Vögele, Anna Wacker, Julia E. Weigand, Julia Wirmer-Bartoschek, Jens Wöhnert, and Harald Schwalbe**

Abstract: *SARS-CoV-2 contains a positive single-stranded RNA genome of approximately 30 000 nucleotides. Within this genome, 15 RNA elements were identified as conserved between SARS-CoV and SARS-CoV-2. By nuclear magnetic resonance (NMR) spectroscopy, we previously determined that these elements fold independently, in line with data from in vivo and ex-vivo structural probing experiments. These elements contain non-base-paired regions that potentially harbor ligand-binding pockets. Here, we performed an NMR-based screening of a poised fragment library of 768 compounds for binding to these RNAs, employing three different ^{1}H-based 1D NMR binding assays. The screening identified common as well as RNA-element specific hits. The results allow selection of the most promising of the 15 RNA elements as putative drug targets. Based on the identified hits, we derive key functional units and groups in ligands for effective targeting of the RNA of SARS-CoV-2.*

Introduction

Since early 2020, enormous scientific efforts are geared towards antiviral treatment for SARS-CoV-2 (SCoV-2). Impressive advancements are reported in vaccine development. Furthermore, experimental approaches are being actively pursued in drug repurposing and in design and synthesis of new drugs. Such studies are supported by virtual screening and molecular docking campaigns.[1–7] Compound libraries with more than 10^6 molecules have been screened virtually. So far, most efforts focus on targeting proteins to inhibit viral propagation, while few attempts have been reported for direct targeting of the large viral RNA genome.[8] This focus on proteins as drug targets comes despite the fact that structured RNA elements have been increasingly recognized as drug targets in recent years,[9–16] including SARS-CoV (SCoV). In SCoV, for example, the pseudoknot (PK) element involved in ribosomal frameshifting was identified as a potent drug target.[17] The abundance of viral RNA in infected host cells opens a further opportunity for pharmaceutical intervention: for coronaviruses in general and for SCoV-2 infected cells in particular, viral RNA accounts for approximately two thirds of the total RNA.[18] In fact, the transcriptome of SCoV-2 is discussed to contain numerous potential druggable sites.[19,20] Currently, however, only few experimental screenings for ligands directly targeting the

[*] Dr. S. Sreeramulu,[+] Dr. C. Richter,[+] H. Berg, M. A. Wirtz Martin, B. Ceylan, T. Matzel, J. Adam, N. Altincekic, J. K. Bains, Dr. J. Ferner, Dr. B. Fürtig, Prof. Dr. M. Göbel, J. T. Grün, Dr. M. Hengesbach, K. F. Hohmann, D. Hymon, B. Knezic, J. N. Martins, K. R. Mertinkus, Dr. A. Niesteruk, D. J. Pyper, Dr. N. S. Qureshi, Dr. U. Scheffer, Dr. R. Schnieders, E. Stirnal, A. Sudakov, A. Tröster, Dr. A. Wacker, Dr. J. Wirmer-Bartoschek, Prof. Dr. H. Schwalbe
Institute for Organic Chemistry and Chemical Biology, Center for Biomolecular Magnetic Resonance (BMRZ), Johann Wolfgang Goethe-University
Max-von-Laue-Str. 7+9, 60438 Frankfurt/M. (Germany)
E-mail: schwalbe@nmr.uni-frankfurt.de

Dr. A. Schlundt, J. Vögele, Prof. Dr. J. Wöhnert
Institute for Molecular Biosciences, Center for Biomolecular Magnetic Resonance (BMRZ), Johann Wolfgang Goethe-University
Max-von-Laue-Str. 7+9, 60438 Frankfurt/M. (Germany)

S. A. Peter, Dr. J. E. Weigand
Department of Biology, Technical University of Darmstadt
Schnittspahnstr. 10, 64287 Darmstadt (Germany)

Dr. K. Azzaoui, Dr. M. J. J. Blommers
Saverna Therapeutics
Pumpmattenweg 3, 4105 Biel-Benken (Switzerland)

Dr. N. S. Qureshi
Present address: EMBL Heidelberg
Meyerhofstr. 1, 69117 Heidelberg (Germany)

[+] These authors contributed equally to this work.

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under: https://doi.org/10.1002/anie.202103693.

structured RNA elements in SCoV-2 have been reported.[8,21,22]

Here, we report a holistic NMR-based screening campaign investigating previously identified structured RNA elements.[23] We determine the druggability of these regulatory RNA elements by NMR-based screening,[24] using a previously implemented workflow.[25] We use a well-characterized,[26] highly diverse fragment library poised for follow-up chemistry. The screen of 768 fragments conducted here yielded a high hit rate, despite the stringent hit definition from three independent NMR experiments, demonstrating the general druggability of the viral RNA targets.

Previously, we identified secondary structures in these RNA targets (Supporting Information, Figure S1) by NMR spectroscopy.[23] Our selection of 15 SCoV-2 RNA targets from the genome was based on functional and structural conservation among Betacoronaviruses.

In the context of screening the viral RNA targets, we briefly describe the functional roles of the 15 RNA elements. In the 5′-genomic end, stem loop 1 (5_SL1) is associated with viral mRNA escape from Nsp1-mediated translational shutdown.[27] 5_SL2 + 3 harbors a stringently conserved stem-loop (Supporting Information, Figure S1)[28] and the transcription regulatory sequence regulating the production of subgenomic RNAs that is embedded within a second, more labile stem-loop. It further comprises a nucleotide stretch involved in long-range interactions.[29] The stem-loop 5_SL4 is indispensable for replication and contains a short upstream open reading frame (uORF).[30] We dissected the large SL5 element into the three sub-elements 5_SL5a, 5_SL5b + c and 5_SLstem, the latter of which contains the ORF1 start codon. The functions of SL6, 7 and 8 are less well defined, but their high degree of structure conservation between SCoV and SCoV-2 suggests regulatory roles as well.

The frameshifting region is a well-established target for binding of ligands of low molecular weight.[17,31,32] We thus screened the attenuator hairpin (att HP) and PK. Within the 3′-UTR, a putative structural switch is supposed to play a role in the initiation events of (−)-strand synthesis.[33] The individual elements forming this elaborate RNA architecture are 3_SL1, 3_SL2 and 3_SL3base. Furthermore, the hypervariable region of the 3′-UTR harbours the 3_s2m motif, which is recurring among a variety of RNA viruses, and seems to be involved in coronavirus pathogenicity.[34] For each of these 15 RNA elements chosen for fragment-based screening, their existence and structural integrity within the SCoV-2 genome was demonstrated by us[23] and others.[35–37] Correct folding prior to screening experiments was ensured as detailed in the Material and Methods section.

Results and Discussion

We analyzed the non-canonical RNA elements (Supporting Information, Table S1), revealing four different classes of non-canonical segments:

(i) Capping loops (blue and grey in Supporting Information, Figure S1, Table S1): In general, loops are ca. 6 nts long and uridine enriched, while adenosine depleted, with respect to the genomic nucleobase composition (Supporting Information, Table S3). We here distinguish the capping loops by their different degrees of structure, with the loops of 5_SL2 and 5_SL5c adopting typical tetraloop conformations.[28] A hexaloop 5′-UUUCGU-3′ found in 5_SL5a also caps 5_SL5b. It is interesting to note that one of the few viral mutations described since early 2020 includes a C to U mutation at position 4 in almost all SCoV-2 strains.[38] The loops of the PK, 3_s2m and especially 3_SL2 are exceptionally large (9, 9 and 13 nts, respectively). The 3_SL2 loop has been proposed to engage in a 3′-pseudoknot structure.[33,39]

ii) Bulges/internal loops (green in Supporting Information, Figure S1, Table S1). Bulges and internal loops dictate the relative orientations of helical segments. They are key elements in RNA tertiary architecture, and are often interaction sites for proteins, RNAs, metal ions or small molecules. Frequently, the helical segments of the SCoV-2 RNA elements are interrupted either by a single unpaired nucleotide or by stretches of unpaired residues. We classify single and tandem mismatched residues separately (under iii). With this classification, the remaining internal loops are asymmetric. The average size of internal loops is ca. 7 nts, and only 5_SL6, 3_SL1 and 3_s2m contain larger internal loops. In contrast to the capping loops, the internal loops and bulges are uridine depleted and adenosine enriched with respect to the genomic base composition (Supporting Information, Table S3). Most bulges consist of only one nucleotide, but up to four bulged-out nucleotides are found.

(iii) Non-canonical base-pairs/mismatches (orange in Supporting Information, Figure S1). Next to the well-known non-canonical G-U/U-G base-pair, which disrupts A-helical conformation only marginally,[40] the RNA elements show a variety of mismatched residues. Interestingly, tandem pyrimidine-pyrimidine (Y-Y) base pairs are identified in 3_SL1 and 3_SL3base by their base-pair connectivities, whereas isolated Y-Y mismatches in 5_SL4, 5_SL5a, 5_SL5b and 5_SL8 do not form base pairs with detectable H-bonds.

(iv) Three-helix junctions (pink in Supporting Information, Figure S1). Single-stranded regions connecting three helices are found in the PK and 3_SL3base and are structurally flexible. Helix junctions are generally considered high-potential binding pockets for small molecules, reminiscent to bacterial riboswitches.[41]

Supporting Information, Table S1 gives an overview of loop and bulge sequences for the 15 RNAs. Column 3 shows the frequency of a sequence in the SCoV-2 genome[42] compared to its abundance in predicted structured regions therein.[20] Column 4 correlates each sequence to the number of actual loops or bulges predicted (or, if deviating, experimentally determined[23]) for this sequence. To estimate the relevance of a given loop or bulge sequence, its conservation (resistance to mutation) is stated in column 5.[43] Supporting Information, Table S1 also provides information on unique loop or bulge motifs within the structured RNA genome as defined by Rangan et al.[20]

The DSI-poised library (DSI-PL, Supporting Information, Table S2. DSI-PL-768_Ligands) is the second generation of the initially developed poised library,[44] composed of 768 highly diverse and poised fragments specifically designed to facilitate easy downstream synthesis.

This library has already been successfully used to screen the main protease nsp5 from SCoV-2.[45] Ligand-observed NMR-based screening is a versatile method routinely adapted in the discovery of fragments binding to a target. Screening individual fragments for individual targets is, however, time-consuming from a library with hundreds of fragments. Instead, evaluation of binding in mixtures ranging from 5 to 20 compounds is highly advantageous. One of the prerequisites in designing an ideal mixture is minimal signal overlap between the ^{1}H-NMR signals of the individual compounds so that spectral changes in the presence of the RNA target can be associated to a single compound. Previously, we performed molecular clustering analysis of the DSI-PL fragment library. We found a total of ca. 400 distinguished chemical clusters, from which ca. 200 clusters contained singletons, suggesting a high chemical diversity of the library.[26] We hypothesized that the chemical diversity of ligands might also result in significantly diverse ^{1}H-NMR spectra of the fragments in each sub-library mix. In total, we generated 64 mixtures with 12 fragments (Supporting Information, Figure S2C) each from the 768 compounds of the DSI-PL. Indeed, the ^{1}H-NMR spectra of each of the randomly chosen 12 fragments is significantly different from one another (Supporting Information, Figure S2A and B; black spectra), so that the signals of the individual compounds could be traced back in the ^{1}H-NMR spectrum (Supporting Information, Figure S2A and B; blue spectra).

We applied ligand-observed ^{1}H-NMR experiments against the 15 RNAs ranging in size from 29–90 nts depicted in Supporting Information, Figure S1 and five additional larger RNAs with sizes between 118–472 nts that comprise several RNA elements (Supporting Information, Table S1 and Files "1–21_Hits" detailing all experimental hits). In screening experiments, changes in the ^{1}H signals of the ligand in the presence and absence of the RNA served as readout for binding. After optimization (Supporting Information, Figure S3), we recorded ^{1}H 1D spectra at $T = 293$ K to detect chemical shift perturbations (CSPs) or line broadening (LB) (Figure 1 A), waterLOGSY[46,47] and T_2-relaxation[48] experiments (Supporting Information, Table S4). Previously, these experiments have successfully been utilized to identify small molecule binders for RNA.[49,50] NMR signals of each of the 64 fragment mixtures were monitored at a ratio of each ligand to one RNA target of 20:1. 190 µl of a 10 µM RNA in screening buffer (25 mM KPi, 50 mM KCl, pH 6.2) was manually pipetted into 3 mm NMR tubes. 10 µl of the fragment mixture was added using a pipetting robot to a final concentration of 200 µM for each fragment. In waterLOGSY, the signal from a bound fragment is positive while the signal of a free ligand is negative. T_2-relaxation-based experiments take advantage of the fact that large biomolecular targets including all RNAs studied here relax faster than small fragments (Figure 1 A).

For identifying binders within the mixtures, we first compared spectra from all three of the above experiments

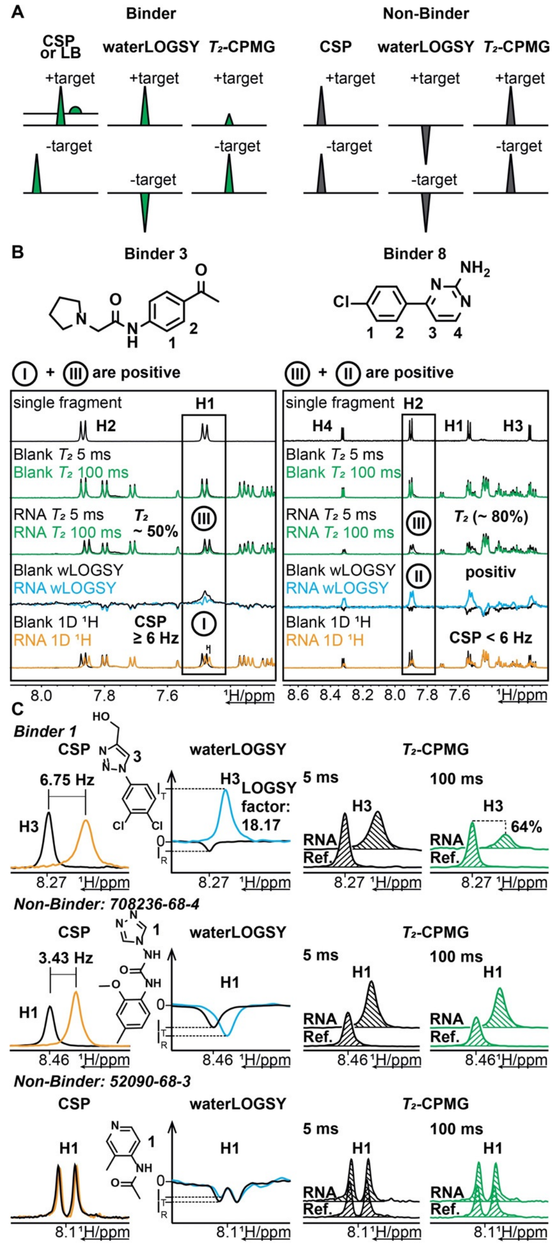


Figure 1. NMR-based fragment screening and hit identification: A, Schematic representation of the experiments and criteria used to identify binders/ non-binders. B, Chemical structures of Binder 3 and Binder 8 along with an overlay of their corresponding NMR spectra [1D ^{1}H (i), waterLOGSY (ii) and T_2-CPMG (iii); (5 ms and 100 ms)] of the mixture in absence (blank) and presence of RNA. The top 1D ^{1}H-spectrum corresponds to the single fragment used for chemical shift deconvolution in the mixture. The inset highlights the signal corresponding to the hit used for analysis. Visual inspection of the overlay indicates that Binder 3 displays CSP (≥6 Hz) and T_2 reduction of ca. 50%. Binder 8 shows no CSP, but a clear sign change in waterLOGSY and T_2-reduction of ca. 80%. C, 1D ^{1}H-NMR spectral regions of fragments with strong (top row), weak (middle row) or no binding (lower row) to 3_SL2. Binder 1 shows a CSP of 6.75 Hz, a positive waterLOGSY signal and a T_2-reduction of 64% suggesting that it binds to 3_SL2 (top row). 708236–68-4 shows a minor CSP of 3.43 Hz, but neither a positive waterLOGSY signal nor a T_2-reduction, indicating either very weak or no binding to 3_SL2 (middle row). 52090–68-3 does not bind to 3_SL2 (lower row of the mixture (Supporting Information, Figure S2, blue spectrum).

and analysed differences by visual inspection (Figure 1B). As stringent criteria, CSPs ≥ 6 Hz or severe line broadening, sign change in the waterLOGSY or ≥ 20% decrease of signal intensity in the T_2-relaxation experiment was used to identify binders. Ligands were assigned as a hit if two of the three criteria were satisfied. For example, binder **3** was defined as a hit by a CSP ≥ 6 Hz, reduced signal intensity of T_2 (ca. 50%), while the waterLOGSY effect cannot unambiguously identified (Figure 1B, left). Similarly, binder **8** qualifies as a hit, showing changes in waterLOGSY and T_2, but not CSP (Figure 1B, right). In the second stage of analysis, for the hits thus identified, we quantified CSP, LOGSY effect[51,52] and T_2-reduction (Q^{bind}) in intensity.[25]

Previously, several reports attempted to delineate common scaffolds binding to RNA.[9,53,54] Our results here show that 40 different fragments are found to bind to the 15 RNA elements and an additional 29 fragments bind to the five multi-element RNAs. In total, we observed 108 binding events. For 5_SL2 + 3 and 5_SL5a, no fragment binding was detected. For three RNAs, we identified hits that only bind to this specific RNA element (5_SL5b + c, 5_SL8, 3_SL1). Furthermore, we identified 48 fragments that recognize more than one target RNA. The range of binding promiscuity ranges from binding of two RNA targets (ligands 35–48, Figure 2) up to 18 out of the 20 screened RNA targets (ligand 1, Figure 2).

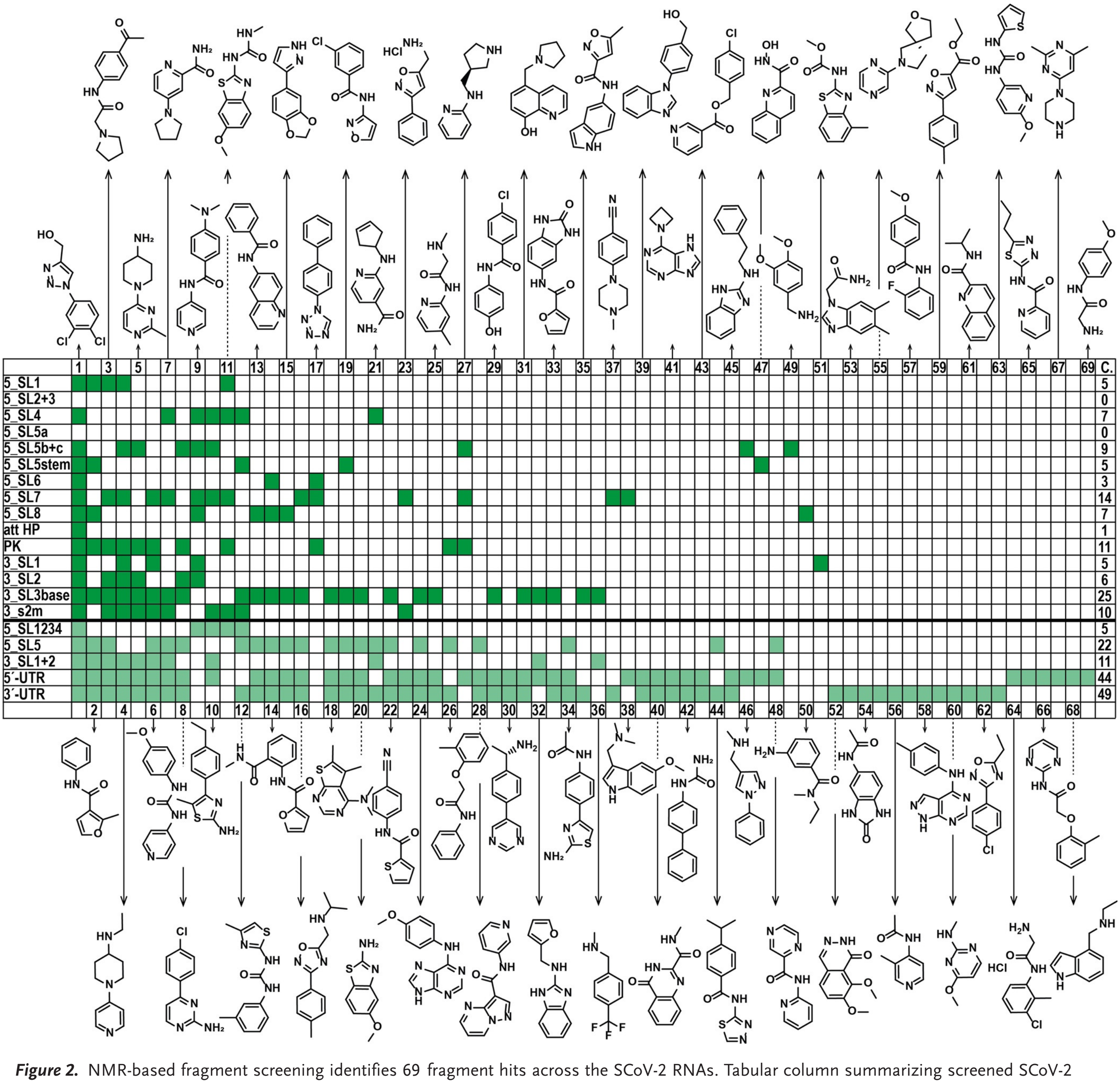


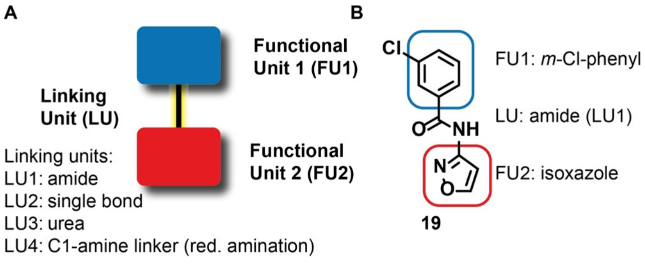
Figure 2. NMR-based fragment screening identifies 69 fragment hits across the SCoV-2 RNAs. Tabular column summarizing screened SCoV-2 RNA elements and their corresponding hits. Dark green indicates hits for individual RNA elements; light green indicates hits for larger RNAs. The last column (C.) lists the number of binders to the investigated RNA target. Chemical structures of respective fragment hits are shown.

We observed no clear correlation between the number of non-helical residues (those forming loops, bulges and mismatches) in an RNA and the number of hits, although the non-targeted RNAs 5_SL2+3 and 5_SL5a belong to the smallest RNAs investigated. Strikingly, 5_SL7 appears to harbor little non-helical structure, but shows 14 hits. By contrast, 3_SL1 contains roughly three times as much non-helical space and has only four hits. However, the RNA elements with the most complex structures, 3_SL3base and PK, are also among those with the highest hit rates (Supporting Information, Table S5).

Inspection of all hits reveals the most significant molecular descriptors distinguishing between binders and non-binders are the number of aromatic rings and the sp^3-character of the compounds. In general, binders have one to three aromatic rings and less sp^3-carbon atoms than non-binders (see Supporting Information, Table S6).

We noticed that 30 out of 40 fragments have a modular architecture: They consist of two functional units, which are mainly non-, mono- or di-substituted rings of the size 5 and 6 (Scheme 1) and connected by four different types of linker units. The mostly aromatic or heteroaromatic core motifs are often modified with a small set of substituents, which are key elements for generating affinity towards the RNA target. The remaining 10 fragments contain a single functional unit (class 1), often substituted with a functional group. Analysis of the additional 29 hits for which binding is detected only to the larger RNA elements are given in the Supporting Information, Figure S5.

Within the functional units, six functional groups are frequently found (Schemes 1 and 2). In the majority of hits, the two functional units are connected either by an amide bond or by a single bond. We further find urea and carbamates as linkers. They contain hydrogen bond donors and acceptors. Additionally, functional groups are often linked by a C1-linker resulting from reductive amination. Furthermore, similar simultaneous donor and acceptor properties are found in the six-membered (pyridine-acetamide, pyrimidine-2-amine) as well as five membered heterocyclic aromatic rings (2-amine-thiazole, 2-amine-imidazole). Primary, secondary and tertiary amine groups are often found and their categorization into functional group or linker unit is somewhat arbitrary. Some of these amines have pK_a values that allow for a change in protonation state under physiological conditions, such as imidazole with a pK_a of 6.95. The classification of the fragments into functional units (see Scheme 2) leads to 26 different functional units (Supporting Information, Figure S6).

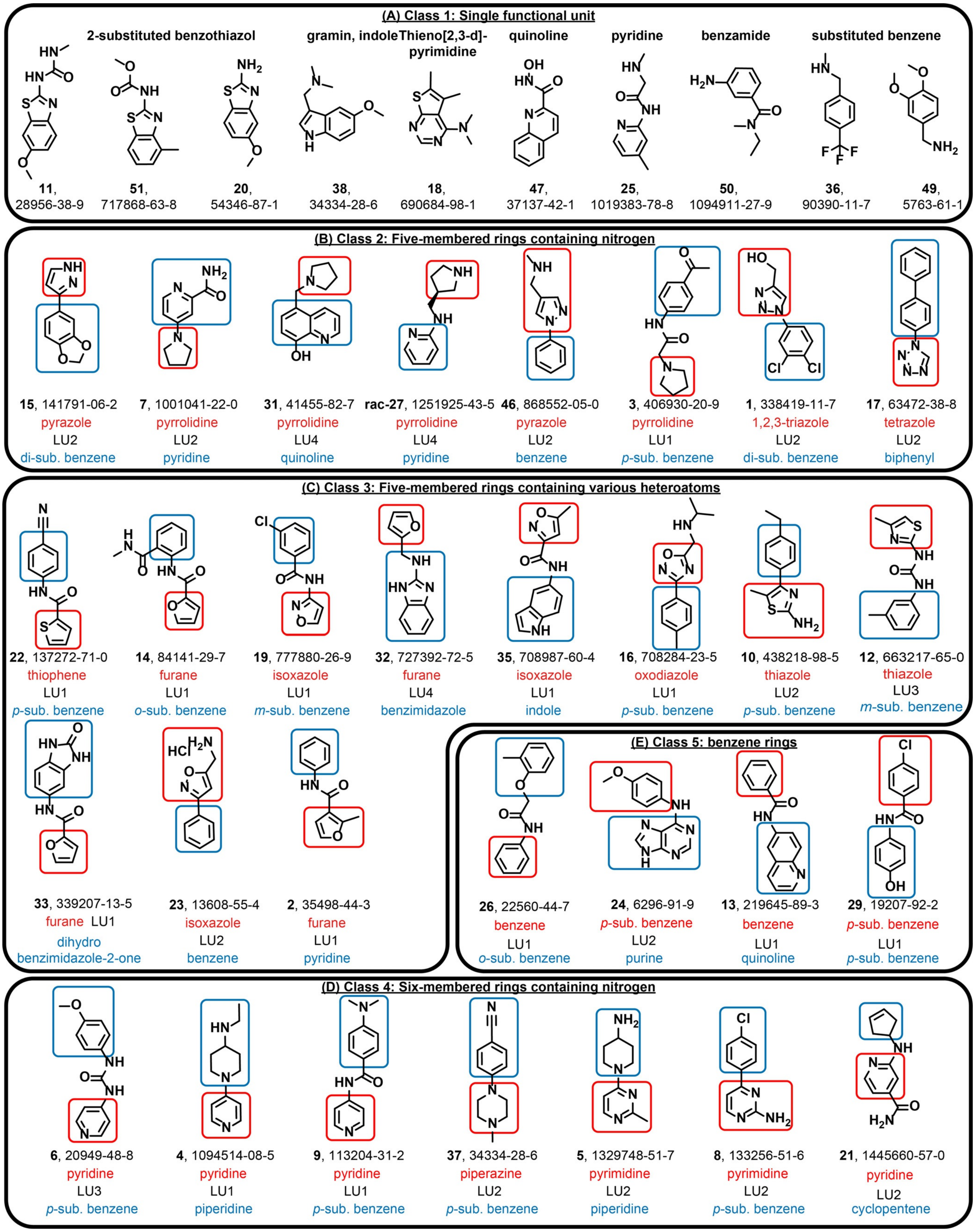


Scheme 1. Molecular descriptors and chemical diversity of RNA binders in the DSI-poised library. About 75% of the fragments showing binding to SCoV-2 RNA contain two functional units, connected by four different linking units. In case of LU3, the urea functional group likely contributes to binding affinity.

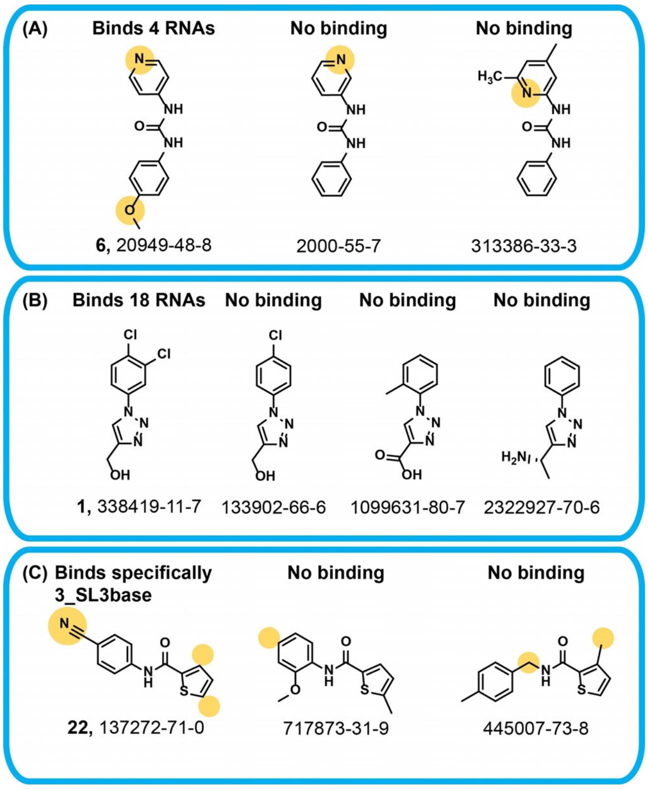
Next to the analysis of binders, comparison to functionally closely related fragments that do not bind is insightful (Scheme 3). Neither fragments containing carboxylic (842971-05-5, for compound identity, Supporting Information, Table S2. DSI-PL-768_Ligands) nor benzoic acid (3303-18-2, 1152510-62-7, 693776-70-4) functional groups, nor sulfone (24092-75-9) or sulfamide (1423029-76-8, 1388691-56-2), nor fragments with hydroxyl groups (74548-62-2, 1849283-80-2), and none of the 17 acetamides within DSI-PL are found among the hits. The library contains not only the para-substituted pyridine derivative 6, but also ortho- and meta-substituted pyridines 2000 55-7 and 313386-33-3 (Scheme 3A), but only para-substituted pyridines bind. Only thiophene-containing fragment 22, but not methyl-substituted thiophenes (717873-31-9 and 445007-73-8 shows binding to three RNA targets. A particularly intriguing example of surprisingly high selectivity for binding over not binding are four triazole-containing ligands (Scheme 3B). While triazole 1 is the most promiscuous binder, the closely related ligands 133902-66-6, 1099631-80-7, and 2322927-70-6 do not bind at all.

Hits containing two functional units can be classified into four additional classes. Class 2: five-membered heterocyclic rings containing a single nitrogen heteroatom, Class 3: five-membered rings with several heteroatoms, in particular nitrogen, oxygen and sulfur, Class 4: six-membered heterocyclic rings containing nitrogen atom(s), and Class 5: substituted and unsubstituted benzenes. In class 1, we find three differently 2-substituted benzo-thiazoles with either an amino group (compound **20**) or two different linker units attached (methyl urea linker (**11**) and methyl carbamate (**51**)). Related to the benzothiazoles are the fragment hits thieno[2,3-*d*]pyrimidine (18) and the C7-methoxy-substituted gramin (**37**). A last single functional unit-containing fragment is quinoline, substituted at C2 with a hydroxamic acid functional group (**47**). Furthermore, we find two methylene-amino-substituted benzenes (**36** and **49**).

For a subset of hits (Table 1) that bound PK and 3_s2m, we determined the dissociations constants using ligand-observed ^{1}H NMR-based titrations. In general, the determined dissociation constants K_D for the fragment hits ranged between 64 and 1318 μM (Table 1 and Supporting Information, Figure S7). Fragment 4 bound with highest affinity both towards PK and 3_s2m (Figure 3A). In addition to the quantification of the T_2-reduction obtained from the primary screen in mixtures at 20-fold excess of ligand over RNA ([RNA] = 10 μM) (Supporting Information, Figure S4), we determined the transverse relaxation time T_2 for the binding ligands at 100 μM concentration at an RNA concentration of 35 μM. Within the subset of the fragment hits, the ligands with the highest degree of T_2-reduction also showed the strongest affinity. For ligands with a K_D < 100 μM, binding affinity correlated with the T_2-reduction (Q^{bind}). No correlation was, however, observed for K_D > 100 μM. This finding suggests that T_2 is a qualitative indicator for binding strength of


(A) Class 1: Single functional unit
2-substituted benzothiazol
gramin, indole
Thieno[2,3-d]-pyrimidine
quinoline
pyridine
benzamide
substituted benzene
11, 28956-38-9
51, 717868-63-8
20, 54346-87-1
38, 34334-28-6
18, 690684-98-1
47, 37137-42-1
25, 1019383-78-8
50, 1094911-27-9
36, 90390-11-7
49, 5763-61-1
(B) Class 2: Five-membered rings containing nitrogen
15, 141791-06-2 pyrazole LU2 di-sub. benzene
7, 1001041-22-0 pyrrolidine LU2 pyridine
31, 41455-82-7 pyrrolidine LU4 quinoline
rac-27, 1251925-43-5 pyrrolidine LU4 pyridine
46, 868552-05-0 pyrazole LU2 benzene
3, 406930-20-9 pyrrolidine LU1 p-sub. benzene
1, 338419-11-7 1,2,3-triazole LU2 di-sub. benzene
17, 63472-38-8 tetrazole LU2 biphenyl
(C) Class 3: Five-membered rings containing various heteroatoms
22, 137272-71-0 thiophene LU1 p-sub. benzene
14, 84141-29-7 furane LU1 o-sub. benzene
19, 777880-26-9 isoxazole LU1 m-sub. benzene
32, 727392-72-5 furane LU4 benzimidazole
35, 708987-60-4 isoxazole LU1 indole
16, 708284-23-5 oxodiazole LU1 p-sub. benzene
10, 438218-98-5 thiazole LU2 p-sub. benzene
12, 663217-65-0 thiazole LU3 m-sub. benzene
33, 339207-13-5 furane LU1 dihydro benzimidazole-2-one
23, 13608-55-4 isoxazole LU2 benzene
2, 35498-44-3 furane LU1 pyridine
HCl
(E) Class 5: benzene rings
26, 22560-44-7 benzene LU1 o-sub. benzene
24, 6296-91-9 p-sub. benzene LU2 purine
13, 219645-89-3 benzene LU1 quinoline
29, 19207-92-2 p-sub. benzene LU1 p-sub. benzene
(D) Class 4: Six-membered rings containing nitrogen
6, 20949-48-8 pyridine LU3 p-sub. benzene
4, 1094514-08-5 pyridine LU1 piperidine
9, 113204-31-2 pyridine LU1 p-sub. benzene
37, 34334-28-6 piperazine LU2 p-sub. benzene
5, 1329748-51-7 pyrimidine LU2 piperidine
8, 133256-51-6 pyrimidine LU2 p-sub. benzene
21, 1445660-57-0 pyridine LU2 cyclopentene

Scheme 2. Classification into five classes of the 40 ligands that bind to the 15 regulatory RNAs. LU = linking unit.



Scheme 3. Comparison of binding and non-binding fragments with closely related chemical structure.

fragments to one specific target for binding in fast exchange and with a low micromolar K_D.

As a follow-up and in the quest for finding molecules with improved affinity, we used the fragment hits obtained for five RNAs (3_s2m, 3_SL1, 5_SL4, PK and 5′-UTR) to identify commercially available potential binders and obtained one compound (D01) that bound to both PK and 3_s2m in single-digit μM affinity (Figure 3B). At the same time, D01 showed no saturable binding towards control RNAs including the most stable RNA UUCG-tetraloop (Supporting Information, Figure S8). For fragments binding with K_D above 200 μM, it is our experience[55,56] that ligand binding sites can often not be detected even at maximum possible excess of ligand (limited by solubility) over RNA for a number of reasons. In fact, we show this lack of binding site detection also for one of the fragments that binds with a K_D of 46 μM (fragment 8, 133256-51-6) in the Supporting Information, Figure S9A. By contrast, for the high-affinity binder D01, we provide experimental mapping of its binding site to PK (Figure 3), which is in line with the binding site determined by in-line probing (Supporting Information, Figure S9C). NMR titration experiments were performed using ^{1}H, ^{15}N/^{13}C -HSQC of PK. In these experiments, many of the signals showed either severe line-broadening or CSPs strongly indicating binding. Mapping of these signals onto the cryo-electron microscopy structure of PK shows that D01 mainly binds to the stem 3 (Figure 3C and Supporting Information, Figure S9B).

Conclusion

We conducted here an NMR-based fragment screening against all regulatory RNA elements of the viral genome of SCoV-2. Our results show that the fragments of the DSI-PL library can cover almost the entire structural space of the viral RNA genome. Thus, the SCoV-2-RNA genome can be targeted by ligands of low molecular weight. We detect that several fragments already show specificity at this very first stage of drug development (Supporting Information, Figure S6). Assuming that increasing structural complexity enhances the potential to achieve specificity for small molecules,[32,57] among the 15 RNAs, PK and 3_SL3base represent the most promising targets for follow-up development of lead molecules.

The current state-of-the-art discovery of small molecules targeting RNAs (RNA drug discovery) and the current status of this field have been described by Juru and Hargrove,[53] by Meyer et al[9] and by Warner et al.[57] The screening campaign conducted here supports previous reports that identified the existence of a privileged chemical space for targeting RNA.[58] Delineating RNA target space into secondary structure motifs, definable from studies of isolated RNA elements, finds further support by a recent paper from the Das group that provides FARFAR models for all conserved viral RNA elements.[59]

Here, we use a fragment library that has previously been used to screen proteins,[44] but also 14 different RNAs, and five DNAs.[55] We use NMR as primary screen because it not only provides information on chemical purity of both ligand and RNA target, but also on affinity and binding site. Identifying RNA fragments with affinities between 60–400 μM allows us to identify known and commercially available binders containing similar structural motifs. A thus identified binder, D01, binds with 6 μM affinity to PK and 3_s2m, allowing fragment linking approaches.[32,60]

It is apparent that all ligands contain at least one, but most often two aromatic or heteroaromatic

Table 1: Affinities and T_2 relaxation times of the RNA binders.

ID	PK			3_s2m		
	T_2 [Q^{bind}] primary screen[a]	T_2 [ms] determined[b]	K_D [μM][c]	T_2 [Q^{bind}] primary screen[a]	T_2 [ms] determined[b]	K_D [μM][c]
4	54	83 ± 6	90 ± 13	48	125 ± 12	64 ± 8
5	38	68 ± 12	160 ± 6	15	98 ± 15	206 ± 5
3	39	84 ± 6	336 ± 18	39	106 ± 7	212 ± 30
23	n.a.	n.a.	n.a.	26	131 ± 10	678 ± 25
2	31	87	456 ± 33	n.a.	n.a.	n.a.
D01	n.a.	n.a.	6 ± 1.6	n.a.	n.a.	4.4 ± 2.3

[a] Primary screen; ligand in mixtures ([RNA]:[Ligand] = 1:20), [RNA] = 10 μM. [b] Single ligand measurements ([RNA]:[Ligand] = 1:3), [RNA] = 35 μM. [c] ^{1}H NMR-based ligand observed titrations [Ligand 100 μM; 0 to 250 μM].

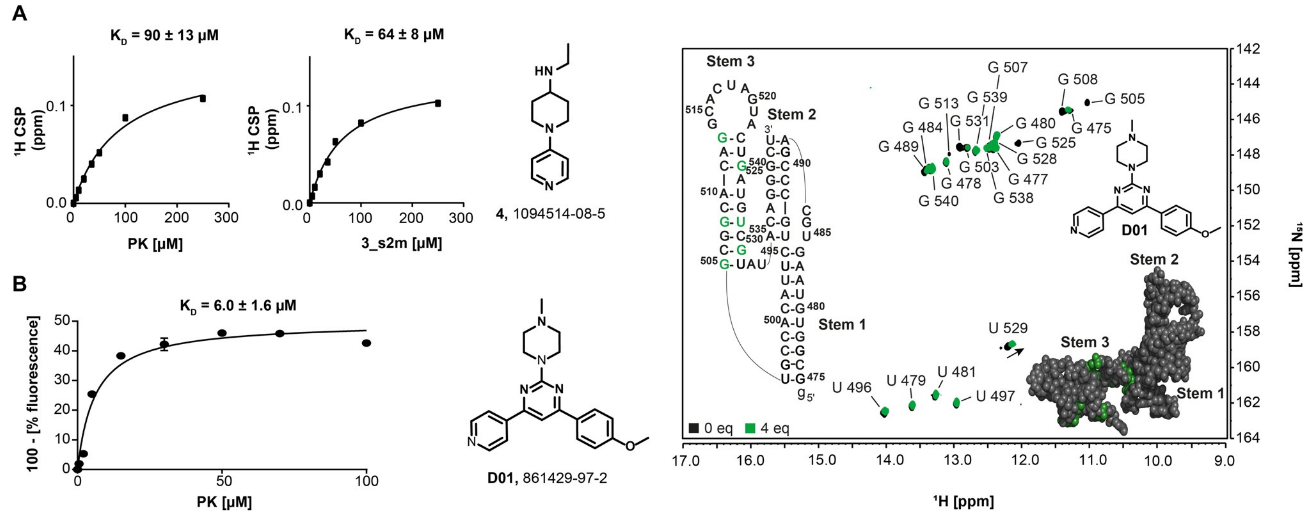


Figure 3. Affinities of binders and mapping of binding sites in SCoV-2 RNAs. A) NMR-based (ligand-observed; protons chosen to follow CSP are indicated) titration curves for the interaction of fragment **4** with PK and 3_s2m. B) Fluorescence-based titration curve and dissociation constant of **D01** with PK. C) Interaction of **D01** with PK monitored by RNA-observed NMR. ^{1}H, ^{15} N-HSQC overlay of PK alone and in presence of four equivalents of **D01** (green). Addition of **D01** to PK results in CSPs or line broadening beyond detection. Affected nucleotides are mapped onto the PK secondary and 3D structures (PDB:7O7Z[61]).

rings (Supporting Information, Figure S6). Heteroaromatic rings with two condensed rings including indoles, purines, quinolones (class I, Scheme 2) are found but not particularly enriched. Many of the hits do not have large planar moieties nor very basic sites. For example, **1** as tertiary amine is strongly basic, 2-aminopyridine has a pK_a of around 3.5. By contrast, 1,2,3 triazole is already not very basic but shows very specific binding (Scheme 3). Further, comprehensively comparing the distribution of functional units between the hits and the non-hits across the library suggests that the hits are enriched with pyrimidines and benzimdazoles (Supporting Information, Table S7). Thus, the absence of trivial functional units in the DSI-PL appears particularly striking: fusion of monocyclic aromatic or heteroaromatic ring systems with H-bond donors or -acceptors seems sufficient to reach binding.

The future analysis of our screening results might open new routes towards binding selectivity. The screening campaign was conducted for 20 different RNA elements. It also included RNAs that combined two or more of the individually screened elements (e.g. 5_SL1234) and even the entire 5′- and 3′-untranslated regions of SCoV-2 comprising more than 337 nts for each of the constructs. The size range of these targets is markedly different. At this point, exact quantification of the NMR screening experiments relies on approximations including effective overall rotational correlation times of the RNA targets, relative population of free vs. bound ligand, and their on- and off-rates. These assumptions are, however, no longer fulfilled when one compares pools of RNAs with significantly different molecular weight and potential anisotropic tumbling. Within a given target, however, the T_2-reduction recorded within the NMR screening of all 768 fragments provides a qualitative indication for binding with K_D values below 100 μM (Table 1).

We analyzed the investigated RNA target space towards sequence-derived properties and derived unique target sites within the SCoV-2 genome that are conserved among coronaviruses. Exceptionally rare mutations in SCoV-2 during 2020 have maintained the target space for ligands up to now. By analysis of the experimental hits, we can derive privileged RNA target space from a medicinal chemistry perspective. Further, our approach relies on the hypothesis that the viral genome can be specifically targeted even in the presence of human cellular RNA. The herein provided analysis of secondary structure abundance (Supporting Information, Table S1) is restricted to the viral RNA genome and will have to be extended to the entire cell-specific human transcriptome. However, given all these theoretical considerations it is striking to recognize that small molecules have been developed that target splicing[11,56] in a highly specific manner.

The results as well as the methodological approach presented here will impact medicinal chemistry approaches but also cellular targeting of SCoV-2 RNA. For example, as immediate follow-up, we currently conduct secondary NMR screens with commercially available compounds using the identified ligands as guide. We consider our broad, systematic and coherent identification of binders to be a significant step forward. The focus of our study is to experimentally validate the druggability of the SCoV-2 genome. In fact, this potential had so far only been predicted from theoretical studies.[59] Our efforts extend on-going studies targeting the viral proteome. In conclusion, we establish here the conserved RNA elements as potential space for small molecule targeting towards Coronavirus-specific medication, even beyond SCoV-2.

Acknowledgements

The work has been conducted in the international consortium of Covid19-NMR (covid19-nmr.de). Work at BMRZ is supported by the state of Hesse. Work in Covid19-nmr was supported by the Goethe Corona Funds, by the IWB-EFRE-program 20007375 of state of Hesse, the DFG through CRC902: "Molecular Principles of RNA-based regulation." and through infrastructure funds (project numbers: 277478796, 277479031, 392682309, 452632086, 70653611) and by European Union's Horizon 2020 research and innovation program iNEXT-discovery under grant agreement No 871037. We thank Peter Maas (SPECS) for providing the compound D01 as a kind gift. Open access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: Covid19-nmr · fragment screening · NMR spectroscopy · RNA · SARS-CoV-2

[1] M. Macchiagodena, M. Pagliai, P. Procacci, *Chem. Phys. Lett.* **2020**, *750*, 137489.
[2] S. Barage, A. Karthic, R. Bavi, N. Desai, R. Kumar, V. Kumar, K. W. Lee, *J. Biomol. Struct. Dyn.* **2020**, 1 – 18.
[3] A. Carino, F. Moraca, B. Fiorillo, S. Marchianò, V. Sepe, M. Biagioli, C. Finamore, S. Bozza, D. Francisci, E. Distrutti, et al., *Front. Chem.* **2020**, *8*, 572885.
[4] M. T. J. Quimque, K. I. R. Notarte, R. A. T. Fernandez, M. A. O. Mendoza, R. A. D. Liman, J. A. K. Lim, L. A. E. Pilapil, J. K. H. Ong, A. M. Pastrana, A. Khan, et al., *J. Biomol. Struct. Dyn.* **2020**, 1 – 18.
[5] M. A. White, W. Lin, X. Cheng, *J. Phys. Chem. Lett.* **2020**, *11*, 9144 – 9151.
[6] C. Liu, X. Zhu, Y. Lu, X. Zhang, X. Jia, T. Yang, *J. Pharm. Anal.* **2020**, https://doi.org/10.1016/j.jpha.2020.08.002.
[7] J. O. Ogidigo, E. A. Iwuchukwu, C. U. Ibeji, O. Okpalefe, M. E. S. Soliman, *J. Biomol. Struct. Dyn.* **2020**, 1 – 18.
[8] H. S. Haniff, Y. Tong, X. Liu, J. L. Chen, B. M. Suresh, R. J. Andrews, J. M. Peterson, C. A. O'Leary, R. I. Benhamou, W. N. Moss, et al., *ACS Cent. Sci.* **2020**, *6*, 1713 – 1721.
[9] S. M. Meyer, C. C. Williams, Y. Akahori, T. Tanaka, H. Aikawa, Y. Tong, J. L. Childs-Disney, M. D. Disney, *Chem. Soc. Rev.* **2020**, *49*, 7167 – 7199.
[10] M. L. Kelly, C.-C. Chu, H. Shi, L. R. Ganser, H. P. Bogerd, K. Huynh, Y. Hou, B. R. Cullen, H. M. Al-Hashimi, *RNA* **2020**, rna.076257.120.
[11] S. Campagne, S. Boigner, S. Rüdisser, A. Moursy, L. Gillioz, A. Knörlein, J. Hall, H. Ratni, A. Cléry, F. H.-T. Allain, *Nat. Chem. Biol.* **2019**, *15*, 1191 – 1198.
[12] Y. Wang, R. R. Rando, *Chem. Biol.* **1995**, *2*, 281 – 290.
[13] M. F. Bardaro, Z. Shajani, K. Patora-Komisarska, J. A. Robinson, G. Varani, *Nucleic Acids Res.* **2009**, *37*, 1529 – 1540.
[14] D. Raghunathan, V. M. Sánchez-Pedregal, J. Junker, C. Schwiegk, M. Kalesse, A. Kirschning, T. Carlomagno, *Nucleic Acids Res.* **2006**, *34*, 3599 – 3608.
[15] F. A. Abulwerdi, M. D. Shortridge, J. Sztuba-Solinska, R. Wilson, S. F. J. Le Grice, G. Varani, J. S. J. Schneekloth, *J. Med. Chem.* **2016**, *59*, 11148 – 11160.
[16] M. Zeiger, S. Stark, E. Kalden, B. Ackermann, J. Ferner, U. Scheffer, F. Shoja-Bazargani, V. Erdel, H. Schwalbe, M. W. Göbel, *Bioorg. Med. Chem. Lett.* **2014**, *24*, 5576 – 5580.
[17] S. J. Park, Y. G. Kim, H. J. Park, *J. Am. Chem. Soc.* **2011**, *133*, 10094 – 10100.
[18] D. Kim, J.-Y. Lee, J.-S. Yang, J. W. Kim, V. N. Kim, H. Chang, *Cell* **2020**, *181*, 914 – 921.
[19] I. Manfredonia, C. Nithin, A. Ponce-Salvatierra, P. Ghosh, T. K. Wirecki, T. Marinus, N. S. Ogando, E. J. Snijder, M. J. van Hemert, J. M. Bujnicki, et al., *Nucleic Acids Res.* **2020**, *48*, 12436 – 12452.
[20] R. Rangan, I. N. Zheludev, R. Das, *RNA* **2020**, *26*, 937 – 959.
[21] Z. Martina, H. Christina, L. Le, D.-C. Jesse, Y.-C. Liang, S. M. Christian, A. G. Monaghan, A. A. Kennedy, J. D. Yesselman, R. R. Gifford, et al., *bioRxiv* **2020**, DOI: https://doi.org/10.1101/2020.12.05.409821.
[22] J. A. Kelly, A. N. Olson, K. Neupane, S. Munshi, J. San Emeterio, L. Pollack, M. T. Woodside, J. D. Dinman, *J. Biol. Chem.* **2020**, *295*, 10741 – 10748.
[23] A. Wacker, J. E. Weigand, S. R. Akabayov, N. Altincekic, J. K. Bains, E. Banijamali, O. Binas, J. Castillo-Martinez, E. Cetiner, B. Ceylan, et al., *Nucleic Acids Res.* **2020**, *48*, 12415 – 12435.
[24] M. Pellecchia, I. Bertini, D. Cowburn, C. Dalvit, E. Giralt, W. Jahnke, T. L. James, S. W. Homans, H. Kessler, C. Luchinat, et al., *Nat. Rev. Drug Discovery* **2008**, *7*, 738 – 745.
[25] O. Binas, V. de Jesus, T. Landgraf, A. E. Völklein, J. Martins, D. Hymon, H. Berg, J. K. Bains, T. Biedenbänder, B. Fürtig, et al., *ChemBioChem* **2020**, *22*, 423 – 433.
[26] S. Sreeramulu, C. Richter, T. Kuehn, K. Azzaoui, M. J. J. Blommers, R. Del Conte, M. Fragai, N. Trieloff, P. Schmieder, M. Nazare, et al., *J. Biomol. NMR* **2020**, *74*, 555 – 563.
[27] C. Huang, K. G. Lokugamage, J. M. Rozovics, K. Narayanan, B. L. Semler, S. Makino, *PLoS Pathog.* **2011**, *7*, e1002433 – e1002433.
[28] C. W. Lee, L. Li, D. P. Giedroc, *FEBS Lett.* **2011**, *585*, 1049 – 1053.
[29] O. Ziv, J. Price, L. Shalamova, T. Kamenova, I. Goodfellow, F. Weber, E. A. Miska, *Mol. Cell* **2020**, *80*, 1067 – 1077.
[30] D. Yang, J. L. Leibowitz, *Virus Res.* **2015**, *206*, 120 – 133.
[31] K. Neupane, S. Munshi, M. Zhao, D. B. Ritchie, S. M. Ileperuma, M. T. Woodside, *J. Mol. Biol.* **2020**, *432*, 5843 – 5847.
[32] T. Hermann, *Wiley Interdiscip. Rev. RNA* **2016**, *7*, 726 – 743.
[33] R. Zust, T. B. Miller, S. J. Goebel, V. Thiel, P. S. Masters, *J. Virol.* **2008**, *82*, 1214 – 1228.
[34] S. J. Goebel, T. B. Miller, C. J. Bennett, K. A. Bernard, P. S. Masters, *J. Virol.* **2007**, *81*, 1274 – 1287.
[35] I. Manfredonia, D. Incarnato, *Biochem. Soc. Trans.* **2020**, *48*, 341 – 352.
[36] Z. Miao, A. Tidu, G. Eriani, F. Martin, *RNA Biol.* **2020**, *17*, 1 – 10.
[37] J. Zhao, J. Qiu, S. Aryal, J. L. Hackett, J. Wang, *Viruses* **2020**, *12*, 1473.
[38] J. Navarro Gonzalez, A. S. Zweig, M. L. Speir, D. Schmelter, K. R. Rosenbloom, B. J. Raney, C. C. Powell, L. R. Nassar, N. D. Maulding, C. M. Lee, et al., *Nucleic Acids Res.* **2021**, *49*, D1046 – D1057.
[39] S. J. Goebel, B. Hsue, T. F. Dombrowski, P. S. Masters, *J. Virol.* **2004**, *78*, 669 – 682.
[40] N. B. Leontis, J. Stombaugh, E. Westhof, *Nucleic Acids Res.* **2002**, *30*, 3497 – 3531.
[41] J. E. Barrick, R. R. Breaker, *Genome Biol.* **2007**, *8*, R239.
[42] F. Wu, S. Zhao, B. Yu, Y. M. Chen, W. Wang, Z. G. Song, Y. Hu, Z. W. Tao, J. H. Tian, Y. Y. Pei, et al., *Nature* **2020**, *579*, 265 – 269.
[43] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R. A. Neher, *Bioinformatics* **2018**, *34*, 4121 – 4123.
[44] O. B. Cox, T. Krojer, P. Collins, O. Monteiro, R. Talon, A. Bradley, O. Fedorov, J. Amin, B. D. Marsden, J. Spencer, et al., *Chem. Sci.* **2016**, *7*, 2322 – 2330.
[45] A. Douangamath, D. Fearon, P. Gehrtz, T. Krojer, P. Lukacik, C. D. Owen, E. Resnick, C. Strain-Damerell, A. Aimon, P. Ábrányi-Balogh, et al., *Nat. Commun.* **2020**, *11*, 5047.
[46] C. Dalvit, G. Fogliatto, A. Stewart, M. Veronesi, B. Stockman, *J. Biomol. NMR* **2001**, *21*, 349 – 359.
[47] C. Dalvit, *J. Biomol. NMR* **1998**, *11*, 437 – 444.

[48] P. J. Hajduk, E. T. Olejniczak, S. W. Fesik, *J. Am. Chem. Soc.* **1997**, *119*, 12257 – 12261.

[49] D. R. Calabrese, C. M. Connelly, J. S. Schneekloth, in *Methods Enzymol.* **2019**, *623*, 131 – 149, DOI: https://doi.org/10.1016/bs.mie.2019.05.030.

[50] J. H. Davis, B. F. Dunican, S. A. Strobel, *Biochemistry* **2011**, *50*, 7236 – 7242.

[51] A. Ciulli, G. Williams, A. G. Smith, T. L. Blundell, C. Abell, *J. Med. Chem.* **2006**, *49*, 4992 – 5000.

[52] A.-S. Bretonnet, A. Jochum, O. Walker, I. Krimm, P. Goekjian, O. Marcillat, J.-M. Lancelin, *J. Med. Chem.* **2007**, *50*, 1865 – 1875.

[53] A. Umuhire Juru, A. E. Hargrove, *J. Biol. Chem.* **2021**, *296*, 100191.

[54] N. F. Rizvi, J. P. Santa Maria, A. Nahvi, J. Klappenbach, D. J. Klein, P. J. Curran, M. P. Richards, C. Chamberlin, P. Saradjian, J. Burchard, et al., *SLAS Discovery* **2020**, *25*, 384 – 396.

[55] O. Binas, V. Jesus, T. Landgraf, A. E. Völklein, J. Martins, D. Hymon, J. Kaur Bains, H. Berg, T. Biedenbänder, B. Fürtig, et al., *ChemBioChem* **2020**, *22*, 423 – 433.

[56] A. Garcia-Lopez, F. Tessaro, H. R. A. Jonker, A. Wacker, C. Richter, A. Comte, N. Berntenis, R. Schmucki, K. Hatje, O. Petermann, et al, *Nat. Commun.* **2018**, *9*, 2032.

[57] K. D. Warner, C. E. Hajdin, K. M. Weeks, *Nat. Rev. Drug Discovery* **2018**, *17*, 547 – 558.

[58] B. S. Morgan, J. E. Forte, R. N. Culver, Y. Zhang, A. E. Hargrove, *Angew. Chem. Int. Ed.* **2017**, *56*, 13498 – 13502; *Angew. Chem.* **2017**, *129*, 13683 – 13687.

[59] R. Rangan, A. M. Watkins, J. Chacon, R. Kretsch, W. Kladwang, I. N. Zheludev, J. Townley, G. Thain, R. Das, *Nucleic Acids Res.* **2021**, *49*, 3092 – 3108.

[60] M. D. Disney, A. J. Angelbello, *Acc. Chem. Res.* **2016**, *49*, 2698 – 2704.

[61] P. R. Bhatt, A. Scaiola, G. Loughran, M. Leibundgut, A. Kratzel, R. Meurs, R. Dreos, K. M. O'Connor, A. McMillan, J. W. Bode, V. Thiel, D. Gatfield, J. F. Atkins, N. Ban, *Science* **2021**, *372*, 1306 – 1313.

Manuscript received: March 18, 2021
Revised manuscript received: June 3, 2021
Accepted manuscript online: June 23, 2021
Version of record online: August 6, 2021

5.15 **Research article:** Comprehensive Fragment Screening of the SARS-CoV-2 Proteome Explores Novel Chemical Space for Drug Development

Hannes Berg*, Maria A. Wirtz Martin*, Nadide Altincekic*, Islam Alshamleh, Jasleen Kaur Bains, Julius Blechar, Betül Ceylan, Vanessa de Jesus, Karthikeyan Dhamotharan, Christin Fuks, Santosh L. Gande, Bruno Hargittay, Katharina F. Hohmann, Marie T. Hutchinson, Sophie Marianne Korn, Robin Krishnathas, Felicitas Kutz, Verena Linhard, Tobias Matzel, Nathalie Meiser, Anna Niesteruk, Dennis J. Pyper, Linda Schulte, Sven Trucks, Kamal Azzaoui, Marcel J J Blommers, Yojana Gadiya, Reagon Karki, Andrea Zaliani, Philip Gribbon, Marcius da Silva Almeida, Cristiane Dinis Anobom, Anna Lina Bula, Matthias Buetikofer, Ícaro Putinhon Caruso, Isabella Caterina Felli, Andrea T Da Poian, Gisele Cardoso de Amorim, Nikolaos K Fourkiotis, Angelo Gallo, Dhiman Ghosh, Francesco Gomes-Neto, Oksana Gorbatyuk, Bing Hao, Vilius Kurauskas, Lauriane Lecoq, Yunfeng Li, Nathane Cunha Mebus-Antunes, Miguel Mompean, Thais Cristtina Neves-Martins, Marti Ninot-Pedrosa, Anderson S Pinheiro, Letizia Pontoriero, Yulia Pustovalova, Roland Riek, Angus Robertson, Marie Jose Abi Saad, Miguel A. Treviño, Aikaterini C. Tsika, Fabio C.L. Almeida, Ad Bax, Katherine Henzler-Wildman, Jeffrey C Hoch, Kristaps Jaudzems, Douglas V. Laurents, Julien Orts, Roberta Pieratelli, Georgios A. Spyroulias, Elke Duchardt-Ferner, Jan Ferner, Boris Fuertig, Martin Hengesbach, Frank Löhr, Nusrat Qureshi, Christian Richter, Krishna Saxena, Andreas Schlundt, Sridhar Sreeramulu, Anna Wacker, Julia E Weigand, Julia Wirmer-Bartoschek, Jens Woehnert and Harald Schwalbe; *Angewandte Chemie International Edition* **46**, e202205858 (2022)

* These authors contributed equally to this work.

In this work, an NMR-based fragment screening against 25 SARS-CoV-2 proteins was performed with 768 compounds using ^{1}H-1D NMR binding assays. Screening experiments were performed with 64 mixtures containing 12 fragments in each mixture. Ligand-binding analysis was based on changes in the ^{1}H proton signals of the ligand in the absence and presence of protein. Binding affinities of these ligands were demonstrated in the micromolar to millimolar range. These fragment-based screening results provide insight for novel drug development.

H. Berg, M. A. Wirtz Martin, and N. Altincekic performed further NMR-based screening experiments, analyzed the data, and prepared the manuscript. The author of this thesis was part of the Covid-19 NMR consortium and protein production team that prepared and purified various isotope-labeled protein samples for NMR spectroscopy.

https://doi.org/10.1002/anie.202205858

A Journal of the Gesellschaft Deutscher Chemiker

Angewandte Chemie

GDCh

International Edition

www.angewandte.org

Accepted Article

Title: Comprehensive Fragment Screening of the SARS-CoV-2 Proteome Explores Novel Chemical Space for Drug Development

Authors: Hannes Berg, Maria A. Wirtz Martin, Nadide Altincekic, Islam Alshamleh, Jasleen Kaur Bains, Julius Blechar, Betül Ceylan, Vanessa de Jesus, Karthikeyan Dhamotharan, Christin Fuks, Santosh L. Gande, Bruno Hargittay, Katharina F. Hohmann, Marie T. Hutchinson, Sophie Marianne Korn, Robin Krishnathas, Felicitas Kutz, Verena Linhard, Tobias Matzel, Nathalie Meiser, Anna Niesteruk, Dennis J. Pyper, Linda Schulte, Sven Trucks, Kamal Azzaoui, Marcel J J Blommers, Yojana Gadiya, Reagon Karki, Andrea Zaliani, Philip Gribbon, Marcius da Silva Almeida, Cristiane Dinis Anobom, Anna Lina Bula, Matthias Buetikofer, Ícaro Putinhon Caruso, Isabella Caterina Felli, Andrea T Da Poian, Gisele Cardoso de Amorim, Nikolaos K Fourkiotis, Angelo Gallo, Dhiman Ghosh, Francesco Gomes-Neto, Oksana Gorbatyuk, Bing Hao, Vilius Kurauskas, Lauriane Lecoq, Yunfeng Li, Nathane Cunha Mebus-Antunes, Miguel Mompean, Thais Cristtina Neves-Martins, Marti Ninot-Pedrosa, Anderson S Pinheiro, Letizia Pontoriero, Yulia Pustovalova, Roland Riek, Angus Robertson, Marie Jose Abi Saad, Miguel A Treviño, Aikaterini C Tsika, Fabio C.L. Almeida, Ad Bax, Katherine Henzler-Wildman, Jeffrey C Hoch, Kristaps Jaudzems, Douglas V Laurents, Julien Orts, Roberta Pieratelli, Georgios A Spyroulias, Elke Duchardt-Ferner, Jan Ferner, Boris Fuertig, Martin Hengesbach, Frank Löhr, Nusrat Qureshi, Christian Richter, Krishna Saxena, Andreas Schlundt, Sridhar Sreeramulu, Anna Wacker, Julia E Weigand, Julia Wirmer-Bartoschek, Jens Woehnert, and Harald Schwalbe

To be cited as: *Angew. Chem. Int. Ed.* **2022**, e202205858

WILEY-VCH

Link to VoR: https://doi.org/10.1002/anie.202205858

Comprehensive Fragment Screening of the SARS-CoV-2 Proteome Explores Novel Chemical Space for Drug Development

Hannes Berg,[a,b‡] Maria A. Wirtz Martin,[a,b‡] Nadide Altincekic,[a,b‡] Islam Alshamleh,[a,b] Jasleen Kaur Bains,[a,b] Julius Blechar,[a] Betül Ceylan,[a,b] Vanessa de Jesus,[a,b] Karthikeyan Dhamotharan,[b,c] Christin Fuks,[a] Santosh L. Gande,[a,b] Bruno Hargittay,[a,b] Katharina F. Hohmann,[a,b] Marie T. Hutchison,[a,b] Sophie Marianne Korn,[b,c] Robin Krishnathas,[a,b] Felicitas Kutz,[a,b] Verena Linhard,[a,b] Tobias Matzel,[a,b] Nathalie Meiser,[a] Anna Niesteruk,[a,b] Dennis J. Pyper,[a,b] Linda Schulte,[a,b] Sven Trucks,[a] Kamal Azzaoui,[d] Marcel J. J. Blommers,[d] Yojana Gadiya,[e,f] Reagon Karki,[e,f] Andrea Zaliani,[e,f] Philip Gribbon,[e,f] Marcius da Silva Almeida,[g,h] Cristiane Dinis Anobom,[i,j] Anna L. Bula,[k] Matthias Bütikofer,[l] Ícaro Putinhon Caruso,[g,m] Isabella Caterina Felli,[n,o] Andrea T. Da Poian,[g] Gisele Cardoso de Amorim,[g,p] Nikolaos K. Fourkiotis,[q] Angelo Gallo,[q,r] Dhiman Ghosh,[l] Francisco Gomes-Neto,[i,s] Oksana Gorbatyuk,[t] Bing Hao,[t] Vilius Kurauskas,[u] Lauriane Lecoq,[v] Yunfeng Li,[t] Nathane Cunha Mebus-Antunes,[g] Miguel Mompeán,[w] Thais Cristtina Neves-Martins,[g] Martí Ninot-Pedrosa,[v] Anderson S. Pinheiro,[j] Letizia Pontoriero,[n,o] Yulia Pustovalova,[t] Roland Riek,[l] Angus J. Robertson,[x] Marie Jose Abi Saad,[y] Miguel Á. Treviño,[w] Aikaterini C. Tsika,[q] Fabio C. L. Almeida,[g,i] Ad Bax,[x] Katherine Henzler-Wildman,[u] Jeffrey C. Hoch,[t] Kristaps Jaudzems,[k] Douglas V. Laurents,[w] Julien Orts,[y] Roberta Pierattelli,[n,o] Georgios A. Spyroulias,[q] Elke Duchardt-Ferner,[b,c] Jan Ferner,[a,b] Boris Fürtig,[a,b] Martin Hengesbach,[a] Frank Löhr,[b,c] Nusrat Qureshi,[a,b] Christian Richter,[a,b] Krishna Saxena,[a,b] Andreas Schlundt,[b,c] Sridhar Sreeramulu,[a,b] Anna Wacker,[a,b] Julia E. Weigand,[z] Julia Wirmer-Bartoschek,[a,b] Jens Wöhnert,[b,c] and Harald Schwalbe*[a,b]

[a,b] H. Berg[‡], M. A. Wirtz Martin[‡], N. Altincekic[‡], Dr. I. Alshamleh, J. K. Bains, J. Blechar, B. Ceylan, V. de Jesus, C. Fuks, Dr. S. L. Gande, B. Hargittay, K. F. Hohmann, M. T. Hutchison, R. Krishnathas, F. Kutz, V. Linhard, T. Matzel, N. Meiser, Dr. A. Niesteruk, D. J. Pyper, Dr. L. Schulte, Dr. S. Trucks, Dr. J. Ferner, Dr. B. Fürtig, Dr. M. Hengesbach, Dr. N. Qureshi, Dr. C. Richter, Dr. K. Saxena, Dr. S. Sreeramulu, Dr. A. Wacker, Dr. J. Wirmer-Bartoschek, Dr. Prof. Dr. H. Schwalbe
[a] Institute for Organic Chemistry and Chemical Biology, Goethe University Frankfurt, Frankfurt am Main, Germany
[b] Center of Biomolecular Magnetic Resonance (BMRZ), Goethe University Frankfurt, Frankfurt am Main, Germany
Max-von-LaueStrasse 7+9, 60438 Frankfurt am Main (Germany)
E-mail: schwalbe@nmr.uni-frankfurt.de

[b,c] K. Dhamotharan, Dr. S. M. Korn, Dr. E. Duchardt-Ferner, Dr. F. Löhr, Dr. A. Schlundt, Prof. Dr. J. Wöhnert
[c] Institute for Molecular Biosciences, Goethe University Frankfurt, Frankfurt am Main, Germany
Max-von-Laue Strasse 7+9, 60438 Frankfurt am Main (Germany)

[d] Dr. K. Azzaoui, Dr. M. J. J. Blommers
Saverna Therapeutics, Pumpmattenweg 3, 4105 Biel-Benken, Switzerland

[e,f] Y. Gadiya, R. Karki, A. Zaliani, Dr. P. Gribbon
[e] Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Screening Port, Schnackenburgallee 114, 22525 Hamburg, Germany
[f] Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), Theodor-Stern-Kai 7, 60596 Frankfurt am Main, Germany

[g,h] M. S. Almeida, A. T. Da Poian, N. C. Mebus-Antunes, T. C. Neves-Martins, F. C. L. Almeida
[g] Institute of Medical Biochemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[h] Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), Theodor-Stern-Kai 7, 60596 Frankfurt am Main, Germany

[i,j] C. D Anobom, A. S. Pinheiro
[i] National Center of Nuclear Magnetic Resonance (CNRMN), CENABIO, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[j] Department of Biochemistry, Institute of Chemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

[k] A. L. Bula, K. Jaudzems
[k] Latvian Institute of Organic Synthesis, Aizkraukles 21, LV-1006 Riga, Latvia

[l] M. Bütikofer, D. Ghosh, R. Riek
[l] ETH, Swiss Federal Institute of Technology, Laboratory of Physical Chemistry, HCI F217, Vladimir-Prelog-Weg 2, 8093 Zürich, Switzerland.

[g,m] Í. P. Caruso
[m] Multiuser Center for Biomolecular Innovation (CMIB), Department of Physics, São Paulo State University (UNESP), São José do Rio Preto, Brazil

[n,o] I. C. Felli, L. Pontoriero, R. Pierattelli
[n] Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, Sesto Fiorentino, 50019, Florence, Italy
[o] Department of Chemistry "Ugo Schiff", University of Florence, Via della Lastruccia 3-13, Sesto Fiorentino, 50019, Florence, Italy

[g,p] G. C. de Amorim
[p] Multidisciplinary Center for Research in Biology (NUMPEX), Campus Duque de Caxias Federal University of Rio de Janeiro, Duque de Caxias, Brazil

[q] N. K. Fourkiotis, A. C. Tsika, G. A. Spyroulias
[q] Department of Pharmacy, University of Patras, Patras, Greece

[q,r] A. Gallo,
[r] Department of Chemistry, University of Torino IT -10126 Torino. Italy

[i,s] F. Gomes-Neto
[s] Laboratory of Toxinology, Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil

[t] O. Gorbatyuk, B. Hao, Y. Li, Y. Pustovalova, J. C. Hoch
[t] Department of Molecular Biology and Biophysics UConn Health 263 Farmington Ave., Farmington, CT 06030-3305 USA

[u] V. Kurauskas, K. Henzler-Wildman
[u] Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706

[v] L. Lecoq, M. Ninot-Pedrosa
[v] Molecular Microbiology and Structural Biochemistry, UMR5086 CNRS/Université Lyon 1, 7, passage du Vercors, 69367 Lyon, France

[w] M. Mompeán, M. Á. Treviño, D. V. Laurents
[w] "Rocasolano" Institute for Physical Chemistry, CSIC
[x] A. J. Robertson, A. Bax
[x] Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases,
[y] M. J. Abi Saad, J. Orts
[y] University of Vienna, Department of Pharmaceutical Sciences, Josef-Holaubek-Platz 2, A-1090 Vienna, Austria
[z] Dr. J. E. Weigand
[z] Department of Biology, Technical University of Darmstadt, Darmstadt, Germany
[‡] These authors contributed equally to this work.

Abstract: SARS-CoV-2 (SCoV2) and its variants of concern pose serious challenges to the public health. The variants increased challenges to vaccines, thus necessitating for development of new intervention strategies including anti-virals. Within the international Covid19-NMR consortium, we have identified binders targeting the RNA genome of SCoV2. We established protocols for the production and NMR characterization of more than 80% of all SCoV2 proteins. Here, we performed an NMR screening using a fragment library for binding to 25 SCoV2 proteins and identified hits also against previously unexplored SCoV2 proteins. Computational mapping was used to predict binding sites and identify functional moieties (chemotypes) of the ligands occupying these pockets. Striking consensus was observed between NMR-detected binding sites of the main protease and the computational procedure. Our investigation provides novel structural and chemical space for structure-based drug design against the SCoV2 proteome.

Introduction

SARS-CoV-2 (SCoV2) is the cause for the COVID-19 pandemic resulting in more than 5 million deaths across the world and continues to pose serious challenges to public health and safety [1]. Countering the continuously evolving virus has not only seen an unprecedented success in the vaccine development but also given birth to several novel campaigns for anti-viral drug discovery [2,3], including the recently approved oral antivirals paxlovid (Pfizer) and molnupiravir (Merck & Co.) [4–6].

The extensively mutated and highly infective variant of SCoV2, Omicron [7], is resistant to several therapeutic antibodies [8,9], evades double immunization [8,10], and dominates the pandemic in 2022, calling for the development of new therapeutic strategies in combating the virus, specifically, by exploiting the conserved features [11,12].

The SCoV2 genome consists of an ~29.9 kb long positive-sense single-stranded RNA [13], two-thirds of which comprises the open-reading frames (ORF) 1a and 1ab. Both ORFs encode polyproteins, which are proteolytically processed into 16 different non-structural proteins (nsp1-nsp16) [14,15]. Four structural proteins: spike (S), envelope (E), membrane (M) and nucleocapsid (N) and nine additional accessory factors are expressed from the 13 ORFs located at the 3´ end of the viral genome. In total, the viral genome encodes for at least 28 peptides or proteins [16–18]. Repurposing of (approved) drugs has been actively pursued as a strategy to counter SCoV2 infections [19–22], however, with little clinical success [23]. Most of repurposed drugs were primarily an outcome of structure-based virtual screening campaigns and solely focused on a small fraction of the proteome, namely proteases (nsp3d, nsp5) or polymerase (nsp12) as targets [24–30]. Within the viral life cycle, the enzymes nsp3 (papain-like protease), nsp5 (main protease), nsp7•nsp8 (primase complex), nsp12 (primary RNA-dependent RNA polymerase (RdRp)), nsp13 (helicase), nsp14 (exoribonuclease) and the methyltransferases nsp14/nsp16 are important components of the replicase-transcriptase complex and hence are also listed as attractive drug targets [16,31]. X-ray crystallography and NMR have been successfully used to screen either fragments, approved drugs, or drugs in clinical trials, against a subset of key SCoV2 protein drug targets like nsp5, nsp3b, nsp13 and nsp14 [32–40].

The current drug development has typically focused its efforts around the two key viral proteins, a protease (nsp5) and a polymerase (nsp12, RdRp), and soon such a monotherapy can result in the virus developing resistance against the first-generation antivirals, thus warranting us to develop new antivirals involving different targets [41]. Recently, using a range of biochemical assays, several drugs were identified as inhibitors against a total of seven enzymes of SCoV2 [42–49]. Therefore, developing drugs or synergistic combinations involving multiple viral targets appears as a viable therapeutic strategy for the treatment of COVID-19 [2,3,50].

Within the Covid19-NMR consortium, we undertook a massive NMR-based ligand screening with the aim of identifying fragments as new chemical entities targeting SCoV2 proteins. Previously, via consorted efforts between NMR groups worldwide we have successfully developed protocols for large-scale production of more than 80% of all SCoV2 proteins [51]. Soon, the availability of proteins and the experience gained from the completion of > 20 screens with the DSI-PL fragment library for binding against the viral RNA [52] positioned us to embark on this massive screening campaign. For this purpose, > 20 SCoV2 proteins (nsp1, nsp2 (CtDR), nsp3a, nsp3b, nsp3b•GS-441524, nsp3c (SUD-N), nsp3c (SUD-MC), nsp3d, nsp3e, nsp3Y, nsp5, $_{GHM}$nsp5, $_{GS}$nsp5, nsp7, nsp8, nsp9, nsp10, nsp10•nsp14, $_{His6}$nsp15, nsp10•nsp16, ORF9a (IDR1-NTD-IDR2), ORF9a (NTD), ORF9a (NTD-SR), ORF9a (CTD), ORF9b; (for definitions see Supporting Information Table 1) were produced in NMR groups at sites all over the world and subsequently shipped to the Frankfurt NMR center (BMRZ) for conducting the NMR screening. We applied ligand-observed ^{1}H-NMR experiments and identified 311 binders across the 25 screened SCoV2 proteins. Further, we used FTMap [53], a computational mapping server which has been proven to be more accurate than the conventional GRID and MCSS methods to identify binding sites (or hot spots) on macromolecules (protein, DNA or RNA). Active sites in enzymes are usually concave surfaces that are suitable for ligand binding and therefore, in our study, binding site, hot spot, and active site are used interchangeably. FTMap predicts chemical scaffolds and functional units occupying these binding pockets. A comparison of the predicted scaffolds and functional units with the constitution of the experimental fragment hits for which we detected binding in our experimental screens showed striking correlation, as exemplified by comparing predicted and experimentally determined binding pockets for the main protease nsp5, the latter obtained both from crystallographic screens [54] as well as NMR protein-based screens conducted here. We thus propose this novel methodology for the analysis of ligand binding capability across multiple protein targets as provided in this work. Such methodology bears excellent potential to act as a unique resource for developing novel inhibitors.

Results and Discussion

We conducted fragment-based screenings for a large number of SCoV2 viral proteins (Table 1 and Supporting Information Table 1). The viral proteins can be classified broadly into three different classes, namely, (i) proteases, (ii) replicase-transcriptase (RT) complex proteins and (iii) other accessory proteins. The main protease (nsp5, Mpro, CLpro) and the Papain-like protease (nsp3d, PLpro) are two important viral proteases that play a functionally important role in viral maturation [55,56]. Nsp5 is responsible for the cleavage of 12 nsps (nsp4-nsp16) and therefore represents one of the most attractive drug targets. We screened three different constructs (nsp5, $_{GS}$nsp5 and $_{GHM}$nsp5) of nsp5. The two ($_{GS}$nsp5) or three ($_{GHM}$nsp5) additional amino acids in the N-terminus resulted from cloning. SEC-MALS analysis of these two proteins revealed that they are monomeric in solution compared to the dimeric wildtype nsp5 [51]. Recently, it has been shown that the monomer-dimer equilibrium is coupled to the catalytic activity of nsp5, with maximum activity associated with the dimeric state [57]. Therefore, identifying small molecules that interfere with the dimer formation is considered as an alternative strategy to impair catalytic activity [58] and so screening of both monomeric and dimeric states of the proteins may act as a valuable tool in identifying and developing allosteric ligands. Nsp3d is responsible for the cleavage of the N-terminus of the polyprotein, releasing nsp1, nsp2 and nsp3 and is therefore also a potential drug target. The RT-complex is composed of multiple enzymes, and we screened the SCoV2 putative primases (nsp7 and nsp8) and the methyltransferases (nsp14 and nsp16) in complex with its co-factor nsp10 (nsp10•nsp14, nsp10•nsp16). The other screened set of proteins included several nsps, various domain constructs of nsp3 and structural and accessory proteins (ORF9a (N-protein) and ORF9b). The molecular weight of the screened proteins ranged between 5 kDa (nsp2 (CtDR)) to 78 kDa (nsp10•nsp14). Further, the 25 screened proteins also included intrinsically disordered proteins (nsp2 (CtDR)), proteins with intrinsically disordered regions (N-protein), and even a protein-inhibitor complex (nsp3b•GS-441524) with the quest to identify ligands binding in close proximity to the nucleotide binding pocket as starting point for fragment growth medicinal chemistry.

The DSI-poised library (DSI-PL, Supporting Information, excel sheet 1 DSI PL Poised Library.xlsx) [59–61] has already been successfully used to screen the druggability of the RNA regulatory elements and the main protease nsp5 from SCoV2 [52,54]. This library is composed of 768 highly diverse and poised fragments specifically designed to facilitate easy downstream synthesis. We applied ligand-observed ^{1}H-NMR experiments and performed the screening with 64 mixes containing 12 fragments each as described previously [52]. In these screening experiments, changes in the ^{1}H signals of the ligand in the presence and absence of the protein served as readout for binding.

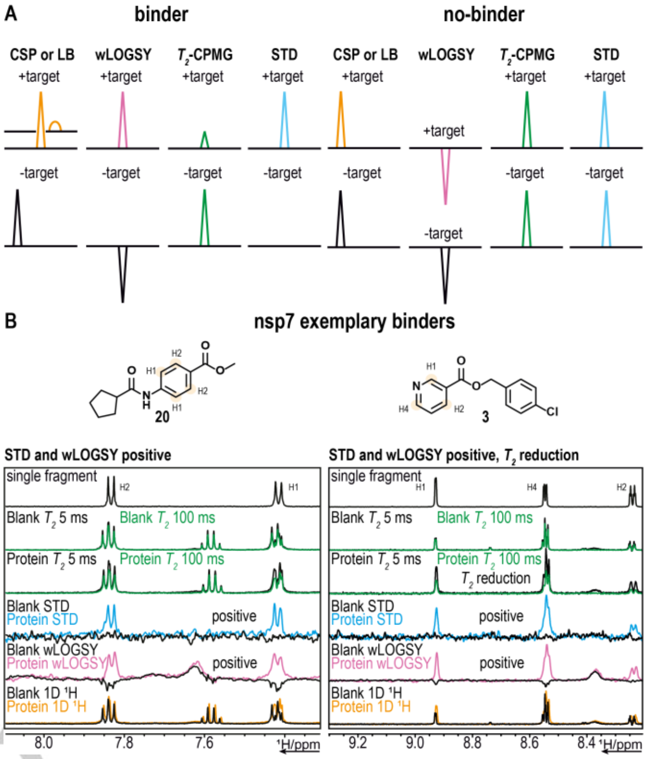


Figure 1. NMR based identification of binding fragments. (A) Schematic representation of all NMR experiments used in the screening that show exemplary effects indicating binding events in the presence of ligand compared to ligand free spectra. (B) NMR spectra (1D ^{1}H, wLOGSY, STD, and T_2-CPMG (5 ms and 100 ms) and chemical structure (binder 20 and binder 3) of two binding fragments identified for nsp7. Single fragment spectra (top) are used for chemical shift deconvolution in the mixture. Binder 20 shows clear sign changes in the STD and wLOGSY in presence of nsp7 protein. Binder 3 also shows signal in the STD and a sign change in the wLOGSY, as well as a T_2 reduction of approximately 50% in presence of nsp7 protein.

For identifying binders within the mixtures, we first compared spectra from four different NMR experiments and analyzed differences by visual inspection. As criteria, chemical shift perturbations (CSPs) or severe line broadening, sign change in the waterLOGSY (wLOGSY), STD signal or significant decrease of signal intensity in a T_2-relaxation experiment were used to identify binders (Figure 1A). Ligands were assigned as a binder if one of the four criteria was satisfied. For example, binder 20 qualifies as a binder, showing changes in wLOGSY and STD, while only minor CSP and change in T_2 (Figure 1B, left). Similarly, binder 3 qualifies as a hit, displaying changes in wLOGSY, STD and T_2, but no CSP (Figure 1B, right).

NMR-based screening resulted in 311 binders across the 25 screened SCoV2 proteins (Figure 2). Our results show that the overall binders identified against a target ranged from 2 (nsp9) to 154 (nsp3c (SUD-MC)). No correlation was observed between the molecular weight of the target and the number of binders (Supporting Information Figure 1). Strikingly, the intrinsically disordered domain of nsp2 (CtDR) shows 19 binders. By contrast, the well folded protein nsp3b has only 3 binders. The protease nsp3d and the nsp3c (SUD) as a didomain with its middle and C-terminus (MC), are amongst those with the largest number of binders (Supporting Information Table 2). The nsp3b (macro domain) is evolutionarily conserved and regarded as a potential drug target. We conducted screening in its apo/free state and in the presence of GS-441524, the active drug and metabolite of remdesivir. We observed one common binder

(binder 41) and two and four unique binders, respectively (Supporting Information Table 2). The main protease nsp5 is a dimeric cysteine protease and its N-terminus forms a part of the dimer interface. Subtle changes in the amino acid sequence at the N-terminus influence the oligomeric state ($_{\mathrm{GS}}$nsp5 and $_{\mathrm{GHM}}$nsp5, monomeric; nsp5, dimeric) of the protein [51]. For the three (nsp5, $_{\mathrm{GS}}$nsp5 and $_{\mathrm{GHM}}$nsp5) screened constructs we identified 78, 12, and 38 binders, respectively. Only 8 binders overlapped (Supporting Information Figure 2) between the three constructs, suggesting that indeed there are differential surfaces exposed for ligand binding, which in turn stems from the monomer/dimer state of the protein constructs [51]. Previously, using the DSI-PL, nsp5 and nsp14 have been screened by crystallography identifying 39 [54] and 41 [38] binders, respectively. In contrast, 78 binders were identified by NMR for the identical construct of nsp5, and for a subset of these identified binders crystallization could be reproduced in house.

A comparison of the binders revealed 6 common binders including two 3-aminopyrimidine-like compounds (21 and 26) that form the chemical starting points within the COVID moonshot initiative [40]. The twice as large number of binders identified by NMR is potentially attributed either to the presence of multiple stable conformations of nsp5 in solution [62] or to the fact that the different NMR-based screening experiments can identify binders within different affinity regimes (low micromolar to high millimolar). For nsp10•nsp14, we identified 44 binders with only one binder (binder 168) overlapping with the X-ray hits, wherein the screening was performed in the absence of nsp10. Further, 7 overlapping binders were found between nsp10•nsp14 and nsp10 NMR screens (Supporting Information Figure 2). Given the fact that significant conformational differences exist between nsp14 and nsp10•nsp14 structures [38], it is not surprising that different sets of binders are identified in X-ray and NMR screens. Further, NMR competition experiments with sinefungin, a methyltransferase inhibitor and structural analog of s-adenosyl methionine (SAM), identified that binder 141 and 146 bind to the SAM binding site.

Table 1. SCoV2 protein constructs screened by NMR

Protein *genome position (nt)* [a]	*Trivial name* Construct expressed	Size (aa) [b]	Boundaries	MW [kDa]	PDB code used for FTMap	Number of binders identified	Crossclusters in Cleft1	Crossclusters in Cleft2
nsp1 *266-805*	*Leader*	180		19.8				
	Globular Domain (GD)	116	13-127	12.7	7k7p	5	0, 7	1, 2, 3
nsp2 *806-2,719*		638		70.5				
	C-terminal IDR (CtDR)	45	557-601	4.9	-	19	-	-
nsp3 *2,720-8,554*		1,945		217.3				
a	Ub-like (UBl) domain	111	1-111	12.4	7kag	14	3, 6	0, 5, 8, 10
b	nsp3b (Macro domain)	170	207-376	18.3	6vxs	10	0, 1, 2, 3, 5	-
b	nsp3b·GS-441524	170	207-376	18.3	6vxs	5	-	-
c	SUD-N	140	409-548	15.4	2w2g	10	0, 2, 4, 5, 6, 9	-
c	SUD-MC	193	551-743	21.5	2kqv	154	1, 2, 3, 4, 6	0
d	Papain-like protease PL^{pro}	318	743-1,060	36	6w9c	150	5, 7	1, 2, 4
e	NAB	116	1,088-1,203	13.4	2k87	21	1, 4 (Cleft 3)	-
Y		286		31.5		81	-	-
nsp5 *10,055-10,972*	*Main protease (M^{pro})*	306		33.8				

Protein *genome position (nt)* [a]	*Trivial name* Construct expressed	Size (aa) [b]	Boundaries	MW [kDa]	PDB code used for FTMap	Number of binders identified	Crossclusters in Cleft1	Crossclusters in Cleft2
	$_{GS}$nsp5	306	1-306	33.8	-	12	-	-
	$_{GHM}$nsp5	306	1-306	33.8	-	38	-	-
	Full-length	306	1-306	33.8	5r83	78	3, 4, 6	1, 2, 7, 8
nsp7 *11,843-12,091*		83		9.2				
	Full-length	83	1-83	9.2	2kys	92	0, 1, 3, 6	-
nsp8 *12,092-12,685*		198		21.9				
	Full-length	198	1-198	21.9	6wiq	35	1, 3, 4, 5, 6	-
nsp9 *12,686-13,024*		113		12.4				
	Full-length	113	1-113	12.4	6w4b	2	1, 3	0, 2, 4
nsp10 *13,025-13,441*		139		14.8				
	Full-length	139	1-139	14.8	6zpe	38	0, 3, 5, 6	-
nsp15 *19,621-20,658*	*Endonuclease*	346		38.8				
	$_{His6}$nsp15	346	1-346	38.8	6w01	42	1, 2	4
nsp10•nsp16 *20,659-21,552*	*Methyltransferase*	298		33.3				
	nsp10·nsp16	298	1-298 (nsp16)	33.3	6w4h	92	3, 4, 5, 7, 8, 10	0
nsp10•nsp14 *18,040-19,620*	*Exoribonuclease*	527		61.4				
	nsp10•nsp14	527	7-527 (nsp14)	61.4	modelled	44	2, 5, 9 (Cleft 3)	-
ORF9a *28,274-29,533*	*Nucleocapsid (N)*	419		45.6				
	IDR1-NTD- IDR2	248	1-248	26.5	6yi3	7	-	-
	NTD-SR	169	44-212	18.1	6yi3	5	-	-
	NTD	136	44-180	14.9	6yi3	32	0, 1, 3, 5, 6, 7	2
	CTD	118	247-364	13.3	7c22	9	1, 2, 6, 8	-

10.1002/anie.202205858

Protein *genome position (nt)* [a]	*Trivial name* Construct expressed	Size (aa) [b]	Boundaries	MW [kDa]	PDB code used for FTMap	Number of binders identified	Crossclusters in Cleft1	Crossclusters in Cleft2
ORF9b *28,284-28,574*		97		10.8				
	Full-length	97	1-97	10.8	6z4u	8	0, 3, 5 (Cleft 3)	-

[a] Genome position in nucleotide (nt) corresponding to SCoV2 NCBI reference genome entry NC_045512.2, identical to GenBank entry MN908947.3. [b] number of amino acids excluding the additional residues due to cloning

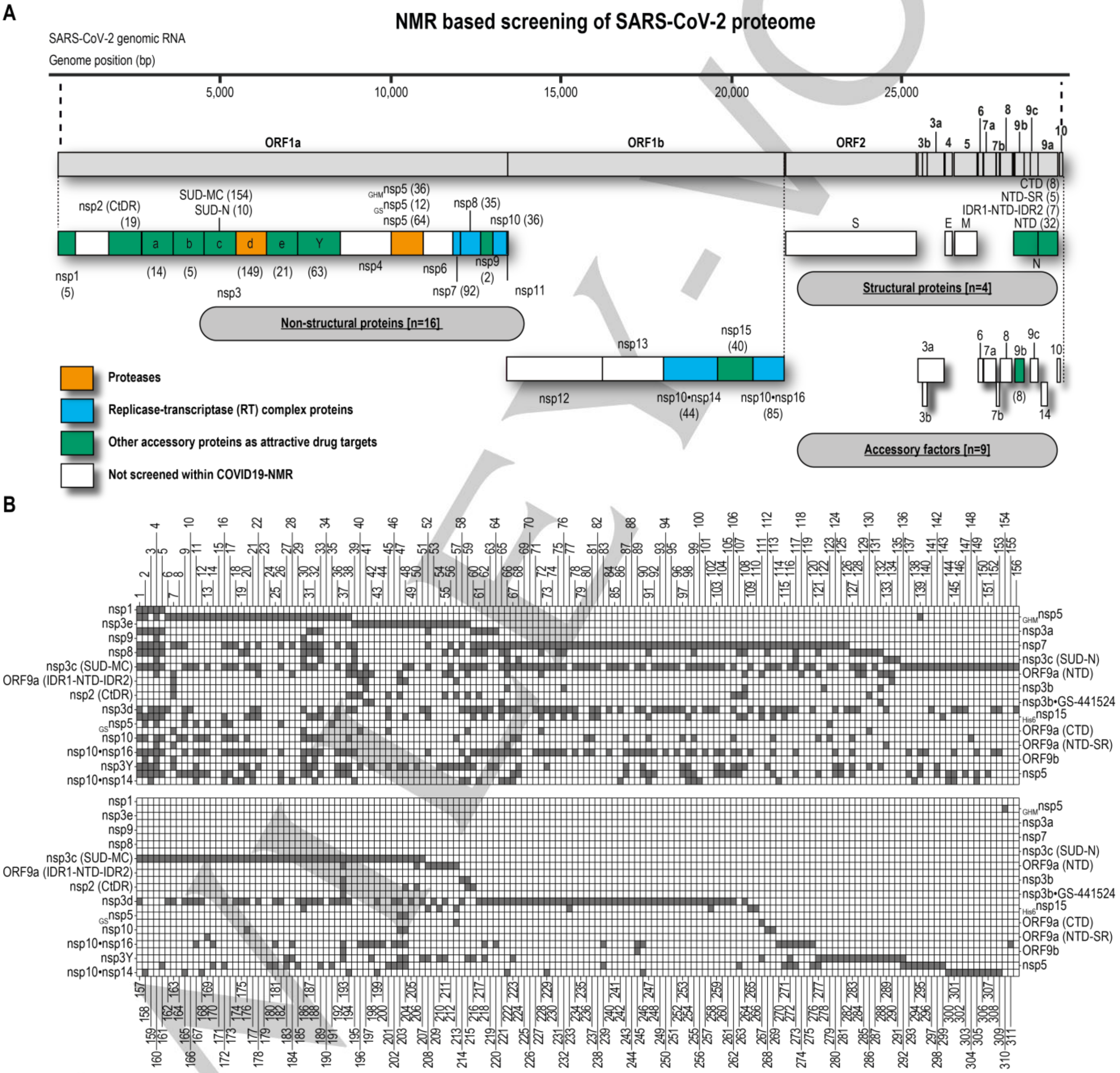


Figure 2. 311 binding fragments identified for SCoV2 proteins from NMR based fragment screening. (A) Schematic representation of the SCoV2 genome (adapted from [16]). (B) The two tables summarize all binding fragments identified in the NMR screening for their corresponding protein (grey). The first table shows binder 1 to 156 (columns) and the corresponding bound proteins (right and left rows). The second table shows binder 157 to 311 and the corresponding proteins (left and right rows).

The relatively diverse and varying number of binders across the screened SCoV2 proteins in this work is likely correlated to the accessible surface of a given protein. In general, proteins that routinely bind to either small molecules or substrates to perform their function have well-defined cavities and pockets. For example, the cysteine protease nsp5 and the nsp3b (macro domain) each have a substrate or endogenous ligand binding cleft that both are currently exploited for designing functional inhibitors. Traditionally, ligand binding pockets in proteins are determined experimentally either by X-ray crystallography or NMR. Such experimental identification of binding pockets for large sets of binders across several targets of SCoV2 reported here would be very time-consuming and sample intensive. Thus, we deduced the ligand binding sites of the SCoV2 proteins using FTMap [53]. FTMap uses 16 small organic molecules (Supporting Information Figure 3) as probes to scan the surface of the protein target and to identify regions that bind multiple of these probes, thus forming a probecluster. Several probeclusters which are in close proximity on the protein surface form one crosscluster, thus defining a consensus site or hot spot. We performed the FTMap analysis for the 18 of the 25 screened proteins for which structural coordinates were available (Supporting Information Table 3). Except for nsp3e, the pdb structures for all proteins were from SCoV2. Further, for structures with multiple chains but with the same sequence (for example: dimer) the FTMap protocol recommends each chain to be independently mapped and therefore a single monomer unit was used for all the proteins except for nsp5, ORF9a (CTD) and ORF9b that is known to exist as a stable dimer in solution and both monomeric and dimeric state were analyzed. Typically, one to three binding sites (Supporting Information Figure 4 to 21) were identified for each of the proteins. For example, the binding sites in monomeric nsp5 clustered mainly around three distinct regions of the protein, including the already known catalytic active site (Supporting Information Figure 20). However, FTMap analysis performed on the dimeric nsp5 does not identify the catalytic site (Supporting Information Figure 22 and Supporting Information Figure 23), which is in line with one of the limitations of FTMap that it works best for single domains. Therefore, monomeric form of nsp5 was utilized for the analysis of druggability. For nsp3b, hot spots clustered mainly in the ADPr binding site (Supporting Information Figure 13). Similarly, we observed the same (previously known and additional binding pockets) trend of hot spot clustering in the other proteins of SCoV2, which facilitated the definition of the relevant clefts on the protein. We used PDBsum [63] to calculate the cleft regions and ranked the clefts according to their volume. Integration of the PDBsum derived cleft information and the FTMap-identified binding sites strikingly revealed that for 13 out of 18 proteins, the hot spots identified by FTMap overlapped with cleft 1, for three proteins with cleft 2 and for three proteins with cleft 3 as identified by PDBsum (Figure 3 and Supporting Information Table 4). Importantly, FTMap analysis together with the cleft analysis for each of the SCoV2 proteins investigated here revealed that indeed, the 18 proteins contain defined potential ligand binding sites and are thus druggable. As a next step, for a given hot spot, we compared and correlated the types of FTMap probes predicted to bind in the binding sites with the chemical substructures present in the experimentally identified fragments in the DSI-PL. For this purpose, we scanned and extracted the number of occurrences of the 16 FTMap probes for all the 768 compounds from the DSI-PL using cheminformatic tools (Supporting Information excel sheet 2 DSI PL Poised Library Characterized into the 16 Probes of FTMap.xlsx). As a next step, for each of the identified binder for a given target, we quantified the overlap of probes between the hits and FTMap probes (Supporting Information excel sheets). We then selected one binder for each target, for which binding effects were observed in one or more NMR experiment. Mapping of the ligand-derived functional units revealed that for 14 out of 18 of these ligands, a 100% correlation was observed with the probes found within one or more of the crossclusters spanning the predicted cleft (Figure 3 and Supporting Information Table 4). For example, binder 21 showed positive binding effects in both wLOGSY and STD NMR experiments for nsp5 and was hence chosen as ligand of choice for this target. FTMap and cleft analysis of nsp5 suggested that crossclusters 1, 2, 7, and 8 were situated within the known active site (cleft 2) of the protease. Binder 21 is composed of mainly three (methanamine, benzene and urea) FTMap probes, and all of them are present in the crosscluster 1 (100%). The crossclusters 2, 7 and 8 each consist of one of the three probes (33%). These observations show that there is a good overlap between the chemical substructures of the FTMap ligands and those experimental fragments that occupy the hot spots, suggesting a likely binding site for this ligand. Further, in order to gain insight into the binding site of the ligand, we performed molecular docking using the Swissdock web server [64,65]. For 50% of the targets, we observed that the top-ranked pose (i.e., the ligand with the lowest binding free energy) of the ligand docks onto the binding site (Figure 3, docked ligand shown in cyan).

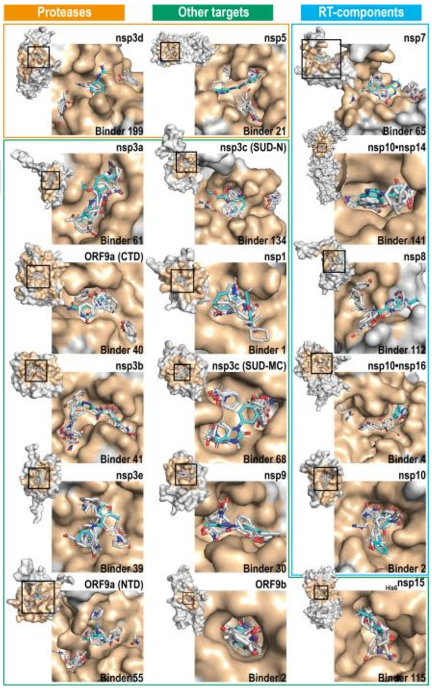


Figure 3. Hot spots identified using FTMap along with the docking of the NMR identified binders for 18 SCoV2 proteins. Proteases are highlighted with an orange box, RT-components with a blue box, and other targets with a green box. Zoom-ins show one of the identified clefts (beige colored) from PDBSum with its corresponding hot spots (and probes in grey sticks) from the FTMap analysis. For each of the targets, one of the binders was docked using SwissDock, shown in cyan.

In order to test the validity of our predicted ligand binding sites, we performed ligand-observed (ORF9a (NTD), nsp3 (SUD-MC) and nsp5) and/or protein observed (nsp5 and nsp10) titrations and determined the dissociation constants for a subset of targets by NMR. In general, the dissociation constants K_D for the fragments ranged from 50 to 2000 µM (Table 2 and Supporting Information Figure 24). Binder 13 (Z979145504) bound to nsp5 with the highest affinity. In addition, we also performed protein-observed titrations for ligands that bind to nsp5 and nsp10. An advantage of protein-observed NMR titrations is that apart from obtaining information on the dissociation constants, it is also possible to visualize the binding site of the ligand by mapping the CSPs, provided the backbone amides are assigned. Previously, within the Covid19-NMR consortium we have achieved the near-to-complete backbone assignments of nsp10 and nsp5 [36,62,66]. Binder 21 was titrated to nsp5 and bound with a K_D of ~500 µM (Figure 4, bottom right). Mapping of the CSPs revealed that apart from remote CSP effects, the residues involved in the binding mainly clustered around the active site (Figure 4, top right, blue regions), which was in good agreement with the binding cleft identified by FTMap. Moreover, FTMap and cleft analysis of nsp5 not only identified the same two sites (S1 and S3) in line with the crystal structure of binder 21 in complex with nsp5 (Figure 4, lower left, orange stick), but also reveals two additional sites (S1` and S2). A similar analysis performed for a weak binder (binder 2, K_D of ~2000 µM) of nsp10 (Supporting Information Figure 25) reveals a striking correlation between the binding site mapped based on NMR CSPs and the FTMap-detected hot spot, thus supporting the robustness and validity of our analysis. Further, FTMap analysis of the 6 and 8 overlapping binders for X-ray/NMR screening and three nsp5 constructs, respectively, suggests, that the active site (cleft 2) is their putative binding site (Supporting Information Table 5 and Supporting Information Table 6). Moreover, the 6 X-ray/NMR overlapping binders revealed identical docking poses for single chains of either monomeric (5r83) or dimeric (7khp) structures as documented in Supplementary Information Figure 26.

Table 2. Affinities of the SCoV2 protein binders.

Ligand observed				Protein Observed		
Binder	ORF9a (NTD)[a]	nsp3c (SUD-MC)[a]	nsp5[a]	Binder	nsp5[a]	nsp10[a]
40	>5	-	-	21	0.46 ±0.04	
129	>5	-	-	32	-	0.44 ± 0.05
209	>5	-	-	2	-	1.70 ± 0.54
68	-	0.45 ± 0.71	-			
30	-	>5	-			
13	-	-	0.02 ± 0.007			
26	-	-	>5			

[a] K_D in millimolar.

The NMR-based fragment hit structures were compared to > 2 million molecules contained in the ChEMBL [67], PubChem [68] and NCATS (https://opendata.ncats.nih.gov/covid19/) associated data resources of bioactive compounds. 2D Tanimoto scoring [69] was used to identify analogues annotated as active in SCoV2 bioassays. To capture "weak associations" between hits and bioactive analogues, a cut-off of 0.65 was set, which revealed 35 hit fragments associated with 50 analogues identified as active in 16 different SCoV2 assays, representing a total of 154 distinct bioactivities (Supporting Information excel sheet 3 Hits to Bioactives.xlsx). A knowledge graph additionally annotated with links to public SCoV2 assay information and relevant metadata on the bioactivities and primary targets of the 154 compounds can be accessed at https://github.com/Fraunhofer-ITMP/COVID-NMR-KG. At a more stringent Tanimoto cut-off of 0.70, a group of 9 hit fragments representing 9 analogues were identified (Table 3). Seven of the analogues, with IC_{50} values between 390 nM and 3190 nM, were identified as inhibitors of protease activity, in the study by Kuzikov et al. [70], who screened a compound repurposing collection in a FRET-based biochemical assay against full-length nsp5. Although the fragment hits binding to nsp5 also binds to at least one additional protein, three (binder 6, 37 and 67) have analogues that inhibit nsp5 activity. Two analogue compounds were also active in phenotypic assays monitoring the anti-cytopathic effect of SCoV2 in Vero E6 cell models (Metoprine, IC_{50} = 2340 nM and Oxyclozanide, IC_{50} = 3710 nM [71]. The NMR hit (binder 74) related to Metroprine, binds multiple proteins (nsp7, nsp3c (SUD-MC), nsp3d, $_{His6}$nsp15, nsp10 and nsp16) whilst the Oxyclozanide related compound (binder 79) targets a smaller group of viral proteins, namely nsp7 and nsp3c (SUD-MC).

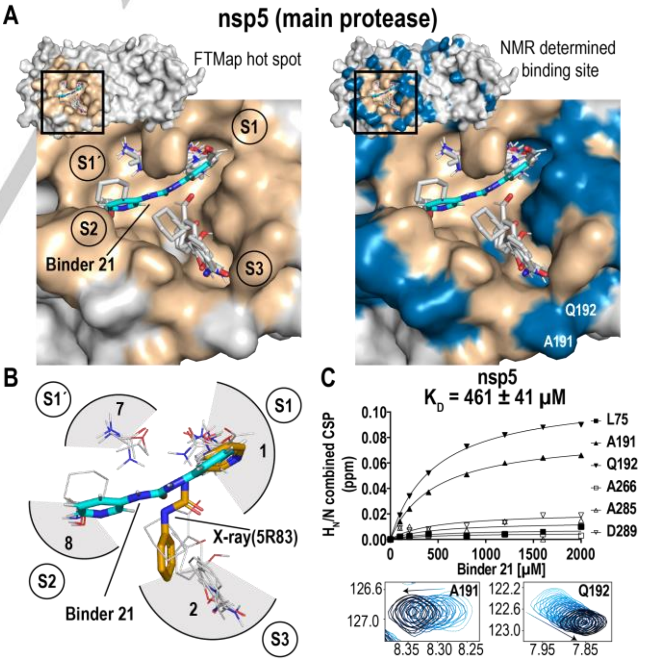
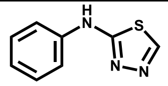
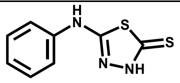
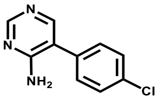
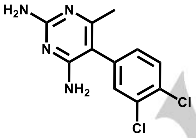
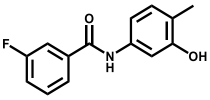
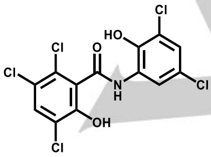
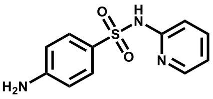
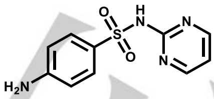
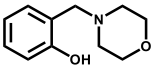
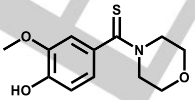
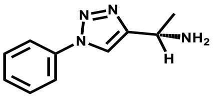
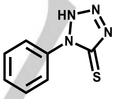
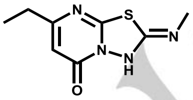
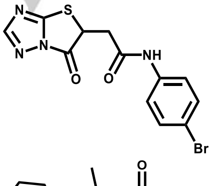
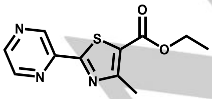
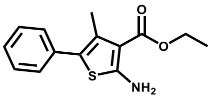
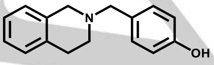
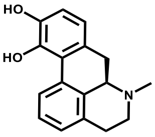


Figure 4. Agreement between bioinformatic and experimental mapping of the binding site. (A) The FTMap identified hot spot for nsp5. The subsites of the active site are labeled as S1, S1´, S2 and S3. The crossclusters (1, 2, 7, and 8) occupying the binding site are shown in grey sticks. The docked pose of binder 21 is shown in cyan. Mapping of the CSPs (in blue) on to the structure of nsp5. (B) Active site of nsp5 with an overlay of a docking (cyan) and X-ray determined (orange) structure of binder 21. (C) The interaction of binder 21 and nsp5 was monitored via NMR titration. Binder 21 binds to nsp5 with a K_D of 461 µM. The inset shows two shifting peaks (A191 and Q192) with increasing concentration of binder 21 (light blue-low to black-high).

Conclusion

Covid19 has triggered enormous research efforts. For the less than 30 viral proteins and 15 conserved RNA regulatory elements, holistic approaches screening almost all viral components can be pursued. X-ray crystallography with recently introduced automatization of fragment screening approaches [33,54] has spearheaded medicinal chemistry approaches focusing on a subset of the viral protein targets. Previously (Sreeramulu, Richter et al.) and here, we exploit the unique advances of NMR spectroscopy for screening of structured elements of the RNA genome as well as the soluble parts of the proteome. The work described thus provides information for 25*768=19200 possible protein-ligand interactions monitored by 4 different ligand-based NMR experiments. The 768 ligands come from a highly privileged fragment library. They have been assembled previously and validated by NMR for their chemical purity and solubility [59,60].

Table 3. Fragment hits from NMR-based screening and related analogues identified as biologically active compounds in SCoV2 related assays in public databases.

Binder	Structure	Binding targets detected by NMR	Bioactive analogue	CHEMBL Compound ID	Name	Tanimoto score	IC_{50} in SCoV2 related BioAssay (nM)	Assay ChEMBL ID	Assay Description
67		nsp7, nsp3c (SUD-MC), nsp3d, nsp10, nsp10•nsp16, nsp5		CHEMBL289356	CL-17107	0.88	390	CHEMBL4495583	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC_{50} FRET format with a peptide substrate
74		nsp7, nsp3c (SUD-MC), nsp3d, $_{\text{His6}}$nsp15, nsp10•nsp16		CHEMBL264373	Metoprine	0.85	2340	CHEMBL4513083	Cell based, SCoV2 induced cytotoxicity of VERO-E6 cells after 48 hours exposure to 0.01 MOI SCoV2 virus by high content imaging
79		nsp7, nsp3c (SUD-MC)		CHEMBL2105450	Oxyclozanide	0.82	3710	CHEMBL4303812	Cell based, Antiviral activity against SCoV2 (viral titer) measured by plaque assay in Vero cells at MOI 0.0125 after 24 hr
150		nsp3c (SUD-MC)		CHEMBL1382627	silver-sulfadiazine	0.81	750	CHEMBL4495583	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC_{50} FRET format with a peptide substrate
50		nsp3e, nsp3y		CHEMBL1380480	VANITIOLIDE	0.74	1320	CHEMBL4495583	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC_{50} FRET format with a peptide substrate
6		$_{\text{GHM}}$nsp5, nsp3d		CHEMBL226652	4-DAMP	0.72	2360	CHEMBL4495583	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC_{50} FRET format with a peptide substrate
157		nsp3c (SUD-MC), nsp3d		CHEMBL243652	PD096194	0.72	2040	CHEMBL4495583	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC_{50} FRET format with a peptide substrate
37		$_{\text{GHM}}$nsp5, nsp10•nsp16, nsp3y, nsp5		CHEMBL566136	PD121351	0.71	3190	CHEMBL4495583	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC_{50} FRET format with a peptide substrate
56		nsp3e, nsp7, nsp3c (SUD-MC), ORF9a (NTD), nsp3y		CHEMBL1616	APOMORPHINE HYDROCHLORIDE	0.71	520	CHEMBL4495583	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC_{50} FRET format with a peptide substrate

The screening identifies 311 hits (1.5% overall hit rate). The work goes, however, beyond reporting these screening results. We delineate a procedure to combine computational methods to validate binding site prediction from FTMap and PDBsum with the experimentally detected binding ligands. This procedure relies on the prediction of chemical sub-moieties essential for binding and the similarity of these substructures in the set of experimental binders. The thus identified and prioritized binding sites allow application of focused docking protocols and further, the experimental cross-validation by protein-based NMR experiments. From these protein-based NMR experiments, we show that dissociation constants of these fragments with proteins range from 80 µM to several millimolar. The determination of binding affinities can be used to prioritize medicinal chemistry campaigns. Using bioinformatics, identification of fragment binders also serves as starting point for database searches of known binders, using chemical similarity scores between fragments and known inhibitors as selection criterion. Thus, the herein developed workflow allows for holistic screening of the majority of the viral proteome. It provides highly valuable data for the day-to-day support of medicinal chemistry campaigns aiming at developing novel drugs applying fragment-based drug discovery. These data will also serve development of artificial intelligence (AI) based algorithms to inform hit-to-lead campaigns.

Acknowledgements

The work has been conducted in the international consortium of Covid19-NMR (covid19-nmr.de). We would like to thank Roderick Lambertz and Katharina Targaczewski for technical assistance. We thank Peter Maas and his team at SPECS for assembling the fragment library. Work at BMRZ is supported by the state of Hesse. Work in Covid19-nmr was supported by the Goethe Corona Funds, by the IWB-EFRE-program 20007375 of state of Hesse, the DFG through CRC902: "Molecular Principles of RNA-based regulation." and through infrastructure funds (project numbers: 277478796, 277479031, 392682309, 452632086, 70653611) and by European Union's Horizon 2020 research and innovation program iNEXT-discovery under grant agreement No 871037. BY-COVID receives funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement number 101046203. "INSPIRED" (MIS 5002550) project, implemented under the Action 'Reinforcement of the Research and Innovation Infrastructure,' funded by the Operational Program 'Competitiveness, Entrepreneurship and Innovation' (NSRF 2014–2020) and co-financed by Greece and the EU (European Regional Development Fund) and the FP7 REGPOT CT-2011-285950 – "SEE-DRUG" project (purchase of UPAT's 700 MHz NMR equipment). The support of the CERM/CIRMMP center of Instruct-ERIC is gratefully acknowledged. This work has been funded in part by a grant of the Italian Ministry of University and Research (FISR2020IP_02112, ID-COVID) and by Fondazione CR Firenze. A.S. is supported by the Deutsche Forschungsgemeinschaft [SFB902/B16, SCHL2062/2-1] and the Johanna Quandt Young Academy at Goethe [2019/AS01]. M.H. and C.F. thank SFB902 and the Stiftung Polytechnische Gesellschaft for the Scholarship. L.L. work was supported by the French National Research Agency (ANR, NMR-SCoV2-ORF8), the Fondation de la Recherche Médicale (FRM, NMR-SCoV2-ORF8), FINOVI and the IR-RMN-THC Fr3050 CNRS. Work at UConn Health was supported by grants from the US National Institutes of Health (R01GM135592 to B.H., P41GM111135 and R01GM123249 to J.C.H.) and the US National Science Foundation (DBI 2030601 to J.C.H.). Latvian Council of Science Grant No. VPP-COVID-2020/1-0014. National Science Foundation EAGER MCB-2031269. This work was supported by the grant Krebsliga KFS-4903-08-2019 and SNF-311030_192646 to J.O. P.G. (ITMP) The EOSC Future project is co-funded by the European Union Horizon Programme call INFRAEOSC-03-2020 - Grant Agreement Number 101017536.

Keywords: COVID19-nmr • drug discovery • fragment screening • NMR spectroscopy • SARS-CoV-2 • protein

References

[1] D. Adam, *Nature* **2022**, *601*, 312–315.

[2] G. J. Kontoghiorghes, S. Fetta, C. N. Kontoghiorghe, *Front Biosci (Landmark Ed)* **2021**, *26*, 1723–1736.

[3] Z. A. Shyr, K. Gorshkov, C. Z. Chen, W. Zheng, *Journal of Pharmacology and Experimental Therapeutics* **2020**, *375*, 127–138.

[4] T. T. Le, J. P. Cramer, R. Chen, S. Mayhew, *Evolution of the COVID-19 Vaccine Development Landscape*, Nature Reviews **2020,** *19*, 667-668.

[5] M. Mei, X. Tan, *Frontiers in Molecular Biosciences* **2021**, *8*, DOI 10.3389/fmolb.2021.671263.

[6] M. Cully, *Nature Reviews Drug Discovery* **2022**, *21*, 3–5.

[7] E. Petersen, F. Ntoumi, D. S. Hui, A. Abubakar, L. D. Kramer, *et. al*, *International Journal of Infectious Diseases* **2022**, *114*, 268–272.

[8] M. Hoffmann, N. Krüger, S. Schulz, A. Cossmann, C. Rocha, *et. al*, *Cell* **2021**, *185*, 447-456, DOI 10.1016/j.cell.2021.12.032.

[9] L. A. VanBlargan, J. M. Errico, P. J. Halfmann, S. J. Zost, J. E. Crowe, *et. al*, *Nature Medicine* **2022**, *28*, 490-495, DOI 10.1038/s41591-021-01678-y.

[10] W. F. Garcia-Beltran, K. J. st. Denis, A. Hoelzemer, E. C. Lam, A. D. Nitido, *et. al*, *Cell* **2022**, *185*, 457-466, DOI 10.1016/j.cell.2021.12.033.

[11] Y.-W. Zhou, Y. Xie, L.-S. Tang, D. Pu, Y.-J. Zhu, *et. al*, *Signal Transduction and Targeted Therapy* **2021**, *6*, 317.

[12] B. Malone, N. Urakova, E. J. Snijder, E. A. Campbell, *Nature Reviews Molecular Cell Biology* **2022**, *23*, 21–39.

[13] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, *et. al*, *Nature* **2020**, *579*, 265–269.

[14] A. R. Fehr, S. Perlman, *Methods Mol Biol* **2015**, *1282*, 1–23.

[15] E. J. Snijder, P. J. Bredenbeek, J. C. Dobbe, V. Thiel, J. Ziebuhr, *et. al*, *J Mol Biol* **2003**, *331*, 991–1004.

[16] D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, *et. al*, *Nature* **2020**, *583*, 459–468.

[17] C. W. Nelson, Z. Ardern, T. L. Goldberg, C. Meng, C. H. Kuo, *et. al*, *Elife* **2020**, *9*, e59633.

[18] A. Pavesi, *Virology* **2020**, *546*, 51–66.

[19] P. Venkatesan, *The Lancet Respiratory Medicine* **2021**, *9*, e63.

[20] B. Cao, Y. Wang, D. Wen, W. Liu, J. Wang, *et. al*, *New England Journal of Medicine* **2020**, *382*, 1787–1799.

[21] M. Wang, R. Cao, L. Zhang, X. Yang, J. Liu, *et. al*, *Cell Research* **2020**, *30*, 269–271.

[22] W.-C. Ko, J.-M. Rolain, N.-Y. Lee, P.-L. Chen, C.-T. Huang, *et. al*, *International Journal of Antimicrobial Agents* **2020**, *55*, 105933.

[23] M. A. Martinez, *Frontiers in Immunology* **2021**, *12*, DOI 10.3389/fimmu.2021.635371.

[24] J. O. Ogidigo, E. A. Iwuchukwu, C. U. Ibeji, O. Okpalefe, M. E. S. Soliman, *Journal of Biomolecular Structure and Dynamics* **2020**, *40*, 2284–2301.

[25] C. Liu, X. Zhu, Y. Lu, X. Zhang, X. Jia, T. Yang, *Journal of Pharmaceutical Analysis* **2021**, *11*, 272–277.

[26] M. A. White, W. Lin, X. Cheng, *The Journal of Physical Chemistry Letters* **2020**, *11*, 9144–9151.

[27] M. T. J. Quimque, K. I. R. Notarte, R. A. T. Fernandez, M. A. O. Mendoza, R. A. D. Liman, *et. al*, *Journal of Biomolecular Structure and Dynamics* **2021**, *39*, 4316–4333.

[28] A. Carino, F. Moraca, B. Fiorillo, S. Marchianò, V. Sepe, *et. al*, *Frontiers in Chemistry* **2020**, *8*, DOI 10.3389/fchem.2020.572885.

[29] S. Barage, A. Karthic, R. Bavi, N. Desai, R. Kumar, *et. al*, *Journal of Biomolecular Structure and Dynamics* **2020**, *40*, 2557-2574.

[30] M. Macchiagodena, M. Pagliai, P. Procacci, *Chemical Physics Letters* **2020**, *750*, 137489.

[31] S. Yazdani, N. de Maio, Y. Ding, V. Shahani, N. Goldman, M. Schapira, *Journal of Proteome Research* **2021**, *20*, 4212–4215.

[32] L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, *et. al*, *Science (1979)* **2020**, *368*, 409–412.

[33] M. Schuller, G. J. Correy, S. Gahbauer, D. Fearon, T. Wu, *et. al*, *Science Advances* **2021**, *7*, eabf8711, DOI 10.1126/sciadv.abf8711.

[34] S. Günther, P. Y. A. Reinke, Y. Fernández-García, J. Lieske, T. J. Lane, *et. al*, *Science (1979)* **2021**, *372*, 642–646.

[35] J. A. Newman, A. Douangamath, S. Yadzani, Y. Yosaatmadja, A. Aimon, *et. al*, *Nature Communications* **2021**, *12*, 4848.

[36] F. Cantrelle, E. Boll, L. Brier, D. Moschidi, S. Belouzard, *et. al*, *Angewandte Chemie International Edition* **2021**, *60*, 25428–25435.

[37] A. L. Kantsadi, E. Cattermole, M.-T. Matsoukas, G. A. Spyroulias, I. Vakonakis, *Journal of Biomolecular NMR* **2021**, *75*, 167–178.

[38] N. Imprachim, Y. Yosaatmadja, J. A. Newman, *bioRxiv* **2022**, 2022.03.11.483836.

[39] V. Napolitano, A. Dabrowska, K. Schorpp, A. Mourão, E. Barreto-Duran, *et. al*, *Cell Chemical Biology* **2022**, *29*, 774-784, DOI 10.1016/j.chembiol.2021.11.006.

[40] T. C. M. Consortium, H. Achdout, A. Aimon, E. Bar-David, H. Barr, *et. al*, *bioRxiv* **2022**, 2020.10.29.339317.

[41] M. Kozlov, *Nature* **2022**, *601*, 496, DOI 10.1038/d41586-022-00112-8.

[42] J. F. X. Diffley, *Biochemical Journal* **2021**, *478*, 2533–2535.

[43] C. T. Lim, K. W. Tan, M. Wu, R. Ulferts, L. A. Armstrong, *et. al*, *Biochemical Journal* **2021**, *478*, 2517–2531.

[44] J. C. Milligan, T. U. Zeisner, G. Papageorgiou, D. Joshi, C. Soudy, *et. al*, *Biochemical Journal* **2021**, *478*, 2499–2515.

[45] S. Basu, T. Mak, R. Ulferts, M. Wu, T. Deegan, *et. al*, *Biochemical Journal* **2021**, *478*, 2481–2497.

[46] B. Canal, R. Fujisawa, A. W. McClure, T. D. Deegan, M. Wu, *et. al*, *Biochemical Journal* **2021**, *478*, 2465–2479.

[47] B. Canal, A. W. McClure, J. F. Curran, M. Wu, R. Ulferts, *et. al*, *Biochemical Journal* **2021**, *478*, 2445–2464.

[48] A. P. Bertolin, F. Weissmann, J. Zeng, V. Posse, J. C. Milligan, *et. al*, *Biochemical Journal* **2021**, *478*, 2425–2443.

[49] J. Zeng, F. Weissmann, A. P. Bertolin, V. Posse, B. Canal, *et. al*, *Biochemical Journal* **2021**, *478*, 2405–2423.

[50] P. Ren, W. Shang, W. Yin, H. Ge, L. Wang, *et. al*, *Acta Pharmacologica Sinica* **2022**, *43*, 483–493.

[51] N. Altincekic, S. M. Korn, N. S. Qureshi, M. Dujardin, M. Ninot-Pedrosa, *et. al*, *Frontiers in Molecular Biosciences* **2021**, *8*, 89.

[52] S. Sreeramulu, C. Richter, H. Berg, M. A. Wirtz Martin, B. Ceylan, *et. al*, *Angewandte Chemie - International Edition* **2021**, *60*, 19191-19200, DOI 10.1002/anie.202103693.

[53] D. Kozakov, L. E. Grove, D. R. Hall, T. Bohnuud, S. E. Mottarella, *et. al*, *Nature Protocols* **2015**, *10*, 733–755.

[54] A. Douangamath, D. Fearon, P. Gehrtz, T. Krojer, P. Lukacik, *et. al*, *Nature Communications* **2020**, *11*, DOI 10.1038/s41467-020-18709-w.

[55] D. Shin, R. Mukherjee, D. Grewe, D. Bojkova, K. Baek, *et. al*, *Nature* **2020**, *587*, 657–662.

[56] K. Anand, J. Ziebuhr, P. Wadhwani, J. R. Mesters, R. Hilgenfeld, *Science (1979)* **2003**, *300*, 1763–1767.

[57] N. T. Nashed, A. Aniana, R. Ghirlando, S. C. Chiliveri, J. M. Louis, *Communications Biology* **2022**, *5*, 160.

[58] B. Goyal, D. Goyal, *ACS Combinatorial Science* **2020**, *22*, 297–305.

[59] O. B. Cox, T. Krojer, P. Collins, O. Monteiro, R. Talon, A. Bradley, O. Fedorov, J. Amin, B. D. Marsden, J. Spencer, F. von Delft, P. E. Brennan, *Chemical Science* **2016**, *7*, 2322-2330, DOI 10.1039/c5sc03115j.

[60] S. Sreeramulu, C. Richter, T. Kuehn, K. Azzaoui, M. J. J. Blommers, *et. al*, *Journal of Biomolecular NMR* **2020**, *74*, 555-563, DOI 10.1007/s10858-020-00327-9.

[61] H. Berg, M. A. Wirtz Martin, A. Niesteruk, C. Richter, S. Sreeramulu, H. Schwalbe, *Journal of Visualized Experiments* **2021**, *172*, e62262, DOI 10.3791/62262.

[62] A. J. Robertson, J. Ying, A. Bax, *Magnetic Resonance* **2021**, *2*, 129–138.

[63] R. A. Laskowski, J. Jabłońska, L. Pravda, R. S. Vařeková, J. M. Thornton, *Protein Science* **2018**, *27*, 129–134.

[64] A. Grosdidier, V. Zoete, O. Michielin, *Journal of Computational Chemistry* **2011**, *32*, 2149–2159.

[65] A. Grosdidier, V. Zoete, O. Michielin, *Nucleic Acids Research* **2011**, *39*, W270–W277.

[66] N. Kubatova, N. S. Qureshi, N. Altincekic, R. Abele, J. K. Bains, *et. al*, *Biomolecular NMR Assignments* **2021**, *15*, 65–71.

[67] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, *et. al*, *Nucleic Acids Research* **2019**, *47*, D930–D940.

[68] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, *et. al*, *Nucleic Acids Research* **2021**, *49*, D1388–D1395.

[69] X. Chen, C. H. Reynolds, *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 1407–1414.

[70] M. Kuzikov, E. Costanzi, J. Reinshagen, F. Esposito, L. Vangeel, *et. al*, *ACS Pharmacology & Translational Science* **2021**, *4*, 1096–1110.

[71] S. Jeon, M. Ko, J. Lee, I. Choi, S. Y. Byun, S. Park, D. Shum, S. Kim, *Antimicrobial Agents and Chemotherapy* **2020**, *64*, DOI 10.1128/AAC.00819-20.

[72] nsp1, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp1, BMRbig, bmrbig48, https://bmrbig.bmrb.io/released/bmrbig48.

[73] _GHM_nsp5, H. Berg, 2022, DSI-PL_COVID19-NMR_GHM_nsp5, BMRbig, bmrbig45, https://bmrbig.bmrb.io/released/bmrbig45.

[74] nsp3e, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp3e, BMRbig, bmrbig56, https://bmrbig.bmrb.io/released/bmrbig56.

[75] nsp3a, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp3a, BMRbig, bmrbig50, https://bmrbig.bmrb.io/released/bmrbig50.

[76] nsp9, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp9, BMRbig, bmrbig61, https://bmrbig.bmrb.io/released/bmrbig61.

[77] nsp7, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp7, BMRbig, bmrbig59, https://bmrbig.bmrb.io/released/bmrbig59.

[78] nsp8, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp8, BMRbig, bmrbig60, https://bmrbig.bmrb.io/released/bmrbig60.

[79] nsp3c (SUD-N), H. Berg, 2022, DSI-PL_COVID19-NMR_nsp3c_SUD-N, BMRbig, bmrbig54, https://bmrbig.bmrb.io/released/bmrbig54.

[80] nsp3c (SUD-MC), H. Berg, 2022, DSI-PL_COVID19-NMR_nsp3c_SUD-MC, BMRbig, bmrbig53, https://bmrbig.bmrb.io/released/bmrbig53.

[81] ORF9a (NTD), H. Berg, 2022, DSI-PL_COVID19-NMR_ORF9a_NTD, BMRbig, bmrbig67, https://bmrbig.bmrb.io/released/bmrbig67.

[82] ORF9a (IDR1-NTD-IDR2), H. Berg, 2022, DSI-PL_COVID19-NMR_ORF9a_IDR1-NTD-IDR2, BMRbig, bmrbig66, https://bmrbig.bmrb.io/released/bmrbig66.

[83] nsp3b, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp3b, BMRbig, bmrbig51, https://bmrbig.bmrb.io/released/bmrbig51.

[84] nsp2 (CtDR), H. Berg, 2022, DSI-PL_COVID19-NMR_nsp2_CtDR, BMRbig, bmrbig49, https://bmrbig.bmrb.io/released/bmrbig49.

[85] nsp3b·GS-441524, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp3b_GS-441524, BMRbig, bmrbig52, https://bmrbig.bmrb.io/released/bmrbig52.

[86] nsp3d, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp3d, BMRbig, bmrbig55, https://bmrbig.bmrb.io/released/bmrbig55.

[87] _His6_nsp15, H. Berg, 2022, DSI-PL_COVID19-NMR_His6_nsp15, BMRbig, bmrbig46, https://bmrbig.bmrb.io/released/bmrbig46.

[88] _GS_nsp5, H. Berg, 2022, DSI-PL_COVID19-NMR_GS_nsp5, BMRbig, bmrbig47, https://bmrbig.bmrb.io/released/bmrbig47.

[89] ORF9a (CTD), H. Berg, 2022, DSI-PL_COVID19-NMR_ORF9a_CTD, BMRbig, bmrbig65, https://bmrbig.bmrb.io/released/bmrbig65.

[90] nsp10, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp10, BMRbig, bmrbig62, https://bmrbig.bmrb.io/released/bmrbig62.

[91] ORF9a (NTD-SR), H. Berg, 2022, DSI-PL_COVID19-NMR_ORF9a_NTD-SR, BMRbig, bmrbig68, https://bmrbig.bmrb.io/released/bmrbig68.

[92] nsp10·nsp16, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp10_nsp16, BMRbig, bmrbig63, https://bmrbig.bmrb.io/released/bmrbig63.

[93] ORF9b, H. Berg, 2022, DSI-PL_COVID19-NMR_ORF9b, BMRbig, bmrbig69, https://bmrbig.bmrb.io/released/bmrbig69.

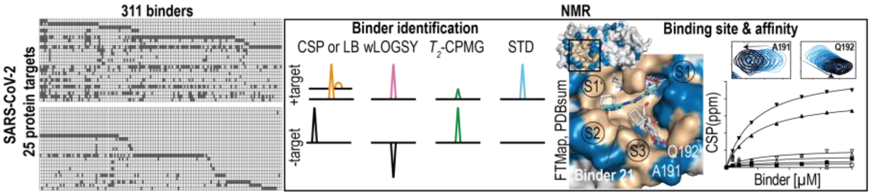
[94] nsp3y, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp3y, BMRbig, bmrbig57, https://bmrbig.bmrb.io/released/bmrbig57.

[95] nsp5, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp5, BMRbig, bmrbig58, https://bmrbig.bmrb.io/released/bmrbig58.

[96] nsp10·nsp14, H. Berg, 2022, DSI-PL_COVID19-NMR_nsp10_nsp14, BMRbig, bmrbig64, https://bmrbig.bmrb.io/released/bmrbig64.

10.1002/anie.202205858

Entry for the Table of Contents



Using a fragment-based screening strategy by NMR, we identified 311 small molecule binders of 25 SARS-CoV-2 proteins, thus expanding the previously unexplored chemical and target space. Further, using experimental and bioinformatic analysis we identify potential binding sites. This comprehensive data would greatly assist medicinal chemistry efforts even beyond COVID-19.

Institute and/or researcher Twitter usernames: ((optional))

5.16 **Research article:** ^{19}F NMR-Based Fragment Screening for 14 Different Biologically Active RNAs and 10 DNA and Protein Counter-Screens

Oliver Binas*, Vanessa de Jesus*, Tom Landgraf*, Albrecht E. Völklein*, Jason Martins, Daniel Hymon, Jasleen Kaur Bains, Hannes Berg, Thomas Biedenbänder, Boris Fürtig, Santosh L. Gande, Anna Niesteruk, Andreas Oxenfarth, Nusrat S. Qureshi, Tatjana Schamber, Robbin Schnieders, Alix Tröster, Anna Wacker, Julia Wirmer-Bartoschek, Maria A. Wirtz Martin, Elke Stirnal, Kamal Azzaoui, Christian Richter, Sridhar Sreeramulu, Marcel Jules José Blommers and Harald Schwalbe; *ChemBioChem* **22**, 423-433 (2021)

* These authors contributed equally to this work.

In this article, the druggability of 24 biological targets, including 14 RNAs, five DNAs, and five proteins, was investigated using a fragment-based screening based on ^{19}F-NMR spectroscopy. A fragment mixture containing 102 ligands was used. Several types of RNA were screened, including small stem-loop structures, ribozymes, tRNAs, aptamers, riboswitch terminators and anti-terminators as well as full-length riboswitches. DNA targets included G-quadruplex structures of various moieties and double-stranded DNAs. Protein screening included enzymes such as kinases and phosphatases as well as RNA-binding proteins.

O. Binas, V. de Jesus, T. Landgraf and A. E. Völklein performed further experiments, analyzed the data, and wrote the manuscript. The author of this thesis contributed to the experiments with the expression and purification of the T7 RNA polymerase.

https://doi.org/10.1002/cbic.202000476

^{19}F NMR-Based Fragment Screening for 14 Different Biologically Active RNAs and 10 DNA and Protein Counter-Screens

Oliver Binas+,[a] Vanessa de Jesus+,[a] Tom Landgraf+,[a] Albrecht Eduard Völklein+,[a] Jason Martins,[a] Daniel Hymon,[a] Jasleen Kaur Bains,[a] Hannes Berg,[a] Thomas Biedenbänder,[a] Boris Fürtig,[a] Santosh Lakshmi Gande,[a] Anna Niesteruk,[a] Andreas Oxenfarth,[a] Nusrat Shahin Qureshi,[a] Tatjana Schamber,[a] Robbin Schnieders,[a] Alix Tröster,[a] Anna Wacker,[a] Julia Wirmer-Bartoschek,[a] Maria Alexandra Wirtz Martin,[a] Elke Stirnal,[a] Kamal Azzaoui,[b] Christian Richter,*[a] Sridhar Sreeramulu,*[a] Marcel Jules José Blommers,*[b] and Harald Schwalbe*[a]

We report here the nuclear magnetic resonance ^{19}F screening of 14 RNA targets with different secondary and tertiary structure to systematically assess the druggability of RNAs. Our RNA targets include representative bacterial riboswitches that naturally bind with nanomolar affinity and high specificity to cellular metabolites of low molecular weight. Based on counter-screens against five DNAs and five proteins, we can show that RNA can be specifically targeted. To demonstrate the quality of the initial fragment library that has been designed for easy follow-up chemistry, we further show how to increase binding affinity from an initial fragment hit by chemistry that links the identified fragment to the intercalator acridine. Thus, we achieve low-micromolar binding affinity without losing binding specificity between two different terminator structures.

Introduction

Proteins constitute the vast majority of validated drug targets. The ribosome, a large RNA-protein complex, is the most prominent RNA drug target. Most antibiotics inhibit protein synthesis by targeting the interface of RNA and proteins in the ribosome.[1] Beyond being target for antibiotics, RNA has for long been considered undruggable. Recently, however, this view has changed and RNA emerged as a potential target for drug discovery as well.[2–6] Clinical success of compounds initially identified as RNA binders[7] inspires thorough exploration of the noncoding RNA target space. Here, potential targets range from RNA involved in oncology and inflammation to RNA involved in bacterial and viral infections, to mention a few areas with unmet medical need. Concurrently, the continuous identification of new regulatory RNAs, including riboswitches or sRNAs further increases potential applications.[8] To combat multidrug-resistant (MDR) bacteria that pose a major health threat for modern human society,[9] Riboswitches in particular have come into focus, which can only be tackled by new antibiotics. The development of drugs targeting riboswitches is therefore an important research focus.[10]

Structure-based drug discovery is a key methodology for rational drug discovery. Here, X-ray crystallography[11] and nuclear magnetic resonance (NMR) spectroscopy[12] provide the essential structural information. Insight from target structures is often supported by computational methods to aid in library design. Virtual screening by *in silico* docking of a compound library against available target model structures guides medicinal chemistry in the development from initial hits towards the generation of lead compounds.[13]

High-throughput screening in drug discovery requires robust detection of binding of a large number of test compounds, a library, to a biological target. Such screening can involve up to 1–2 million compounds from which routinely a very small number of potential lead compounds are identified. Screening large libraries thus requires high preparative and infrastructural effort.

An alternative to this classic approach is fragment-based screening. Fragments are often weak binders and their binding

[a] *O. Binas,+ V. de Jesus,+ T. Landgraf,+ A. E. Völklein,+ J. Martins, D. Hymon, J. Kaur Bains, H. Berg, T. Biedenbänder, Dr. B. Fürtig, Dr. S. Lakshmi Gande, Dr. A. Niesteruk, A. Oxenfarth, Dr. N. Shahin Qureshi, T. Schamber, R. Schnieders, A. Tröster, Dr. A. Wacker, Dr. J. Wirmer-Bartoschek, M. A. Wirtz Martin, E. Stirnal, Dr. C. Richter, Dr. S. Sreeramulu, Prof. Dr. H. Schwalbe*
Institute for Organic Chemistry and Chemical Biology
Center for Biomolecular Magnetic Resonance (BMRZ)
Johann Wolfgang Goethe-University Frankfurt
Max-von-Laue Strasse 7, 60438 Frankfurt am Main (Germany)
E-mail: ric@nmr.uni-frankfurt.de
sridhar@nmr.uni-frankfurt.de
schwalbe@nmr.uni-frankfurt.de

[b] *Dr. K. Azzaoui, Dr. M. J. José Blommers*
Saverna Therapeutics
Gewerbestrasse 24, 4123 Allschwil (Switzerland)
E-mail: mjj.blommers@saverna.com

[+] *These authors contributed equally to this work.*

Supporting information for this article is available on the WWW under https://doi.org/10.1002/cbic.202000476

specificity can be lower than expected from a lead compound.[14] Thus, fragment-based drug discovery requires that the initial hits are further processed into lead compounds by chemical modification such as growing the compound to fit the desired number of binding interactions with the target or linkage of two fragments which bind to binding sites in spatial proximity. Fragment-based drug discovery is nowadays the basis of many hit-to-lead research programs.[15] In fact, it has been shown that FDA-approved drugs targeting proteins can be developed starting from fragment screens.[15–21]

Methods used to screen RNA include fluorescence-based assays,[22,23] mass spectrometry,[24,25] small-molecule microarrays (SMM),[26,27] microscale thermophoresis (MST)[28,29] and NMR spectroscopy.[30] However, most of these studies focus only on a single target RNA.[31,32] In contrast, we report herein the screening of a library of 102 fragments against 14 different RNAs of different sizes and different architectures by using ^{19}F NMR as the main method for hit identification. The targeted RNAs include small stem loop structures, aptamer domains of riboswitches, full-length riboswitches, terminators and antiterminators of riboswitches, ribozymes as well as tRNAs, traditionally serving as control RNAs in screening (Figure 1). Thus, our results allow us to delineate specificities of fragments towards different RNAs. We further counter-screened against five DNAs and five proteins to test whether the fragment library can target for these different classes of biomacromolecules or whether the library is biased towards binding a subset of biomacromolecules. The DNA targets include regular double-stranded DNA as well as G-quadruplex structures of different morphology, and the proteins include RNA-binding proteins as well as the important enzyme classes kinases and phosphatases that bind to phosphorylated moieties as part of the enzymatic function. With our multi-target approach, we show that selective fragments for an RNA target can be found but often a good hit is broadly binding to RNAs of the same size and structural complexity. Also we show that the selective targeting of RNA over other classes of biomacromolecules is possible with this library.

By way of example, we further show that by fast follow-up chemistry, that involves the linkage of an RNA-binding fragment with the intercalator acridine, we obtain low micromolar RNA binders with more than tenfold specificity towards different RNAs.

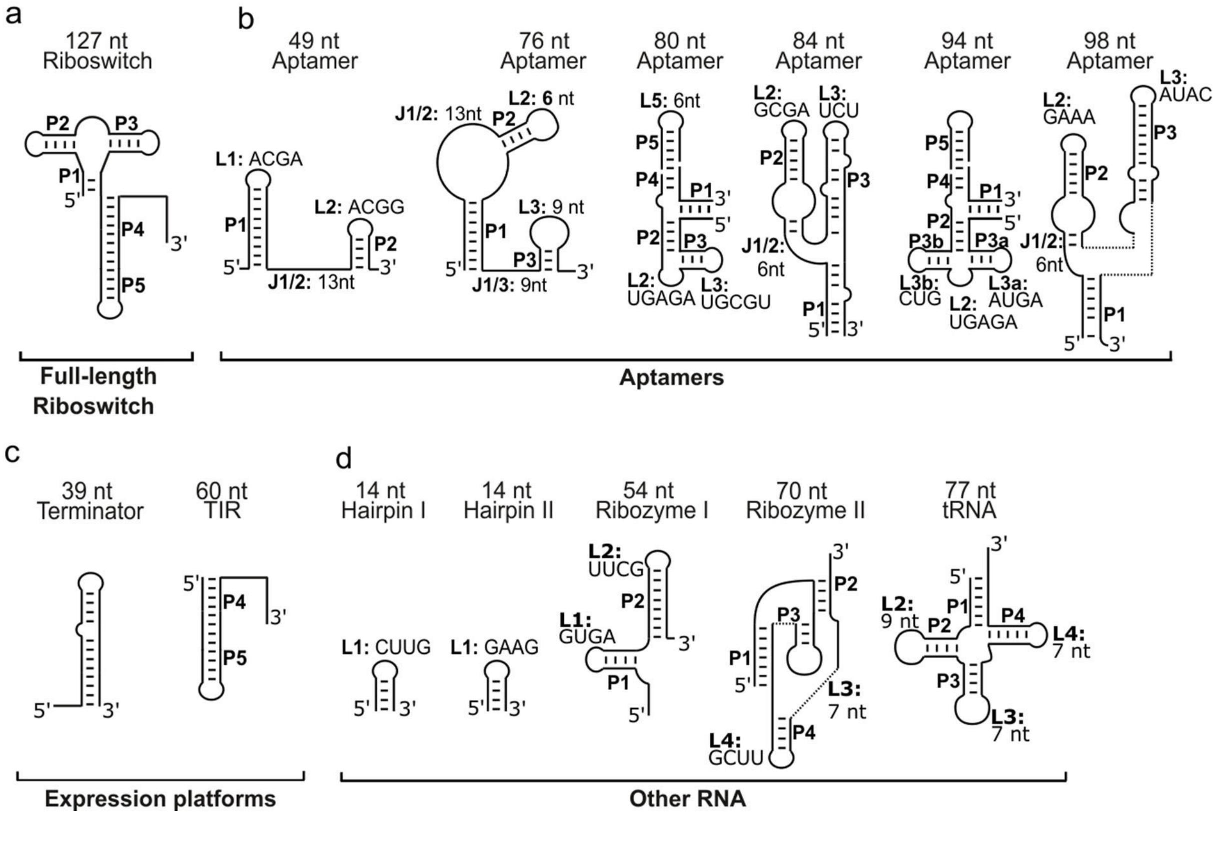


Figure 1. Overview of RNA targets. Schematic secondary structures of the RNA targets investigated by ^{19}F fragment binding studies (FBS). Stems (P), loops (L) and junctions (J) are annotated. Tri-, tetra- and pentaloop sequences are listed explicitly. Rational ligand design led to the development of compounds that specifically bind to an RNA loop region.[37] In our study, we included two 14-nt stem-loop structures exhibiting a GAAG and a CUUG tetraloop, respectively in order to detect fragments binding to this abundant secondary structure motif. Also, we included the guanidine-sensing riboswitch as an example of a functional RNA with hairpin structures. Loop–loop interactions[38] are part of the stabilizing function with purine-sensing riboswitches that are part of the RNA targets in this study (Figure S3).

Results and Discussion

Target choice is an important step for any screening, especially in a multi-target approach in which a broad spectrum of biologically relevant target molecules is crucial to success. Therefore, we showcase our measures of target choice in the following section. All RNA constructs screened are summarized in Table 1.

RNA hairpin structures

Stem-loop/hairpin structures represent the most common small secondary structure motifs in RNA.[33] Common loop lengths range from three to seven nucleotides, but more than 50% of all loops are tetraloops.[34] Tetraloops are not only very abundant, they also exhibit a high thermodynamic stability as they are usually stabilized by hydrogen bonding and stacking interactions. When drug-screening approaches are employed on biologically relevant stem-loops, further characterization of the binding mode is needed to distinguish stem-binding[35] from loop-binding ligands,[27] to understand the structural basis of the ligand-induced change in biological function.[36]

RNA bulges, internal loops and pseudoknots

Helix-junction-helix (HJH) structure elements occur between two helices or three- and four-way junctions. They can be divided into bulges and internal loops, where the first is characterized by short single-stranded intersections on one side of an RNA stem and the latter features unpaired regions on

Table 1. List of all biomolecules used in the study listed with their biological host organism (if applicable), PDB accession codes of X-ray structures and primary publication. *only homologue structures available. #only aptamer structures or single domains available.

	Organism	X-ray	NMR
Riboswitches and Aptamers			
Guanidine (Gdn-II)-sensing riboswitch (49 nt)	*Escherichia coli*	5NDI[73]	
ZMP-sensing riboswitch (76 nt)	*Thermosinus carboxydivorans*	4ZNP[74]	
thiM TPP-sensing riboswitch (80 nt)	*E. coli*	2GDI[75]	
pilM 3′, 3′-cGAMP-sensing riboswitch (84 nt)	*Geobacter metallireducens*	4YAZ*[76]	
TenA TTP- sensing riboswitch (94 nt)	*Staphylococcus aureus*		
cyclic di-GMP-1 riboswitch (98 nt)	*Clostridium difficile*	3MXH*[77] 3IRW*[78] 3IWN*[79]	
Adenine-sensing riboswitch (127 nt)	*Vibrio vulnificus*	1y26#[80] 5E54#[81] 4TZX#[82]	[49] [69]
Riboswitch Elements			
2′-deoxyguanosine-sensing-riboswitch terminator (39 nt)	*Mesoplasma florum*		[83]
SAM-sensing riboswitch anti-terminator (38 nt)	*Bacillus subtilis*		
Adenine-sensing riboswitch terminator (51 nt)	*B. subtilis*		[84]
Adenine-sensing riboswitch expression platform (60 nt)	*V. vulnificus*		[69] [85]
Other RNAs			
RNA with GAAG tetraloop (14 nt)	*artificial*		2F87*[86]
RNA CUUG tetraloop (14 nt)	*artificial*		1RNG*[87]
Hammerhead ribozyme (54 nt)		1MME[88]	
Hepatitis delta virus ribozyme (70 nt)	*Hepatitis delta virus*	1DRZ[89]	
tRNAfMet (77 nt)	*E. coli*	[90]	
DNA			
cMyc G-quadruplex (22 nt)	*Homo sapiens*		1XAV[91] 2L7V[92]
cKit G-quadruplex (24 nt)	*H. sapiens*	3QXR*[93] 2WO2	2O3M[94]
DNA duplex (24 nt)	*artificial*	1BNA[95]	
Tel26 G-quadruplex (26 nt)	*H. sapiens*		2HY9[96] 5Z80[97]
wtTel26 G-quadruplex (26 nt)	*H. sapiens*		5MVB[98] 2JPZ[99]
Proteins			
Mycobacterium tuberculosis protein tyrosine phosphatase A (MptpA, 18 kDa)	*Mycobacterium tuberculosis*	1U2P[100]	2LUO[101]
Protein tyrosine kinase A (PtkA, 30 kDa)	*M. tuberculosis*		6F2X[102]
Receptor tyrosine kinase EphA2 (34 kDa)	*H. sapiens*	5I9U[103]	
Ribosomal protein S1 (61 kDa)	*V. vulnificus*		2MFI*#[104] 2KHI*#[105]
T7 RNA polymerase (100 kDa)	*Escherichia phage T7*	1MSW[106]	

both sides of the stem.[39,40] Bulges and internal loops constitute conformational hinges, allowing helices to adopt different conformations with respect to each other. Formation of these structures allows for inter helix motions such as dynamic nucleobase stacking or rotation in and out of a junction.[39] They can often be targeted by low-molecular-weight ligands as previously shown for the Tat-TAR interaction, which was mimicked by arginamide.[41,42] In drug screening approaches, internal loops and bulges can be valuable targets for ligand design since their potential for ligand recognition is relatively high. Examples of virtual screenings directed on HIV-1 trans-activation response element (TAR) show that sampling of the entirety of the allowed topological space leads to an ensemble of discrete conformations that can bind different ligands.[36,43,44]

In our screening pool, several examples of small and large bulge regions, internal loops and pseudoknots are present in RNAs (Figure 1).

Riboswitches

Riboswitches are structured RNA elements which regulate gene expression by allosteric structural re-arrangements of an expression platform element in response to sensing environmental changes by an aptamer element. Most riboswitches respond to changes in concentration of small molecules, mostly metabolites, which they bind with remarkable specificity.[45,46] Examples showing the observation of binding via homonuclear and heteronuclear 2D NMR spectroscopy are displayed in Figures S2 and S3 in the Supporting Information. Intricate tertiary structures are formed by most aptamers to achieve high-affinity binding required for optimal sensitivity, alongside with sufficient discrimination against noncognate ligands. These binding pockets feature a closely defined chemical space, which is usually thoroughly described by structural data (Figure S1). The complex and specific chemical environments aid development of specific ligands and therefore especially fragment-based drug discovery approaches, which rely on chemical adaption in particular. Thus, it is not surprising that riboswitch aptamer domains have been identified as excellent drug targets very early on.[47,48] Of the 14 RNAs screened, eight are derived from riboswitches including natural ligand binding aptamer domains (*vide infra*).

In this study, we screened the aptamer domains of riboswitches from the second-messenger-sensing class, the guanidinium-sensing class, the purine-sensing class and the thiamin-pyrophosphate-(TPP)-sensing riboswitch. The TPP-sensing riboswitch represents the most abundant riboswitch found in different prokaryotes and even eukaryotes (Table 1). For purine-sensing riboswitches, operating either on the transcriptional or the translational level, we have previously reported on the mechanism of full-length riboswitch function.[49,50]

Counter screens

To maximize the coverage of conformational space and to rule out unspecific binding, we added five other RNAs ranging from 14 to 77 nt in length to the pool of target RNAs. We screened tRNA$^{\mathrm{fMet}}$ produced in house. tRNAs, being ubiquitously present in all kingdoms of life, are used as counter screen RNA in many applications and especially in high-throughput screens of RNA molecules, such as an RNA G-quadruplex[51] and the TAR RNA.[52] To rule out binders not specific to RNA, we additionally screened five DNAs (Figures S24–S31), including four G-quadruplexes and five proteins with molecular weights ranging from 18 to 100 kDa (Figure S32–S39).

^{19}F-CPMG based screening by NMR spectroscopy

NMR spectroscopy is well suited for the identification of initial fragment hits as it is fast and reliable and offers the possibility to detect weak binding in solution. There are a large number of NMR experiments to detect binding, including detection of NOEs, chemical shift perturbation, saturation transfer difference,[53] WaterLOGSY[54] or T$_2$-relaxation spectroscopy.[55] By these methods, interactions characterized by dissociation constants in the range of 10 mM down to low-nanomolar can be detected, depending on experimental setup. Further improvements are currently developed with more sophisticated methods of dynamic nuclear polarization (DNP) or hyperpolarization,[56] decreasing the lower detection limit. Additionally, NMR offers the possibility to observe the interaction of the target with fragments in a mix of several fragments at the same time, greatly reducing operational effort. Often, NMR screenings use ^{1}H detection. The high number of hydrogen atoms in fragment compounds, however, leads to severe overlap of NMR signals, reducing the number of fragments within one mixture that can be screened in a single experiment.

^{19}F detection is an attractive alternative to ^{1}H-detection.[57] NMR signals in ^{19}F spectra show a much higher chemical shift dispersion covering a range of around 50 kHz (83 ppm) compared to around 6 kHz (10 ppm) for protons. Furthermore, if ^{1}H decoupling can be performed, which depends on the spectrometer configurations and NMR probe head used, each ^{19}F resonates at a single resonance frequency, allowing the observation of several fragments in a single mixture. Figure 2 shows ^{19}F-1D spectra demonstrating the design of the fragment mixtures containing 20 or 21 different ligands per mixture. An overview of all 101 fragments screened is available in Table S1. In this study, we measured ^{19}F transverse relaxation experiments which apply CPMG pulse trains[55,58] for varying relaxation delays.

CPMG T_2 measurements exploit the different relaxation properties of unbound fragments in comparison to (transiently) bound fragments to biological targets. Low-molecular-weight fragments with short rotational correlation times (τ_c) in solution will show changes in CPMG T_2 values upon (transient) binding to a high-molecular weight macromolecule (4–100 kDa), which exhibits much slower τ_c and consequently faster T_2 relaxation.

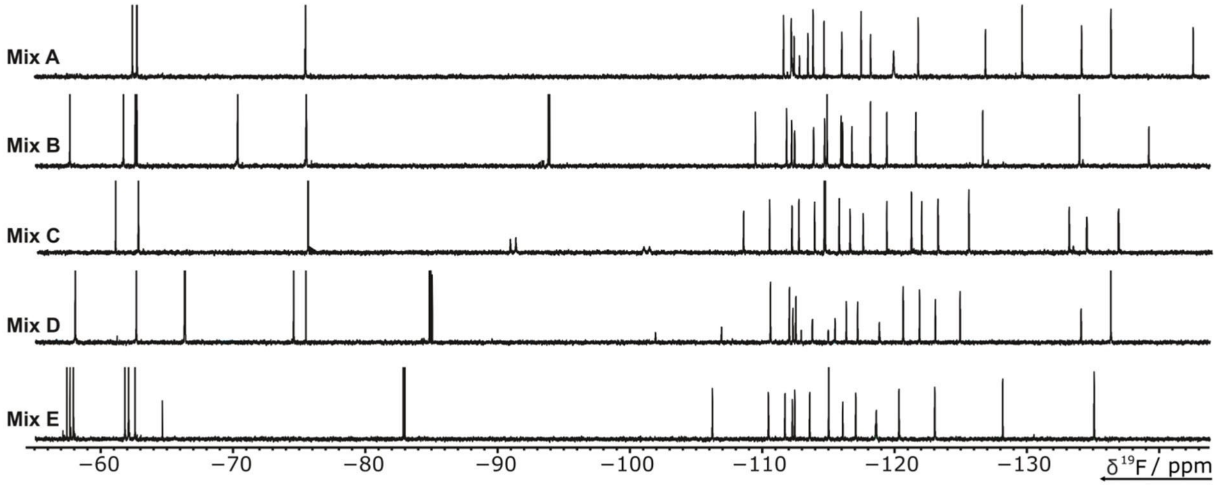


Figure 2. ^{19}F 1D NMR-spectra of the ^{19}F library fragment mixtures. The ^{19}F library contains 101 compounds (Table S1). Five mixtures of either 20 or 21 ligands were generated to avoid signal overlap. The spectra of the mixtures (A–E) in the screening buffer are displayed.

This relaxation effect is observed even if the population of bound fragment is 1 % or lower. We chose this method since it is very sensitive to even low-affinity interactions and is therefore beneficial to fragment-based approaches.

Validation of hits from ^{19}F screens

After this initial broad screening, follow-up screens integrate cheminformatic-based searches for similar ligands that are commercially available but also cross-validation of binding to other targets. In fact, some of the fluorine containing ligands show binding to almost all RNAs (e.g. fragment 57). Screening of fragments against RNA targets yielded several hits with a hit rate of up to 26 %. These hits were roughly estimated to bind with an at least low-millimolar K_D or tighter to the target RNA. This estimation stems from the observation that most RNA signals did not show large chemical shifts changes in the follow up 2D ^{1}H,^{15}N-correlation experiments, upon addition of fragments. This assumption is made, although it is not clear whether changes in chemical shifts are strictly correlated with changes in the chemical environment. After the initial ^{19}F screening the hits can be confirmed, K_D values determined and information on the fragment binding site obtained.

Mapping binding to a specific site in RNAs is usually performed by analysis of chemical shift perturbations (CSPs) that fragment binding causes on the RNA resonances in proximity to the binding epitope. In this approach, care has to be taken to distinguish direct binding induced CSPs from remote effects whose origins are sometimes difficult to assess. Most of the RNAs studied are riboswitches, which bind to metabolites of low molecular weight with an affinity several orders of magnitude higher than the expected affinity of the fluorinated fragments. Thus, orthosterically binding fragments can be detected in a competition experiment. By adding the natural ligand to an RNA sample containing the hit fragment, the natural ligand will compete for the RNA binding site and eventually displace the lower-affinity binding fragment. Accordingly, if a CPMG-experiment with long mixing time is recorded, fragment signals will be recovered upon addition of the natural ligand. Although these experiments can provide evidence for ligands that interact with the natural ligand binding site, this might not always be the case, since structural rearrangement upon binding of the natural ligand can also obscure other binding sites, which were formerly accessible to fragments. To evaluate the T_2-modulated 1D NMR spectra, we measured the fragment signal integrals and calculated the ratios between 200 ms CPMG and 0 ms CPMG applied to ^{19}F NMR spectra.

The quotient Q^{bind} of the integral ratios

$$Q^{bind} = \frac{Intensity\ Ratio^{\ +Target}}{Intensity\ Ratio^{\ -Target}}$$

with

$$Intensity\ Ratio = \frac{Peak\ Integral^{\ CPMG(200\ ms)}}{Peak\ Integral^{\ CPMG(0\ ms)}}$$

defines a quantifiable factor to classify the ligand-target interaction into no binding, strong or weak binding Figure 3. The exact effect, however, depends on the overall rotational tumbling time τ_c that increases with increasing molecular weight. Q^{bind} is a rough criterion, as we did not differentiate between aromatic and aliphatic bound flourines, which we however deem sufficient for the task. The quotient (Peak Integral$^{+\text{Target}}$/Peak Integral$^{-\text{Target}}$) was automatically calculated; all potential hits were then manually checked. The results are summarized in Figure 4 and spectra regions of all hits are displayed in Figures S5–S39. We obtained hits for all targets. The results show a clear trend that riboswitch RNAs show a high hit rate, ranging from 7 to 26 hits per riboswitch. Riboswitches contain aptamer domains that bind metabolites in

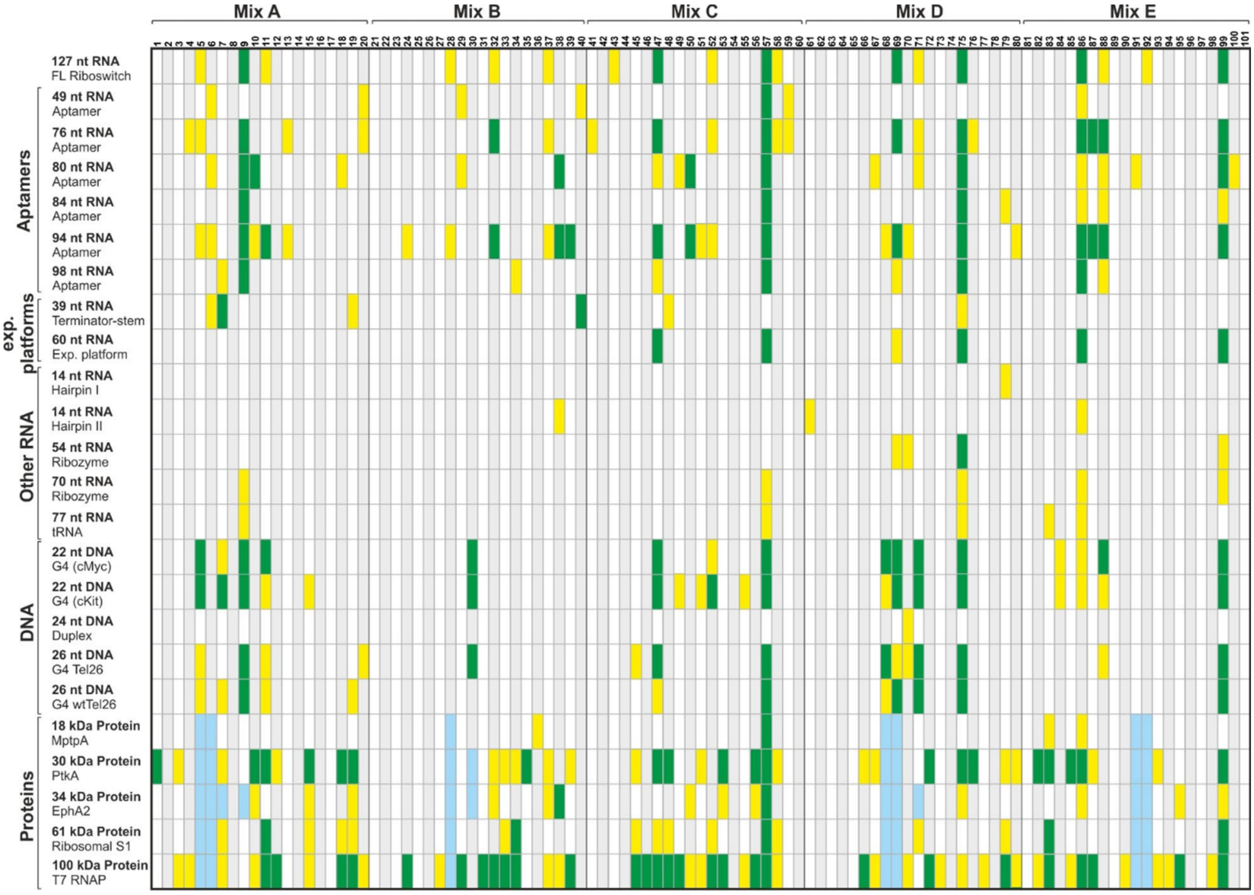


Figure 3. Interaction table of all fragments and biological targets screened. Hits were classified into no binding ($Q^{bind} > 0.67$, alternating gray and white), weak ($Q^{bind} = 0.66–0.33$, yellow) or strong binding ($Q^{bind} < 0.32$, green) in ^{19}F CPMG experiments. For protein screens, hits for ~5% of the ligands could not unambiguously be assigned (light blue).

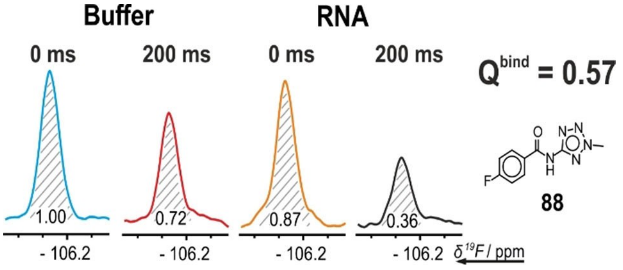


Figure 4. Determination of Q^{bind}. Four ^{19}F CPMG experiments were recorded to determine the binding factor Q^{bind} from peak integrals as discussed in the main text. The relaxation loss at 200 ms relaxation dephasing time relative to 0 ms dephasing for the ^{19}F signal of the ligand was recorded in the presence and absence of biomolecular target.

the same size range as the fragments or even lower (e.g., F^--sensing[59] or Mg^{2+}-sensing[60] riboswitches). Only one to six hits could be determined for all other RNAs. Although CPMG measurements as relaxation-based experiments are biased towards increased sensitivity for larger constructs, increased affinities of riboswitch targets are still striking in comparison to the 77 nt tRNA.

DNA structures which were included to sample other nucleic acid structures, namely duplex and G-quadruplex, showed very different behavior. For the duplex, only a single fragment showed binding. For G-quadruplexes, we observed between 12 and 20 hits with some overlap to hits that also bind to riboswitch RNAs. Of the five proteins investigated, four showed a large number of hits, ranging from 16 up to 55 (Table 1).

The 18 kDa phosphatase MptpA showed only four hits, in line with the difficult druggability of phosphatases. In general, for 101 fragments screened across 24 different biomolecular targets involving either DNA/RNA/proteins, approximately 5% of the data were not analyzable. This is a result of the necessity to optimize buffer conditions in particular for proteins. The different buffer conditions can lead to solubility issues and chemical stability issues for this subset of ligands. The remaining fragments show a broad variety of target selectivity from fragments binding exclusively a single target (fragment 100) to highly promiscuous binding behavior (fragment 57).

From the pool of the screened biological targets, we chose the aptamer domains containing 76, 84 and 98 nt of the secondary-messenger-sensing riboswitches for follow-up investigation. Weak hits were omitted. To verify hits and rule out effects of fragment mixing we confirmed binding of hits for single fragments. For all investigated fragment and RNA combinations, we were able to observe the same strong effects as in the mixtures. To validate the observed effects and to characterize the binding epitope, we analyzed the effect of fragment 75 addition to the 76 nt riboswitch (Figure 5). We used ^{15}N-isotopically labeled RNA to conduct ^{15}N-correlated 2D spectroscopy on imino hydrogens (Figure 5a) to detect possible chemical shift perturbation of RNA signals introduced by the addition of the screening hit. In these spectra, only helical imino hydrogen signals are visible, which shift distinctly in case of helix groove binding fragments. In our case, we could only observe very small shifts and sometimes additional small signals in the spectra. More pronounced effects could be observed on the signals of aromatic hydrogens in ^{1}H,^{1}H TOCSY spectra (Figure 5b). Here, clear signal shifts over 10 Hz on H5-H6 cross peaks of pyrimidine residues were detected.

At concentrations appropriate for screening (in this case 50 μM), and for large RNAs, such as the riboswitches investigated, only the strongest H5-H6 peaks are visible. In the example shown for compound 75 (Figure 5b), dose-dependent chemical shift perturbation was detected for three signals. Combined with the data obtained from ^{15}N-correlation spectroscopy, which showed only minor alteration, we can conclude that the respective binding site is located in a flexible region of the RNA.

Additional information on the binding site was obtained from a competitive binding assay. We added the fragment under investigation to the RNA target and characterized the effect of addition of native ligand on the T_2-modulated signal. As observed earlier, the signal is completely suppressed upon addition of the RNA, but addition of native ligand leads to signal recovery by approximately 15 %. The incomplete recovery of fragment signals points to the possibility that orthosteric binding of the fragment hit to the binding site of the cognate ligand is accompanied by additional nonspecific binding. The major population binds allosterically, and a smaller population binds orthosterically to the same binding site as cognate ligand. We found that signals of other compounds could be recovered by 83 % (compound 47, Figure 5d) after addition of native ligand, pointing to a larger influence on the binding site. The most convenient way to obtain affinity data by NMR is the observation of the ^{19}F fragment signal, since it does not require isotopic enrichment and there is only one signal. ^{19}F signals in general are very sensitive to changes in their chemical environment and thus, presumably, show also a substantial CSP for the ^{19}F upon addition of ligand. In contrast, the chemical shift dispersion of the aromatic hydrogens is smaller and the largest CSP in ^{1}H RNA signals was only around 5–8 Hz. From ^{19}F CSP data we could obtain affinity constants in the high-micromolar range, such as 400 μM for fragment 75.[62]

The screening of a large number of biomolecular targets demonstrates that the fragment library exhibits the highest hit rate to proteins, followed by RNA and then DNA. RNA hits are observed predominantly for RNAs containing loop regions, bulges, and internal loops. Some of the fragments in the library are promiscuously binding to all three different classes of biomolecular targets. The overlap for hits binding to proteins

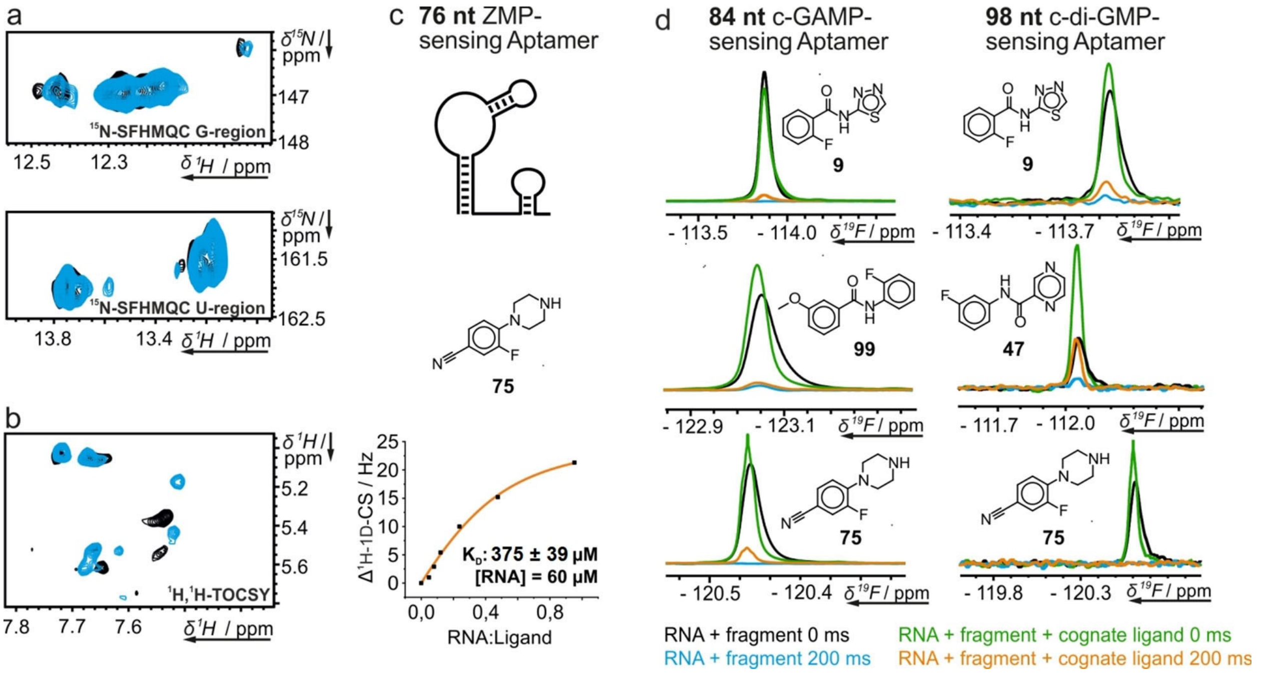


Figure 5. Hit validation and competition experiments. Validation of ^{19}F CPMG screening hits for the aptamer domains of the three secondary-messenger-sensing riboswitches. a) Spectral regions with signals from guanosine (top) and uridine (bottom) residues of the ^{1}H,^{15}N correlation experiment of the 76-nt riboswitch with (blue) and without (black) 75 shown under c. b) ^{1}H,^{1}H TOCSY spectrum with (blue) and without 75 (black). c) ^{19}F 1D NMR titration of 75 with the RNA. K_D was determined according to Williamson.[61] d) (Partially) competitive binding of fragments to the 84- and 98-nt riboswitch observed in T_2-modulated 1D ^{1}H experiments.

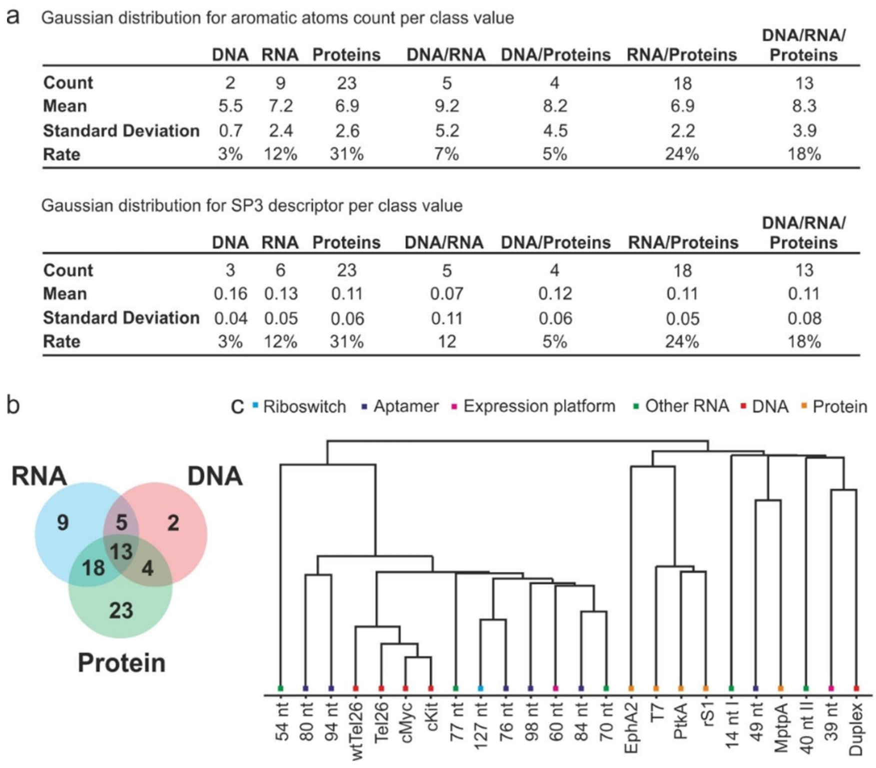
and RNA is also around 20%. The most promising result, however, is the observation that each biomolecular target class can be specifically targeted (Figure S4).

Cheminformatic analysis of hit data

The 69 hit compounds were chemically clustered using Hierarchical Clustering (DistMatrix, Morgan fingerprint, distance threshold 0.6, using Knime software 4.0.2) resulting in 38 singletons and 4 chemical families sharing a closely related scaffold. The largest cluster contains 5 members that were binders for proteins, DNA/RNA, and DNA/RNA/proteins targets. No cluster seems to be specific to any of the target families screened.

In order to check if there are any correlations between chemical structures and the number of targets that bind to it, we generated a number of molecular descriptors linked to the shape, electrostatic and hydrophobic interactions. The correlation matrix of all data shows no significant correlations between the number of target hits and molecular descriptors. The lowest and the highest correlations found are respectively with the number of aromatic atoms ($R = +0.27$) and SP3 descriptor (sp^3 carbon atom count/total carbon atom count reflecting the flatness of the molecules; $R = -0.23$).

The statistical analysis of the number of aromatic atoms for each category of binders shows higher average value of this descriptor for compounds hitting DNA/RNA/proteins (9.2) after the DNA/RNA binders (10.4) but in this last case the number of hits is too small to draw a conclusion. The SP3, like the other molecular descriptors, shows no significant difference between the category for hits (Figure 6). The substructure counting of popular and frequent motifs in organic molecules reported in the Table 6a, also shows no significant enrichment in different categories of binders. Because of the small size of the fragments in the current ^{19}F-library, there is no privileged class of compounds or relevant physicochemical properties that can be specific to a family of biological targets (RNA/DNA/proteins). The ^{19}F-fragment library is therefore suitable for RNA, DNA, and proteins and can be used to generate starting points for follow-up screens and to examine the presence of potential binding pockets for ligands in the individual targets. Moreover, correlation analysis (Figure 6c and Figure S4) shows striking clustering of hits between riboswitches and aptamers, DNA and proteins, respectively. Currently research focuses on the development of libraries solely suited for RNA, taking the repetitive nature of the RNA backbone, the higher charges and RNA dynamics into account.[63,64] While the here used library of ^{19}F-fragments was shown to be suited for proteins and RNA likewise, these recent developments could be taken into account in order to enrich current fragment libraries into a more RNA-focused library.



a Gaussian distribution for aromatic atoms count per class value

	DNA	RNA	Proteins	DNA/RNA	DNA/Proteins	RNA/Proteins	DNA/RNA/ Proteins
Count	2	9	23	5	4	18	13
Mean	5.5	7.2	6.9	9.2	8.2	6.9	8.3
Standard Deviation	0.7	2.4	2.6	5.2	4.5	2.2	3.9
Rate	3%	12%	31%	7%	5%	24%	18%

Gaussian distribution for SP3 descriptor per class value

	DNA	RNA	Proteins	DNA/RNA	DNA/Proteins	RNA/Proteins	DNA/RNA/ Proteins
Count	3	6	23	5	4	18	13
Mean	0.16	0.13	0.11	0.07	0.12	0.11	0.11
Standard Deviation	0.04	0.05	0.06	0.11	0.06	0.05	0.08
Rate	3%	12%	31%	12	5%	24%	18%

Figure 6. Cheminformatic analysis of hit data for all RNA, DNA and protein biomolecules. a) Gaussian distributions for aromatic atoms and SP3 descriptor over categories of biomolecules. SP3 descriptor (sp^3 carbon atom count/total carbon atom count) reflects the flatness of the fragment molecules. b) Visualization of categories in a Venn diagram. c) Euclidian distribution of hits to the target biomolecules.

Follow-up chemistry

After hit identification, the next major aspect in fragment-based drug discovery is to link initial hits to increase binding affinity. Such an improvement can be achieved by linkage of the fragment to an RNA binder, as shown for a neomycin-acridine conjugate.[65] To generate affinity we used an acridine moiety, which is a well-known intercalator that also allows fast follow-up chemistry.[66,67] Herein we describe a three step synthesis to generate an acridine moiety followed by linking it to a fragment. We chose a fragment from an additional ^{1}H screening for the 39-nt terminator stem (P2D11) with a similar structure to the hits from ^{19}F screening because of better availability and facilitated synthesis. On linking this fragment to the acridine, a binding in low-micromolar range was observed as described in the Supporting Information.

Conclusion

In summary, we outline the screening of ^{19}F-containing libraries to 14 different RNA targets. Commercially available fragment library can be used to identify low-molecular-weight fragments. We also show the general versatility of the used poised library, allowing straight-forward follow-up chemistry to increase binding affinity to as low as 1 μM, while observing a 15-fold selectivity between different RNA targets. Identified hits report on the general druggability of RNA targets.[68] Cheminformatics allow delineation of features within the fragments that are specific for each class of biomolecular target. Our study will aid our current effort to identify new ligands with antiviral activity in the context of combatting the COVID-19-pandemic (Covid19-nmr.de).

Experimental Section

Initial checks: All fragments in the library utilized for screening were checked for inconsistencies by running 1D ^{1}H and ^{19}F spectra of each compound. Spectra were analyzed by hand, and compounds showing wrong or additional signals were ruled out.

RNA preparation: All RNAs were prepared in house through in vitro transcription with T7 RNAP.[69] DNA templates included the necessary T7 promotor and were either obtained from linearizing plasmid containing the sequence of interest or PCR run-off. Transcription conditions were optimized for yield and sample purity and in vitro transcription was performed in 10 to 20 mL scale dependent on the expected yield. Purification was performed either by HPLC, preparative PAGE or buffer exchange to NMR buffer if necessary.[70] Concentration and purity of the samples were analyzed by UV/Vis spectroscopy and analytical PAGE respectively. For some of the follow-up experiments uniformly ^{15}N-labeled RNA was prepared with the same procedure using isotopically ^{15}N-labeled rNTPs.

Sample generation: Each fragment mixture contained 20 or 21 fragments at 2.5 mM concentration in 90% [D$_6$]DMSO with 10% D_2O. Mixtures were designed (minimize signal overlap) in an excel sheet, using the chemical shift obtained from individual compound measurement. SamplePro Tube robot was used for automated pipetting of the samples into the NMR tubes. The final sample volume was 170 μL with 5% D_2O as locking solvent. For each target, two samples (with and without target) were prepared per mixture. ^{19}F screening was performed at a ratio of 1:1 with respect to RNA and fragments. The final [RNA]([protein])-ligand concentration was around 50 μM. The screening buffer was 25 mM KPi, pH 6.2, 50 mM KCl, 5 mM $MgCl_2$ in 94% H_2O/5% D_2O/1% [D$_6$]DMSO in a 3 mm tube. For DNA, buffer conditions were 25 mM KPi, pH 7.0, 70 mM KCl, in 94% H H_2O/5% D_2O/1% [D$_6$]DMSO. For proteins MptpA and PtkA buffer conditions were 25 mM HEPES/NaOH, pH 7.0, 150 mM NaCl, 10 mM DTT and 50 mM HEPES/NaOH, pH 7.5, 300 mM NaCl, 10 mM DTT, 10 mM $MgCl_2$ respectively. For EphA2 buffer conditions were 20 mM Tris pH 8, 200 mM NaCl, 5 mM $MgCl_2$, 3 mM TCEP. For T7 and rS1 buffer conditions were 25 mM KPi (pH 7.2), 150 mM KCl, 5 mM DTT. NMR screening data of ^{19}F 1D and ^{19}F CPMG T_2 measurements were recorded with mixing times of 0, 200, and 400 ms for the CPMG experiments. Strong hits from screening of three investigated RNAs were chosen for follow-up experiments. Samples of these fragments with the respective RNA were prepared in the same way omitting mixture generation, to confirm binding of the single fragment. In the same way, samples for the competition experiments were generated. Native ligand was added at an equimolar concentration to the RNA. Additionally, samples of ^{15}N-labeled RNA were prepared with an RNA concentration between 50 and 100 μM and fragment concentration of 1.25 mM and were used in follow-up experiments. Reference samples contained no fragment.

NMR spectroscopy: Spectra acquisition was carried out on a Bruker AVIIIHD-600 NMR spectrometer equipped with a five mm ^{1}H/^{19}F [^{13}C,^{15}N]-TCI prodigy cryo-probe and high throughput sample changer for 579 samples with temperature option for sample storage. For the screening process the following spectra were acquired: ^{19}F 1D, water-suppressed proton 1D and ^{19}F 1D with CPMG spinlock (0, 200 and 400 ms). All ^{19}F spectra were recorded at room temperature, without ^{1}H decoupling and processed with line broadening function of 10 Hz. The CPMG spin lock was applied using an adiabatic WURST pulse with a bandwidth of 120 ppm and a length of 2 ms. The pulse is calculated on-the-fly by wavemaker software in Topspin. The interpulse delay of the CPMG spin lock was set to 9 ms. Screening data were analyzed by integration with Topspin 4.0 (Bruker Biospin) and manually checked using the integrated fragment-based screening software tool. Strong hits for three RNAs investigated were chosen for follow-up experiments. ^{19}F 1D CPMG spectra were run on single fragment samples with the same parameters as in screening. The following spectra were acquired: ^{15}N SFHMQC, proton 1Ds with excitation sculpting[71] or jump-and-return echo[72] scheme for water suppression, and ^{1}H,^{1}H TOCSY with excitation sculpting.

Synthesis: All experimental details for the synthesis and characterization of compound **1** are described in the Supporting Information.

Acknowledgements

The work was supported by funds from DFG in collaborative research center 902: "Molecular principles of RNA-based regulation", by EU-funded project iNEXT-discovery. Work at BMRZ is supported by the state of Hesse. V.D.J. was supported by Boehringer Ingelheim Fonds. R.S. was supported by Fonds der Chemischen Industrie. Open access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: DNA · ^{19}F · FBS · fluorine · fragment-based screening · proteins · RNA

[1] Q. Vicens, E. Westhof, *ChemBioChem* **2003**, *4*, 1018–1023.
[2] K. D. Warner, C. E. Hajdin, K. M. Weeks, *Nat. Rev. Drug Discovery* **2018**, *17*, 547–558.
[3] A. Garcia-Lopez, F. Tessaro, H. R. A. Jonker, A. Wacker, C. Richter, A. Comte, N. Berntenis, R. Schmucki, K. Hatje, O. Petermann, G. Chiriano, R. Perozzo, D. Sciarra, P. Konieczny, I. Faustino, G. Fournet, M. Orozco, R. Artero, F. Metzger, M. Ebeling, P. Goekjian, B. Joseph, H. Schwalbe, L. Scapozza, *Nat. Commun.* **2018**, *9*, 1–12.
[4] M. Sivaramakrishnan, K. D. McCarthy, S. Campagne, S. Huber, S. Meier, A. Augustin, T. Heckel, H. Meistermann, M. N. Hug, P. Birrer, A. Moursy, S. Khawaja, R. Schmucki, N. Berntenis, N. Giroud, S. Golling, M. Tzouros, B. Banfai, G. Duran-Pacheco, J. Lamerz, Y. Hsiu Liu, T. Luebbers, H. Ratni, M. Ebeling, A. Cléry, S. Paushkin, A. R. Krainer, F. H. T. Allain, F. Metzger, *Nat. Commun.* **2017**, *8*, 1–13.
[5] A. Donlic, A. E. Hargrove, *Wiley Interdiscip. Rev.: RNA* **2018**, *9*, e1477.
[6] M. Matsui, D. R. Corey, *Nat. Rev. Drug Discovery* **2017**, *16*, 167–179.
[7] J. Palacino, S. E. Swalley, C. Song, A. K. Cheung, L. Shu, X. Zhang, M. Van Hoosear, Y. Shin, D. N. Chin, C. G. Keller, M. Beibel, N. A. Renaud, T. M. Smith, M. Salcius, X. Shi, M. Hild, R. Servais, M. Jain, L. Deng, C. Bullock, M. McLellan, S. Schuierer, L. Murphy, M. J. J. Blommers, C. Blaustein, F. Berenshteyn, A. Lacoste, J. R. Thomas, G. Roma, G. A. Michaud, B. S. Tseng, J. A. Porter, V. E. Myer, J. A. Tallarico, L. G. Hamann, D. Curtis, M. C. Fishman, W. F. Dietrich, N. A. Dales, R. Sivasankaran, *Nat. Chem. Biol.* **2015**, *11*, 511–517.
[8] L. S. Waters, G. Storz, *Cell* **2009**, *136*, 615–628.
[9] D. van Duin, D. L. Paterson, *Inf. Disp.* **2016**, *30*, 377–390.
[10] M. S. Butler, M. A. Cooper, *Curr. Drug Targets* **2012**, *13*, 373–387.
[11] D. R. Davies, *Structural Genomics and Drug Discovery*, Humana Press, New York, NY, **2014**, pp 315–323.
[12] T. Sugiki, K. Furuita, T. Fujiwara, C. Kojima, *Molecules* **2018**, *23*, 148.
[13] X. Liu, D. Shi, S. Zhou, H. Liu, H. Liu, X. Yao, *Expert Opin. Drug Discovery* **2018**, *13*, 23–37.
[14] D. A. Erlanson, *in Top. Curr. Chem.*, **2011**, pp. 1–32.
[15] D. A. Erlanson, I. J. P. de Esch, W. Jahnke, C. N. Johnson, P. N. Mortenson, *J. Med. Chem.* **2020**, acs.jmedchem.9b01581.
[16] S. B. Shuker, P. J. Hajduk, R. P. Meadows, S. W. Fesik, *Science* **1996**, *274*, 1531–1534.
[17] J. Tsai, J. T. Lee, W. Wang, J. Zhang, H. Cho, S. Mamo, R. Bremer, S. Gillette, J. Kong, N. K. Haass, K. Sproesser, L. Li, K. S. M. Smalley, D. Fong, Y.-L. Zhu, A. Marimuthu, H. Nguyen, B. Lam, J. Liu, I. Cheung, J. Rice, Y. Suzuki, C. Luu, C. Settachatgul, R. Shellooe, J. Cantwell, S.-H. Kim, J. Schlessinger, K. Y. J. Zhang, B. L. West, B. Powell, G. Habets, C. Zhang, P. N. Ibrahim, P. Hirth, D. R. Artis, M. Herlyn, G. Bollag, *Proc. Mont. Acad. Sci.* **2008**, *105*, 3041–3046.
[18] P. A. Brough, W. Aherne, X. Barril, J. Borgognoni, K. Boxall, J. E. Cansfield, K.-M. J. Cheung, I. Collins, N. G. M. Davies, M. J. Drysdale, B. Dymock, S. A. Eccles, H. Finch, A. Fink, A. Hayes, R. Howes, R. E. Hubbard, K. James, A. M. Jordan, A. Lockie, V. Martins, A. Massey, T. P. Matthews, E. McDonald, C. J. Northfield, L. H. Pearl, C. Prodromou, S. Ray, F. I. Raynaud, S. D. Roughley, S. Y. Sharp, A. Surgenor, D. L. Walmsley, P. Webb, M. Wood, P. Workman, L. Wright, *J. Med. Chem.* **2008**, *51*, 196–218.
[19] S. Howard, V. Berdini, J. A. Boulstridge, M. G. Carr, D. M. Cross, J. Curry, L. A. Devine, T. R. Early, L. Fazal, A. L. Gill, M. Heathcote, S. Maman, J. E. Matthews, R. L. McMenamin, E. F. Navarro, M. A. O'Brien, M. O'Reilly, D. C. Rees, M. Reule, D. Tisi, G. Williams, M. Vinković, P. G. Wyatt, *J. Med. Chem.* **2009**, *52*, 379–388.
[20] C.-M. Park, M. Bruncko, J. Adickes, J. Bauch, H. Ding, A. Kunzer, K. C. Marsh, P. Nimmer, A. R. Shoemaker, X. Song, S. K. Tahir, C. Tse, X. Wang, M. D. Wendt, X. Yang, H. Zhang, S. W. Fesik, S. H. Rosenberg, S. W. Elmore, *J. Med. Chem.* **2008**, *51*, 6902–6915.
[21] Y.-S. Wang, C. Strickland, J. H. Voigt, M. E. Kennedy, B. M. Beyer, M. M. Senior, E. M. Smith, T. L. Nechuta, V. S. Madison, M. Czarniecki, B. A. McKittrick, A. W. Stamford, E. M. Parker, J. C. Hunter, W. J. Greenlee, D. F. Wyss, *J. Med. Chem.* **2010**, *53*, 942–950.
[22] E. Largy, F. Hamon, M.-P. Teulade-Fichou, *Anal. Bioanal. Chem.* **2011**, *400*, 3419–3427.
[23] P. N. Asare-Okai, C. S. Chow, *Anal. Biochem.* **2011**, *408*, 269–276.
[24] S. Laughlin, W. Wilson, *Int. J. Mol. Sci.* **2015**, *16*, 24506–24531.
[25] H. Shimizu, F. Jinno, A. Morohashi, Y. Yamazaki, M. Yamada, T. Kondo, S. Asahi, *J. Mass Spectrom.* **2012**, *47*, 1015–1022.
[26] J. Sztuba-Solinska, S. R. Shenoy, P. Gareiss, L. R. H. Krumpe, S. F. J. Le Grice, B. R. O'Keefe, J. S. Schneekloth, *J. Am. Chem. Soc.* **2014**, *136*, 8402–8410.
[27] C. M. Connelly, R. E. Boer, M. H. Moon, P. Gareiss, J. S. Schneekloth, *ACS Chem. Biol.* **2017**, *12*, 435–443.
[28] M. H. Moon, T. A. Hilimire, A. M. Sanders, J. S. Schneekloth, *Biochemistry* **2018**, *57*, 4638–4643.
[29] M. Jerabek-Willemsen, T. André, R. Wanner, H. M. Roth, S. Duhr, P. Baaske, D. Breitsprecher, *J. Mol. Struct.* **2014**, *1077*, 101–113.
[30] M. Pellecchia, I. Bertini, D. Cowburn, C. Dalvit, E. Giralt, W. Jahnke, T. L. James, S. W. Homans, H. Kessler, C. Luchinat, B. Meyer, H. Oschkinat, J. Peng, H. Schwalbe, G. Siegal, *Nat. Rev. Drug Discovery* **2008**, *7*, 738–745.
[31] M. Garavís, B. López-Méndez, A. Somoza, J. Oyarzabal, C. Dalvit, A. Villasante, R. Campos-Olivas, C. González, *ACS Chem. Biol.* **2014**, *9*, 1559–1566.
[32] M. Zeiger, S. Stark, E. Kalden, B. Ackermann, J. Ferner, U. Scheffer, F. Shoja-Bazargani, V. Erdel, H. Schwalbe, M. W. Göbel, *Bioorg. Med. Chem. Lett.* **2014**, *24*, 5576–5580.
[33] D. R. Groebe, O. C. Uhlenbeck, *Nucleic Acids Res.* **1988**, *16*, 11725–11735.
[34] J. P. Sheehy, A. R. Davis, B. M. Znosko, *RNA* **2010**, *16*, 417–429.
[35] J. L. Childs-Disney, J. Hoskins, S. G. Rzuczek, C. A. Thornton, M. D. Disney, *ACS Chem. Biol.* **2012**, *7*, 856–862.
[36] A. C. Stelzer, A. T. Frank, J. D. Kratz, M. D. Swanson, M. J. Gonzalez-Hernandez, J. Lee, I. Andricioaei, D. M. Markovitz, H. M. Al-Hashimi, *Nat. Chem. Biol.* **2011**, *7*, 553–559.
[37] C. Hong, M. Hagihara, K. Nakatani, *Angew. Chem. Int. Ed.* **2011**, *50*, 4390–4393; *Angew. Chem.* **2011**, *123*, 4482–4485.
[38] J. P. Marino, R. S. Gregorian, G. Csankovszki, D. M. Crothers, *Science* **1995**, *268*, 1448–1454.
[39] M. H. Bailor, A. M. Mustoe, C. L. Brooks, H. M. Al-Hashimi, *Curr. Opin. Struct. Biol.* **2011**, *21*, 296–305.
[40] M. H. Bailor, C. Musselman, A. L. Hansen, K. Gulati, D. J. Patel, H. M. Al-Hashimi, *Nat. Protoc.* **2007**, *2*, 1536–1546.
[41] F. Hamy, V. Brondani, A. Flörsheimer, W. Stark, M. J. J. Blommers, T. Klimkait, *Biochemistry* **1998**, *37*, 5086–5095.
[42] A. S. Brodsky, J. R. Williamson, *J. Mol. Biol.* **1997**, *267*, 624–639.
[43] A. T. Frank, A. C. Stelzer, H. M. Al-Hashimi, I. Andricioaei, *Nucleic Acids Res.* **2009**, *37*, 3670–3679.
[44] N. N. Patwardhan, L. R. Ganser, G. J. Kapral, C. S. Eubanks, J. Lee, B. Sathyamoorthy, H. M. Al-Hashimi, A. E. Hargrove, *MedChemComm* **2017**, *8*, 1022–1036.
[45] M. Mandal, B. Boese, J. E. Barrick, W. C. Winkler, R. R. Breaker, *Cell* **2003**, *113*, 577–586.
[46] H. Schwalbe, J. Buck, B. Fürtig, J. Noeske, J. Wöhnert, *Angew. Chem. Int. Ed.* **2007**, *46*, 1212–1219; *Angew. Chem.* **2007**, *119*, 1232–1240.
[47] S. D. Gilbert, S. J. Mediatore, R. T. Batey, *J. Am. Chem. Soc.* **2006**, *128*, 14214–14215.
[48] E. R. Lee, K. F. Blount, R. R. Breaker, *RNA Biol.* **2009**, *6*, 187–194.
[49] A. Reining, S. Nozinovic, K. Schlepckow, F. Buhr, B. Fürtig, H. Schwalbe, *Nature* **2013**, *499*, 355–359.
[50] C. Helmling, D. Klötzner, F. Sochor, R. A. Mooney, A. Wacker, R. Landick, B. Fürtig, A. Heckel, H. Schwalbe, *Nature Commun.* **2018**, 9, **n.d.**
[51] M. Garavís, B. López-Méndez, A. Somoza, J. Oyarzabal, C. Dalvit, A. Villasante, R. Campos-Olivas, C. González, *ACS Chem. Biol.* **2014**, *9*, 1559–1566.
[52] L. R. Ganser, J. Lee, A. Rangadurai, D. K. Merriman, M. L. Kelly, A. D. Kansal, B. Sathyamoorthy, H. M. Al-Hashimi, *Nat. Struct. Mol. Biol.* **2018**, *25*, 425–434.
[53] D. W. Begley, S. O. Moen, P. G. Pierce, E. R. Zartler, in *Current Protocols in Chemical Biology*, Wiley, Hoboken, **2013**, pp. 251–268.
[54] C. Dalvit, G. Fogliatto, A. Stewart, M. Veronesi, B. Stockman, *J. Biomol. NMR* **2001**, *21*, 349–359.
[55] H. Y. Carr, E. M. Purcell, *Phys. Rev.* **1954**, *94*, 630–638.
[56] Y. Kim, C. Hilty, *PPAR Res.* **2019**, 501–526.

[57] C. Dalvit, P. E. Fagerness, D. T. A. Hadden, R. W. Sarver, B. J. Stockman, *J. Am. Chem. Soc.* **2003**, *125*, 7696–7703.
[58] S. Meiboom, D. Gill, *Rev. Sci. Instrum.* **1958**, *29*, 688–691.
[59] J. L. Baker, N. Sudarsan, Z. Weinberg, A. Roth, R. B. Stockbridge, R. R. Breaker, *Science* **2012**, *335*, 233–235.
[60] C. E. Dann, C. A. Wakeman, C. L. Sieling, S. C. Baker, I. Irnov, W. C. Winkler, *Cell* **2007**, *130*, 878–892.
[61] M. P. Williamson, *Prog. Nucl. Magn. Reson. Spectrosc.* **2013**, *73*, 1–16.
[62] S. H. Rüdisser, N. Goldberg, M.-O. Ebert, H. Kovacs, A. D. Gossert, *J. Biomol. NMR* **2020**, *7*, DOI: 10.1007/s10858-020-00325-x.
[63] S. G. Rzuczek, M. R. Southern, M. D. Disney, *ACS Chem. Biol.* **2015**, *10*, 2706–15.
[64] B. S. Morgan, J. E. Forte, R. N. Culver, Y. Zhang, A. E. Hargrove, *Angew. Chem. Int. Ed.* **2017**, *56*, 13498–13502; *Angew. Chem.* **2017**, *129*, 13683–13687.
[65] S. R. Kirk, N. W. Luedtke, Y. Tor, *J. Am. Chem. Soc.* **2000**, *122*, 980–981.
[66] P. Prasher, M. Sharma, *MedChemComm* **2018**, *9*, 1589–1618.
[67] J. F. Arambula, S. R. Ramisetty, A. M. Baranger, S. C. Zimmerman, *Proc. Mont. Acad. Sci.* **2009**, *106*, 16068–16073.
[68] P. J. Hajduk, J. R. Huth, C. Tse, *Drug Discovery Today* **2005**, *10*, 1675–1682.
[69] J. K. Bains, J. Blechar, V. de Jesus, N. Meiser, H. Zetzsche, B. Fürtig, H. Schwalbe, M. Hengesbach, *Methods* **2019**, *153*, 22–34.
[70] C. Helmling, S. Keyhani, F. Sochor, B. Fürtig, M. Hengesbach, H. Schwalbe, *J. Biomol. NMR* **2015**, *63*, 67–76.
[71] T. L. Hwang, A. J. Shaka, *J. Magn. Reson. Ser. A* **1995**, *112*, 275–279.
[72] G. Marius Clore, B. J. Kimber, A. M. Gronenborn, *J. Magn. Reson.* **1983**, *54*, 170–173.
[73] L. Huang, J. Wang, D. M. J. Lilley, *Cell Chem. Biol.* **2017**, *24*, 695–702.e2.
[74] A. Ren, K. R. Rajashankar, D. J. Patel, *Structure* **2015**, *23*, 1375–1381.
[75] A. Serganov, A. Polonskaia, A. T. Phan, R. R. Breaker, D. J. Patel, *Nature* **2006**, *441*, 1167–1171.
[76] A. Ren, X. C. Wang, C. A. Kellenberger, K. R. Rajashankar, R. A. Jones, M. C. Hammond, D. J. Patel, *Cell Rep.* **2015**, *11*, 1–12.
[77] K. D. Smith, S. V. Lipchock, A. L. Livingston, C. A. Shanahan, S. A. Strobel, *Biochemistry* **2010**, *49*, 7351–7359.
[78] K. D. Smith, S. V Lipchock, T. D. Ames, J. Wang, R. R. Breaker, S. A. Strobel, *Nat. Struct. Mol. Biol.* **2009**, *16*, 1218–1223.
[79] N. Kulshina, N. J. Baird, A. R. Ferré-D'amaré, *Nat. Struct. Mol. Biol.* **2009**, *16*, 1212–1217.
[80] A. Serganov, Y.-R. Yuan, O. Pikovskaya, A. Polonskaia, L. Malinina, A. T. Phan, C. Hobartner, R. Micura, R. R. Breaker, D. J. Patel, *Chem. Biol.* **2004**, *11*, 1729–1741.
[81] J. R. Stagno, Y. Liu, Y. R. Bhandari, C. E. Conrad, S. Panja, M. Swain, L. Fan, G. Nelson, C. Li, D. R. Wendel, T. A. White, J. D. Coe, M. O. Wiedorn, J. Knoska, D. Oberthuer, R. A. Tuckey, P. Yu, M. Dyba, S. G. Tarasov, U. Weierstall, T. D. Grant, C. D. Schwieters, J. Zhang, A. R. Ferré-D'Amaré, P. Fromme, D. E. Draper, M. Liang, M. S. Hunter, S. Boutet, K. Tan, X. Zuo, X. Ji, A. Barty, N. A. Zatsepin, H. N. Chapman, J. H. Spence, S. A. Woodson, Y. X. Wang, *Nature* **2017**, *541*, 242–246.
[82] J. Zhang, A. R. Ferré-D'Amaré, *Structure* **2014**, *22*, 1363–1371.
[83] C. Helmling, A. Wacker, M. T. Wolfinger, I. L. Hofacker, M. Hengesbach, B. Fürtig, H. Schwalbe, *J. Am. Chem. Soc.* **2017**, *139*, 2647–2656.
[84] R. Rieder, K. Lang, D. Graber, R. Micura, *ChemBioChem* **2007**, *8*, 896–902.
[85] R. Schnieders, C. Richter, S. Warhaut, V. de Jesus, S. Keyhani, E. Duchardt-Ferner, H. Keller, J. Wöhnert, L. T. Kuhn, A. L. Breeze, W. Bermel, H. Schwalbe, B. Fürtig, *J. Biomol. NMR* **2017**, *69*, 31–44.
[86] K. Okada, M. Takahashi, T. Sakamoto, G. Kawai, K. Nakamura, A. Kanai, *Nucleosides Nucleotides Nucleic Acids* **2006**, *25*, 383–395.
[87] F. M. Jucker, A. Pardi, *Biochemistry* **1995**, *34*, 14416–14427.
[88] W. G. Scott, J. T. Finch, A. Klug, *Cell* **1995**, *81*, 991–1002.
[89] A. R. Ferré-D'Amaré, K. Zhou, J. A. Doudna, *Nature* **1998**, *395*, 567–574.
[90] N. H. Woo, B. A. Roe, A. Rich, *Nature* **1980**, *286*, 346–351.
[91] A. Ambrus, D. Chen, J. Dai, R. A. Jones, D. Yang, *Biochemistry* **2005**, *44*, 2048–2058.
[92] J. Dai, M. Carver, L. H. Hurley, D. Yang, *J. Am. Chem. Soc.* **2011**, *133*, 17673–17680.
[93] D. Wei, G. N. Parkinson, A. P. Reszka, S. Neidle, *Nucleic Acids Res.* **2012**, *40*, 4691–4700.
[94] A. T. Phan, V. Kuryavyi, S. Burge, S. Neidle, D. J. Patel, *J. Am. Chem. Soc.* **2007**, *129*, 4386–4392.
[95] H. R. Drew, R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura, R. E. Dickerson, *Proc. Natl. Acad. Sci. USA* **1981**, *78*, 2179–2183.
[96] J. Dai, C. Punchihewa, A. Ambrus, D. Chen, R. A. Jones, D. Yang, *Nucleic Acids Res.* **2007**, *35*, 2440–2450.
[97] W. Liu, Y. F. Zhong, L. Y. Liu, C. T. Shen, W. Zeng, F. Wang, D. Yang, Z. W. Mao, *Nat. Commun.* **2018**, *9*, 3496.
[98] J. Wirmer-Bartoschek, L. E. Bendel, H. R. A. Jonker, J. T. Grün, F. Papi, C. Bazzicalupi, L. Messori, P. Gratteri, H. Schwalbe, *Angew. Chem. Int. Ed.* **2017**, *56*, 7102–7106; *Angew. Chem.* **2017**, *129*, 7208–7212.
[99] J. Dai, M. Carver, C. Punchihewa, R. A. Jones, D. Yang, *Nucleic Acids Res.* **2007**, *35*, 4927–4940.
[100] C. Madhurantakam, E. Rajakumara, P. A. Mazumdar, B. Saha, D. Mitra, H. G. Wiker, R. Sankaranarayanan, A. K. Das, *J. Bacteriol.* **2005**, *187*, 2175–2181.
[101] T. Stehle, S. Sreeramulu, F. Löhr, C. Richter, K. Saxena, H. R. A. Jonker, H. Schwalbe, *J. Biol. Chem.* **2012**, *287*, 34569–34582.
[102] A. Niesteruk, H. R. A. Jonker, C. Richter, V. Linhard, S. Sreeramulu, H. Schwalbe, *J. Biol. Chem.* **2018**, *293*, 11823–11836.
[103] S. Heinzlmeir, D. Kudlinzki, S. Sreeramulu, S. Klaeger, S. L. Gande, V. Linhard, M. Wilhelm, H. Qiao, D. Helm, B. Ruprecht, K. Saxena, G. Médard, H. Schwalbe, B. Kuster, *ACS Chem. Biol.* **2016**, *11*, 3400–3411.
[104] P. Giraud, J. B. Créchet, M. Uzan, F. Bontems, C. Sizun, *Biomol. NMR Assign.* **2015**, *9*, 107–111.
[105] P. Salah, M. Bisaglia, P. Aliprandi, M. Uzan, C. Sizun, F. Bontems, *Nucleic Acids Res.* **2009**, *37*, 5578–5588.
[106] Y. W. Yin, T. A. Steitz, *Science* **2002**, *298*, 1387–1395.

Manuscript received: July 15, 2020
Revised manuscript received: August 11, 2020
Accepted manuscript online: August 14, 2020
Version of record online: September 25, 2020

5.17 **Research article:** Combined smFRET and NMR analysis of riboswitch structural dynamics

Jasleen Kaur Bains*, Julius Blechar*, Vanessa de Jesus*, Nathalie Meiser*, Heidi Zetzsche, Boris Fürtig, Harald Schwalbe and Martin Hengesbach; *Methods* **153**, 22-34 (2019)

* These authors contributed equally to this work and are listed in alphabetical order.

This article summarizes the results of folding state and structural dynamics studies of the *V. vulnificus* adenine-sensing riboswitch identified by single-molecule Förster resonance energy transfer (smFRET) microscopy and NMR spectroscopy have been consolidated. Detailed protocols for sample preparation and for both smFRET and NMR measurements were provided. The combination of both techniques allows to the study of helical arrangements using information about base pairing patterns of different secondary structures together with the long-range distance information. Furthermore, the coexistence of two conformations can be studied in detail.

The author of this thesis contributed to the literature screening and manuscript preparation for the the sections on sample preparation and NMR experiments together with V. de Jesus, B. Fürtig and H. Schwalbe. The paragraphs on smFRET sample preparation and measurements were prepared by J. Blechar, N. Meiser and M. Hengesbach.

https://doi.org/10.1016/j.ymeth.2018.10.004

Methods 153 (2019) 22–34

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Combined smFRET and NMR analysis of riboswitch structural dynamics

Jasleen Kaur Bains[1], Julius Blechar[1], Vanessa de Jesus[1], Nathalie Meiser[1], Heidi Zetzsche, Boris Fürtig, Harald Schwalbe, Martin Hengesbach*

Goethe-University Frankfurt, Institute for Organic Chemistry and Chemical Biology, Max-von-Laue-Str. 7, 60438 Frankfurt, Germany

1. Introduction

Analysis of RNA structure and dynamics faces one major challenge: a number of techniques that resolve structural properties often do not provide sufficient dynamic information. There have been numerous attempts to overcome this challenge for both proteins and RNA. Over recent years, integrated approaches have been developed to use several techniques with yield synergistic but also complementary information and reconcile the experimental data in a unifying mechanistic model. Nuclear magnetic resonance (NMR) spectroscopy and single molecule Förster resonance energy transfer (smFRET) are techniques that yield exactly such information. Especially for RNA: While NMR easily resolves canonical and non-Watson-Crick base pairing patterns and small inter-atom distances as well as dynamics up to the second timescale, it suffers from mapping long-range structural dynamics and becomes increasingly difficult to apply for larger RNAs above 200 nucleotides. smFRET on the other hand provides information on coexisting conformational populations and reports dynamics in real-time for a single molecule, thus potentially providing a larger number of different dynamics due to individual sampling of single molecules. As only label positions are observed, smFRET provides limited structural information by assessing only one distance vector.

A very small number of studies have been using NMR and smFRET in a truly collaborative manner to study RNA systems. While studies have reported on Mg^{2+} binding sites by NMR [1], confirmed structural models [2], or qualitatively assess RNA-protein binding [3] there has been little work towards a truly complementary information from both NMR and smFRET experiments. We have recently combined these techniques to identify both folding states and structural dynamics of the three-state adenine-sensing riboswitch from *Vibrio vulnificus* [4,5]. Our approach allowed us to characterize folded conformations and their exchange dynamics by NMR. smFRET confirmed populations and fundamentally slow exchange dynamics and revealed, in addition, that these long-lived states still showed unexpected dynamics in smFRET experiments. Based on this integration of structurally resolved, dynamic states the mechanistic understanding of the functional regulatory pathway of the *add* riboswitch requiring substantial RNA refolding could be established. We found that despite several orders of magnitude of difference in the concentration of the analyzed RNA, the results obtained from both methods were congruent, and offered additional insight that would not have been obtainable from either method alone.

Here, we provide detailed protocols for both NMR and smFRET measurements, which yield the abovementioned complementary data. While neither of these techniques can be exhaustively covered within the scope of this study, we aim at providing tools that facilitate ease of use and integration of both techniques. We furthermore show how integration of results from both techniques provide unprecedented insight into structural dynamics of the adenine-responsive riboswitch from *Vibrio vulnificus*.

1.1. Analysis of conformational dynamics in RNA by smFRET spectroscopy

Single molecule Förster resonance energy transfer (smFRET) microscopy is an excellent tool to study molecular dynamics [6,7]. The measured structural dynamics can be analyzed regarding the number of different states, their dwell times, often also referred to as life times, and population. Resulting FRET efficiencies provide semi-quantitative information on the distance vector between labeling sites, however, achievement of quantitative results is challenging. One of the major advantages of smFRET is the compatibility with many different buffer conditions combined with requirement of only low sample concentrations. In smFRET, selection of labeling sites is challenging. In our application, we focus on measurements of immobilized samples using a total internal reflection setup. Such setup enables millisecond to second resolution which, however, also depends on the cameras frame rate. To prepare immobilized samples for smFRET measurements, the construct has to fulfill several requirements. Labeling of one RNA construct with two different fluorophores is necessary to record intramolecular dynamics. Therefore, labeling sites have to be chosen carefully [8]. To realize labeling with two different fluorophores, the RNA is obtained commercially divided into two (or more) fragments, each containing one single modifiable site, i.e. 5-aminoallyl-modified uridine, on the chosen labelling site. Additionally, one of the end fragments has to be attached to a biotin linker. After coupling of the fluorophores to the

* Corresponding author.

E-mail address: hengesbach@nmr.uni-frankfurt.de (M. Hengesbach).

[1] These authors contributed equally to this manuscript and are listed in alphabetical order.

https://doi.org/10.1016/j.ymeth.2018.10.004
Received 10 September 2018; Received in revised form 8 October 2018; Accepted 9 October 2018
Available online 12 October 2018

correct RNA fragment, the RNA fragments are combined to the full construct via DNA splint ligation. For immobilization of the construct onto the smFRET slides surface, this surface needs to be streptavidin coated (we use BSA-biotin-streptavidin coating). Detailed information on this procedure can be found in the following protocol together with smFRET measurement and data analysis guidelines.

1.2. Analysis of conformational dynamics in RNA by NMR

Nuclear magnetic resonance (NMR) spectroscopy is a powerful technique for the investigation of atomic structure and its dynamic changes of RNAs, proteins and their complexes in solution. Structural dynamics can be resolved over a wide range of timescales [9], extending from processes such as bond vibrations at 10^{-9} s up to large-scale conformational rearrangements of RNAs secondary structure at 10^3 s [10,11]. NMR has the advantage to study RNA molecules in a biological relevant context under near to physiological conditions such as temperature, pH value, ion- and metabolite concentrations. The preparation of ^{13}C, ^{15}N isotope-labelled RNA is crucial and allows the structure determination of RNA fragments about 100 nucleotides [12]. A detailed and optimized RNA preparation and purification protocol for RNA isotope labelling is described in the following section. To investigate the structure and ligand-dependent folding of the translation-regulating *add* adenine riboswitch from the gram-negative human pathogenic marine bacterium *Vibrio vulnificus* (Asw) we outline several powerful NMR spectroscopy methods that yield information on numerous types of conformational states and intermolecular interactions [4,13].

2. Materials and methods

2.1. Sample preparation (smFRET)

The solid 2′-ACE protected RNA oligonucleotide fragments (Dharmacon, Riboswitch fragments) containing a single 5-aminoallyl-modified uridine residue was reconstituted at a concentration of 1 mM in water each. One fragment per construct needs to be biotinylated. 0.1 vol 3 M sodium acetate (pH 5.2) and 2.5 vol abs. ethanol were added to 30 µL oligonucleotide solution and the solution cooled for 10 min at −80 °C. The RNA was pelleted through centrifugation at 13,000 rpm and −4 °C for 15 min and the pellet dried in an Eppendorf Vacufuge plus concentrator (2 min, 30 °C, vacuum setting V_{al}). For dye coupling, the pellet was reconstituted in 20 µL 100 mM sodium bicarbonate (pH 8.2), or for deprotection in 200 µL deprotection buffer (100 mM acetic acid adjusted to pH 3.8 with TEMED; Dharmacon).

2.1.1. Dye coupling

Cy3 or Cy5 Mono-Reactive dye pack vial (Amersham) contents were dissolved in 20 µL DMSO and the dye was mixed with the oligonucleotide in 20 µL 100 mM sodium bicarbonate. The mix was incubated for 90 min, and the reaction protected from ambient light. Precipitation was performed by adding 0.1 vol 3 M sodium acetate (pH 5.2) and 2.5 vol abs. ethanol and cooling for 10 min at −80 °C. The RNA was pelleted through centrifugation at 13000 rpm and −4 °C for 20 min and the pellet air-dried for 5 min. The pellet was redissolved in 200 µL deprotection buffer.

2.1.2. Deprotection

The RNA oligonucleotide (labelled or unlabelled) was heated to 60 °C for 30 min and the biotinylated fragment for 2 h (dye labelled oligonucleotides: protect the reaction from ambient light) and the RNA precipitated by adding 0.1 vol 3 M sodium acetate (pH 5.2) and 2.5 vol abs. ethanol and cooling for 10 min at −80 °C. The RNA was pelleted through centrifugation at 13000 rpm and −4 °C for 20 min and the pellet air-dried for 5 min. The dye labelled RNA was reconstituted in 60 µL TEAA buffer (100 mM trimethylamine adjusted to pH 7.0 with acetic acid) for HPLC purification.

2.1.3. HPLC purification of dye-labelled RNA oligonucleotides

A C8 HPLC column (e.g. Kromasil) was used, with buffers (A) TEAA buffer (100 mM trimethylamine adjusted to pH 7.0 with acetic acid) and (B) acetonitrile for HPLC purification of dye-labelled oligonucleotides. The oligonucleotide sample was heated to 60 °C for 3 min prior to loading. Elution was performed with a linear gradient from 0% B to 100% B over 160 min at a flow rate of 0.5 mL/min. Absorbance was monitored at 260 nm for nucleic acid, 550 nm for Cy3 and 650 nm for Cy5. TEAA buffer was removed by drying the fractions of dye-labelled RNA in Eppendorf Vacufuge plus concentrator (10 min, 30 °C, vacuum setting V_{aq}). Precipitation was achieved by adding 0.1 vol 3 M sodium acetate (pH 5.2) and 2.5 vol abs. ethanol and cooling at −80 °C for 10 min. The RNA was pelleted by centrifugation at 13,000 rpm and −4 °C for 15 min and the pellets air-dried for 5 min. The pellet was reconstituted in water at ~30 A_{260} units. Fractions were analysed by denaturing polyacrylamide gel electrophoresis and by UV/vis spectroscopy and fractions were combined that showed a single band on the denaturing gel and a comparable UV/vis absorption ratio at the fluorophore over the RNA absorption maximum.

2.1.4. DNA-splinted enzymatic RNA ligation

The ligation reactions were performed under exclusion of light on a suitable scale (preferably > 100 pmol). Oligonucleotides and the DNA splint were annealed at 2–10 µM concentration in T4 RNA ligase 2 buffer (New England Biolabs) containing 20% v/v PEG 8000 by heating at 75 °C for 3 min and cooling at room temperature for 10 min. The mixture was incubated with 200 U/mL T4 RNA ligase 2 (New England Biolabs) at 37 °C for 1 h. 70 U/mL T4 RNA ligase 2 were added and the incubation was continued for 30 min. 70 U/mL Turbo DNase (Ambion) were added and the incubation was continued for 30 min. The mixture was extracted three times with 1 vol Roti®-Aqua-P/C/I (Carl Roth) and twice with chloroform. Rapid phase separation was achieved by centrifugation (2000 rpm, 4 °C, 3 min). 1 vol Aqua-Phenol and 3 vol diethyl ether can be used instead of Roti®-Aqua-P/C/I and chloroform. RNA was precipitated from the aqueous phase by addition of 0.1 vol 5 M ammonium acetate (pH 5.2) and 2.5 vol ethanol and cooling at -80 °C for 10 min. The RNA was pelleted by centrifugation at 13,000 rpm and −4 °C for 30 min. The pellet was dried in an Eppendorf Vacufuge plus concentrator (2 min, 30 °C, vacuum setting V_{al}). The pellet was reconstituted in 12 µL water and 12 µL formamide buffer (90% v/v formamide, 90 mM Tris, 90 mM boric acid, 2 mM EDTA) for purification by denaturing polyacrylamide gel electrophoresis.

2.1.5. RNA purification with PAGE

A gel was prepared with 10 mL 10% v/v acrylamide bis-acrylamide (29:1) in TBE buffer (90 mM Tris, 90 mM boric acid, 2 mM EDTA, pH 8.0) with 8 M urea and polymerized by addition of 0.1% w/v ammonium persulfate and 0.1% v/v TEMED. The gel was prerun in TBE buffer at 225 V for 10 min. Crude ligation product was loaded adjacent to a reference lane containing 3% of the ligation product. The gel was run under protection from light at 225 V for 30 min. The target band was identified by eye (preferred) or by scanning a fluorescent band next to the RNA to be isolated for precise localization of the desired fragment. The target band was excised and frozen for ≥10 min at −80 °C. Successful excision of the desired band was verified by imaging the gel via fluorescence detection (Typhoon scanner; GE Healthcare). The RNA was eluted from the gel slice in 400 µL 0.5 M ammonium acetate (pH 5.2) by shaking in an Eppendorf thermoshaker (450 rpm, r.t., overnight). The supernatant was filtered afterwards (e.g. Nanosep spin filter, Pall). The RNA was precipitated by adding 2.5 vol abs. ethanol and incubation at −80 °C for 1.5 h. The RNA was pelleted through centrifugation at 13,000 rpm and −4 °C for 60 min and the pellet dried in an Eppendorf Vacufuge plus concentrator (5 min, 30 °C, vacuum setting V_{al}). The pellet was reconstituted in an appropriate amount

(15–20 µL) water and the concentration determined by UV/vis spectroscopy via the absorbance of the RNA.

2.1.6. smFRET sample preparation

0.5 pmol FRET construct were refolded at 1 nM concentration in smFRET immobilization buffer (25 mM K_2HPO_4/KH_2PO_4, 50 mM KCl, pH 7.0) by denaturing at 85 °C and 15 min cooling at room temperature for single molecule fluorescence measurements. The RNA was diluted to a series of 50–200 pM samples in smFRET immobilization buffer supplemented with different concentrations of $MgCl_2$ and ligand. Samples were kept on ice until measurement within 24 h. Measurement slides were assembled from standard microscope slides and glass coverslips (24 × 60 mm; Carl Roth). The glass microscope slides were washed in 1 M KOH for 15 min and for 15 min in water and the glass microscope slide was dried with compressed air. Glass coverslips were cleaned in N_2 plasma for 10 min (Diener Electronic). Small stripes of parafilm were aligned vertically along the long edge of the glass slide with a ~3 mm spacing to confine measurement channels with ~10 µL volume. The clean surface of the coverslip was pressed gently onto the parafilm stripes. The assembly was fixed by placing the slide onto an 85 °C heatblock for 1 min. A calibration sample with 0.1 µm fluorescent beads was prepared. For this, a channel was flushed with a previously prepared 1:500 dilution of fluorescent beads (e.g. TetraSpeck, Thermo Fisher Scientific) in phosphate buffered saline. For the RNA samples, the channels were flushed shortly before measurement with 30 µL 1 mg/mL biotinylated BSA (Sigma Aldrich) in buffer T50 (10 mM Tris-HCl, pH 8.0, 50 mM NaCl) and incubated for 10 min. The channel was flushed with 100 µL buffer T50. The channel was then flushed with 30 µL 0.2 mg/mL streptavidin (Molecular Probes, Thermo Fisher Scientific) in buffer T50 and incubated for 10 min. The channel waswashed with 100 µL smFRET immobilization buffer (25 mM K_2HPO_4/KH_2PO_4, 50 mM KCl, pH 7.0). BSA-biotin-streptavidin coated slides were used within 12 h. The RNA sample was immobilized directly before measurement. The channel was first flushed with 200 µL smFRET immobilization buffer containing equivalent concentrations of $MgCl_2$ and ligand. 50 µL of the RNA sample were flushed through the channel and incubated for 1 min. The channel was rinsed with 125 µL imaging buffer (25 mM K_2HPO_4/KH_2PO_4, 50 mM KCl, pH 7.0) containing equivalent concentrations of $MgCl_2$ and adenine, an enzymatic oxygen scavenging system (10% glucose, 14 U/mL glucose oxidase, 1000 U/mL catalase) and saturating amounts of trolox. The immobilization procedure, especially the RNA concentration, should be optimized to result in an optimal surface density for data acquisition.

2.1.7. smFRET spectroscopy

Fig. 1 provides an overview of the microscopy setup required for smFRET spectroscopy, including alternating laser excitation. As a detailed description would exceed the boundaries of the methods provided here, we would like to refer the reader to literature providing excellent overviews of equipment and application options [7,14–16].

All measurements were performed with an integration time of 100 ms and maximum EM gain about 3 min after imaging buffer exchange. Acquired movies were saved as *.sif files. One 20 frame long calibration movie of the fluorescent beads (TetraSpeck) sample was recorded. ~20 movies of 20 frame length per sample were measured for smFRET histograms. > 10 movies of ~1500 frame length per sample were obtained for smFRET time traces.

2.2. smFRET analysis

2.2.1. Data preparation

The data was converted from *.sif format acquired datasets into *.raw files. A transformation map for alignment of donor and acceptor channel was generated with the calibration movie file using IDL scripts (Bokinsky et al., 2003), as was the acquired histogram and trace data.

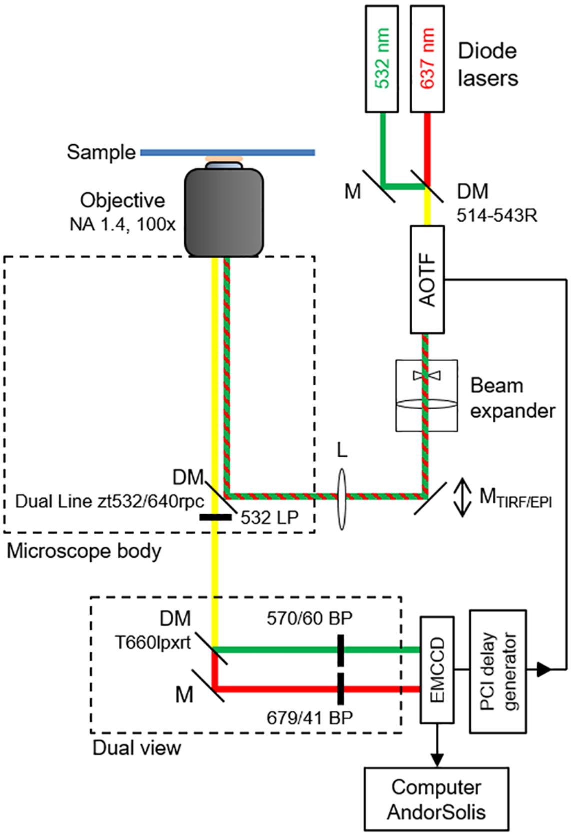


Fig. 1. Typical total internal reflection fluorescence (TIRF) microscopy setup used for smFRET (adapted from [17]). The setup contains one red and one green laser. The combination of dichroitic mirrors (DM), lenses (L) and mirrors (M) focuses the laser beam on the back focal plane of the objective. Depending on the position of the movable mirror TIRF or EPI illumination can be selected. The sample fluorescence is split into FRET donor and acceptor fluorescence. Filters (long pass (LP), band pass (BP)) remove scattered light before the EMCCD camera detects the fluorescence. The AOTF gives the opportunity for alternating laser excitation (e.g. for smALEX). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

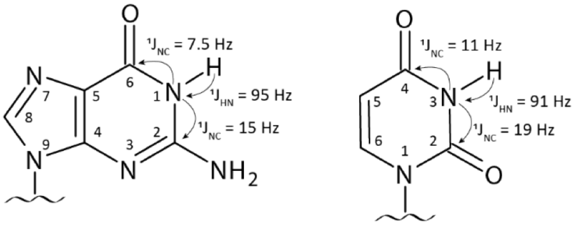


Fig. 2. Depiction of the through bond magnetization transfer during HNCO experiments in guanine and uracil utilising $^1J_{XN}$ couplings.

2.2.2. smFRET histograms

The files generated from the 20 frame movies were used to generate a smFRET histogram. Therefore, an automated Matlab script was used which corrects acceptor fluorescence intensities for donor fluorescence leakage, calculates the FRET efficiency (E) with $E = I_A/(I_A + I_D)$, selects molecules with fluorescence intensities ($I_{sum} = I_A + I_D$) higher than a selected threshold, and bins the calculated FRET efficiency of the selected molecules in desired intervals. The smFRET histogram can be plotted by using the obtained files. In addition, Gaussian distribution fitting can be applied. Subtraction of the Gaussian arising from donor-only molecules can be done afterwards (see Fig. 6).

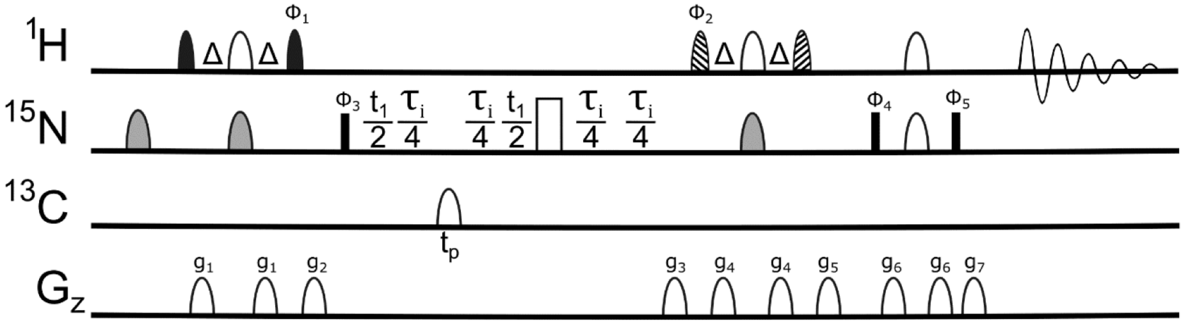


Fig. 3. [^{1}H,^{15}N]-BEST-TRACT pulse sequence. Narrow and wide pulse symbols indicate 90° and 180° rf pulses and semi-ellipses indicate for selective pulses. Unless indicated, all pulses are applied with phase x. All selective 1H pulses are centered at 12.3 ppm with the following shapes and field strengths: polychromatic PC9 [24] (narrow filled pulse symbol) with a field strength of 6.4 kHz, REBURP [25] (wide unfilled pulse symbol) with a field strength of 8 kHz, E-BURP2tr and E-BURP2 (stripped pulse symbols) with a field strength of 5.6 kHz. The selective 13C adiabatic composite smoothed chirp pulse [26] is centered at 110 ppm with a field strength of 89 kHz. All selective ^{15}N pulses are centered at 153.5 ppm with the following shapes: bipolar BIP-720-50-20 [29] (wide grey filled pulse symbol) with a field strength of 21.6 kHz and EBURP REBURP [25] with a field strength of 6.8 kHz. The delays Δ are defined as $\Delta = 1/4 * {}^1J(NH) = 2.7$ ms. T_1 is the ^{15}N chemical shift evolution period. Pulsed field gradients g_1–g_7 are applied along the z-axis (G_z) with the following durations: 250 µs for g_1, g_2 and g_4, 500 µs for g_3, g_5 and g_7, 1 ms for g_6. The gradient pulses are applied with a smoothed square amplitude (SMSQ10.32) with strengths of 2% (g_1), 21% (g_2), −80% (g_3), 5% (g_4), 30% (g_5), 45% (g_6) and 30.13% (g_7) (100% corresponds to 53 G/cm). The phase cycle is: $\Phi_1 = y$, $\Phi_2 = -y$ for alpha plane and $\Phi_2 = y$ for beta plane, $\Phi_3 = x, -x$, $\Phi_4 = y$, $\Phi_5 = -x$ and the receiver phase $\Phi_{rec} = x, -y$. The relaxation delay τ_i is incremented from 12 µs to 40 ms to determine the rotational correlation times τ_c.

2.2.3. smFRET time traces

The files generated from the 1500 frame movies were used to gain dwell times and to plot transition occupancy density plots (TODP). Therefore, custom Matlab scripts and the publicly available HaMMy (Hidden Markov modelling software) [18,19] were used. The final output files contain a TODP, the HMM fits and the corrected HMM fits of E each sorted by static and dynamic molecules. Additionally, a *.txt file with extracted dwell times is created (see Fig. 7).

2.3. Sample preparation (NMR)

2.3.1. Preparation of RNAs by in vitro transcription

DNA templates were prepared by PCR amplification or large-scale plasmid DNA preparation. The DNA template was designed with a T7 promotor and optionally self-cleaving 5′- and 3′-ribozymes [20]. Standard T7 polymerase requires the presence of guanosine at ideally the three first transcribed nucleotides on the 5′-terminus adjacent to the promotor sequence [21]. If the RNA sequence of interest does not match these conditions, a 5′-ribozyme should be introduced to improve transcription yields. Further, the standard T7 polymerase generates

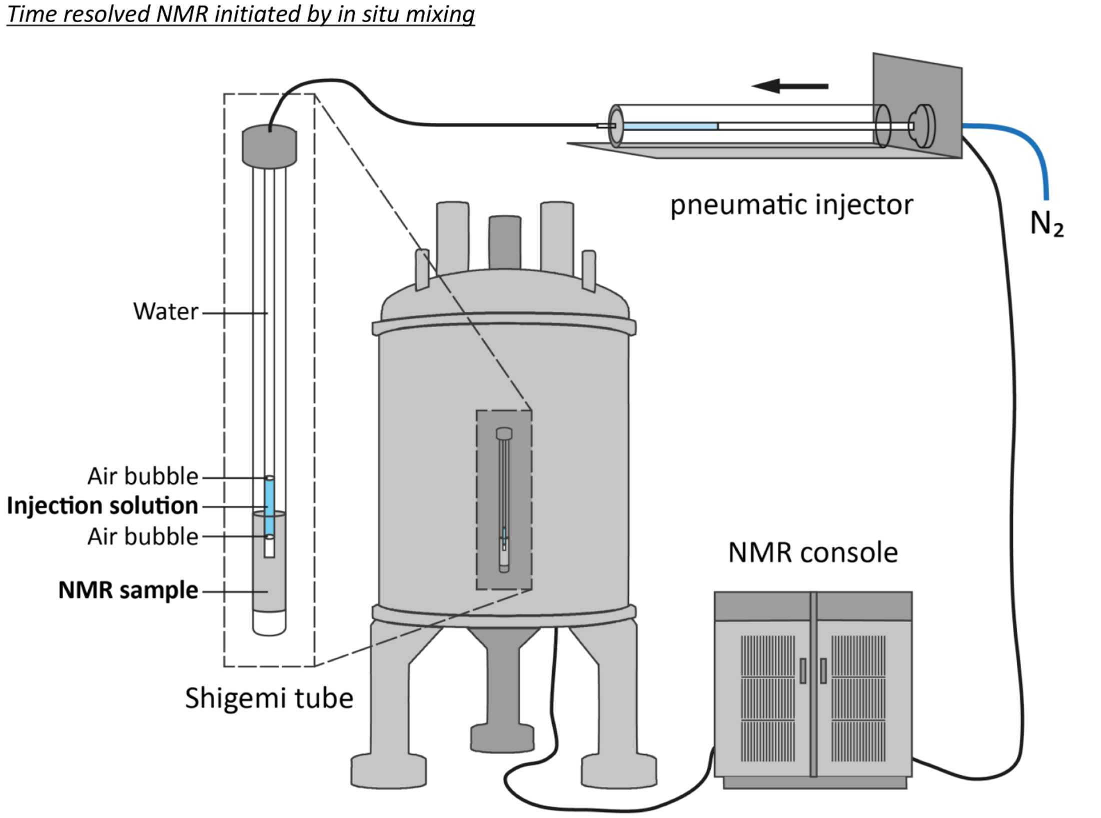


Fig. 4. Schematic representation of real-time NMR mixing setup including modified NMR tube with mounted injector, pneumatic syringe and hardware connectivity.

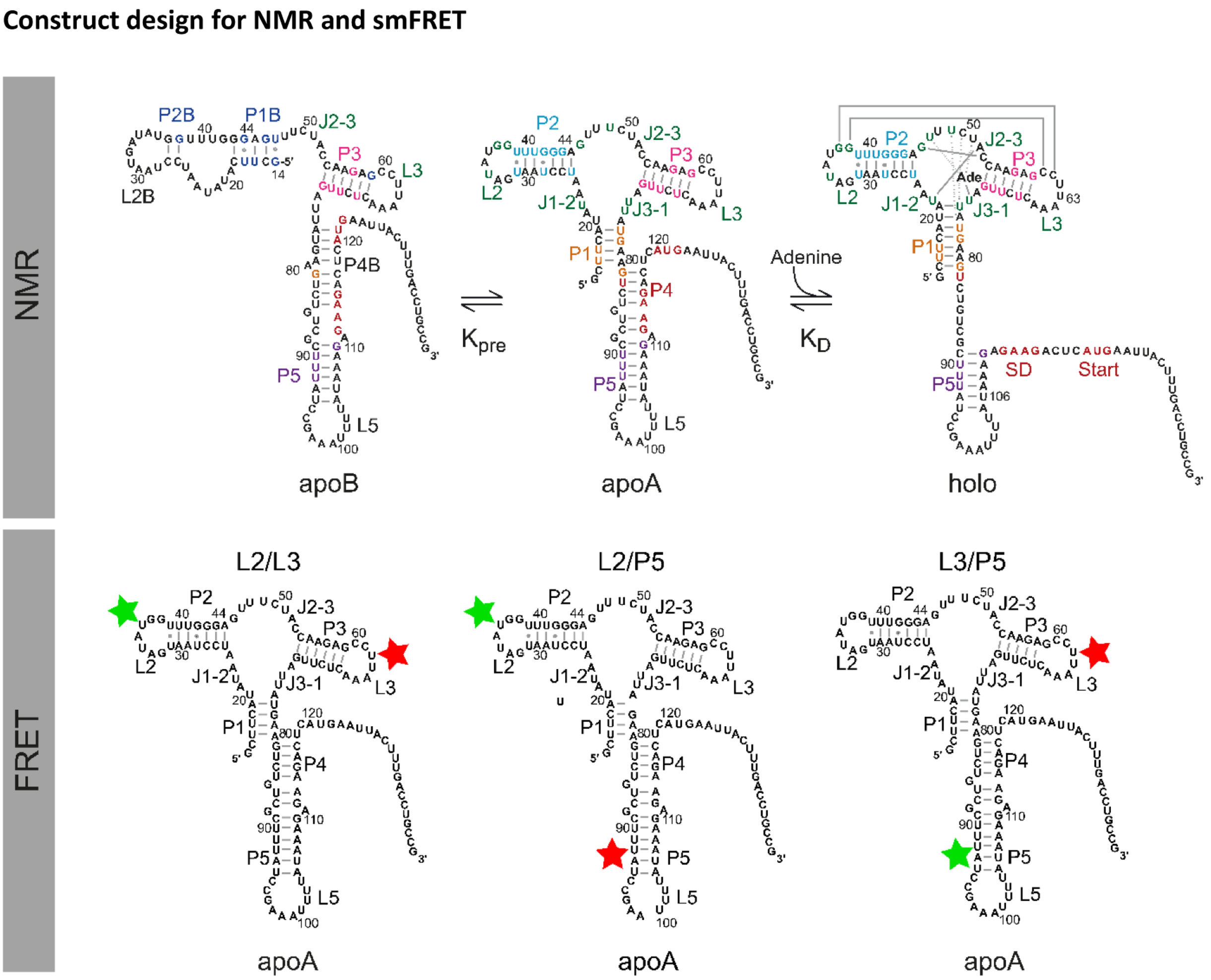


Fig. 5. Three-state model of the adenine-sensing riboswitch from Vibrio vulnificus (127 nucleotides). The three conformations apoB, apoA and holo are linked by two equilibria K_{pre} and K_D. ApoB lacks the ligand-sensing three-way junction of the aptamer domain and has an extended P4B helix. In both apo states, the ribosome binding site which contains the Shine-Dalgarno (SD) sequence is masked by base pairing. In addition, the start codon (AUG) is also base paired in the apoB conformation. The holo state forms a more compact structure with kissing loop interaction between L2 and L3 and with solvent exposed ribosome binding site. For NMR experiments, the RNA is synthesized using ^{15}N-labeled NTPs. For FRET analysis, three labeling schemes are used.

inhomogeneous 3′-ends during run-off transcription. For 3′-homogeneity, it is advantageous to employ a 3′-ribozyme. Commonly used are 5′-hammerhead and 3′-HDV (hepatitis delta virus) ribozymes [22].

In vitro transcription from crude PCR products or linearized plasmids was performed with T7 RNA polymerase. The transcription conditions were optimized in small scale test transcriptions (25 µL with respect to

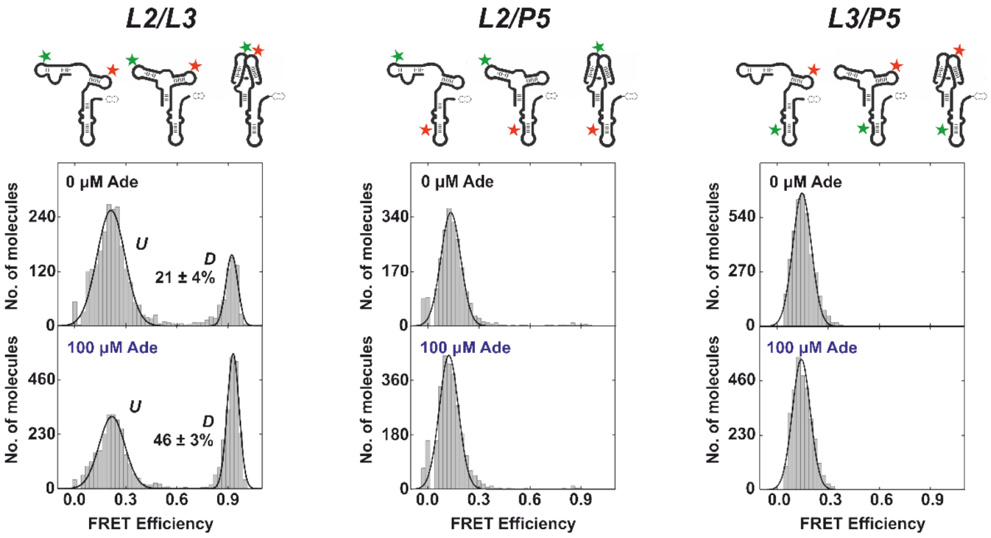


Fig. 6. smFRET histograms of Asw with three different labelling schemes to monitor adenine-induced switching (taken from [13]). Used labelling schemes are indicated within the schematic representation of the Asw above the corresponding histograms. Data was collected in the presence of 2 mM Mg^{2+}. The shown histograms were obtained and Gaussian fitted as described in the protocol. The donor only peak was removed. Fractional population of the docked (D) state is indicated for L2/L3 labeling scheme.

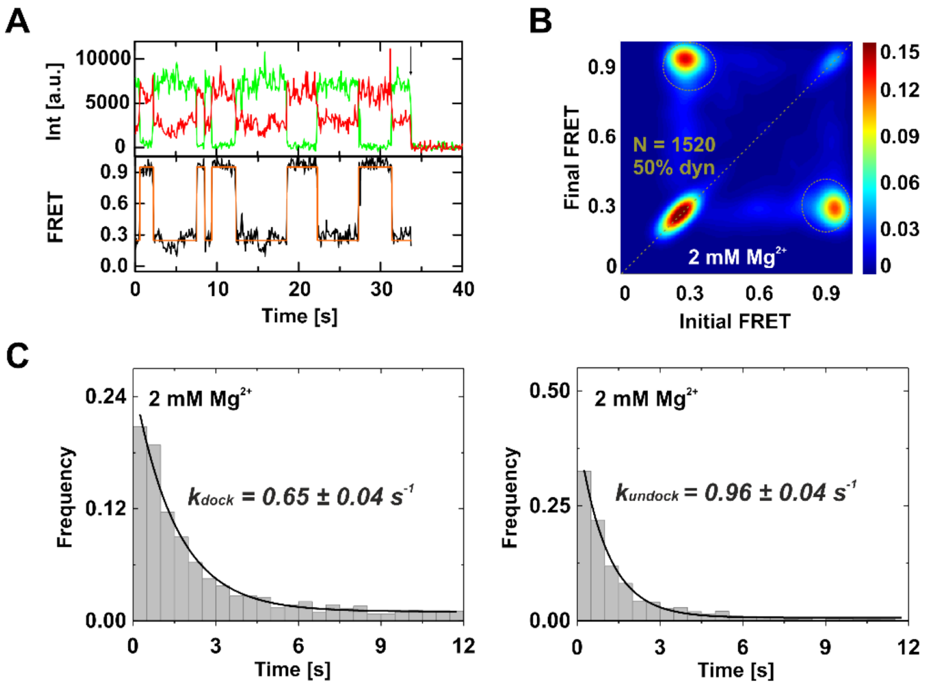


Fig. 7. Representative results for smFRET time trace analysis from full length Asw at 2 mM Mg^{2+} and without adenine (adapted from [13]). A: A time trace (green indicates I_D, red indicates I_A) for one single molecule is shown. The resulting FRET efficiency (black line) is overlaid with the corresponding HaMMy fit (orange line). B: TODP obtained from 1520 different time traces, showing 50% of the molecules being dynamic. C: Dwell time histograms of the undocked (left) and docked state (right) fitted with single exponential decay functions give the indicated rate constants. For analysis, dynamic time traces determined from the TODP were used. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the Mg^{2+} (5 mM–75 mM), the rNTP (2 mM–6 mM each) and DNA concentrations (8%–6% [v/v] for PCR products and 40 ng/μL–120 ng/μL for plasmid template). Test transcriptions were incubated for 4 h at 37 °C and analyzed by denaturing polyacrylamide gel electrophoresis. Preparative transcriptions were performed in 25 mL reactions in a 50 mL tube with parameters optimized in test transcriptions. For isotope-labeled RNA for NMR spectroscopy, 4 mM of ^{15}N- or ^{13}C- ^{15}N-labeled rNTPs (Silantes) were used. The transcription mix was prepared by adding water, transcription buffer (250 mM Tris glutamate, pH 8.1), 2 mM Spermidine, 20 mM DTT, magnesium acetate, DNA template and rNTPs. T7 polymerase was added to the reaction and incubated for 16 h at 37 °C and 150 rpm in a thermoshaker. Further, 0.08 u/μL yeast inorganic pyrophosphatase (YIPP) were added after 2 h in order to increase the transcription yield by enzymatic hydrolysis of pyrophosphate. After incubation, the transcription reactions were stored at −20 °C until purification of the target RNA.

2.3.2. RNA purification by PAGE and folding

The transcription reaction was desalted by solvent exchange into water via centrifugal concentrators. The solvent exchange was performed until a constant UV-absorbance ratio A_{260}/A_{200} is reached. The RNA sample (1 mL, ~1000 A_{260} units) was mixed with an equal volume of formamide. Preparative polyacrylamide gel was prepared using a 16.0 × 47.5 cm 12% denaturing polyacrylamide gel using a TVS1000 sequencer electrophoresis system (Biostep) connected to a high-voltage electrophoresis power supply (PowerPac 3000, Bio-Rad). The gel was prepared with 250 mL 12% [v/v] acrylamide bis-acrylamide (29:1) and 7 M urea in TBE buffer (90 mM Tris, 90 mM boric acid, 2 mM EDTA, pH 8.0). Gel is polymerized by addition of 0.1% [w/v] ammonium persulfate and 0.1% [v/v] TEMED. The RNA was loaded on the gel

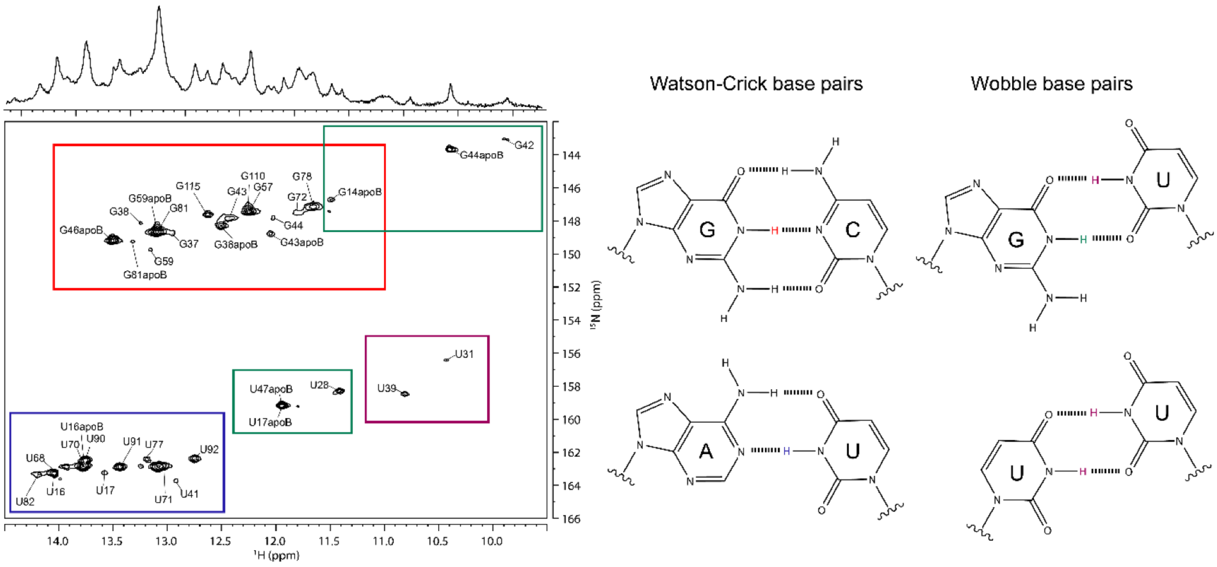


Fig. 8. Exemplary [^{1}H,^{15}N]-BEST-TROSY spectrum for the assignment of imino proton resonances. The characteristic chemical shifts for Watson-Crick base pairs are shown in red (G-C) and blue (A-U). Wobble base pairs are highlighted in green (G-U) and purple (U-U). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

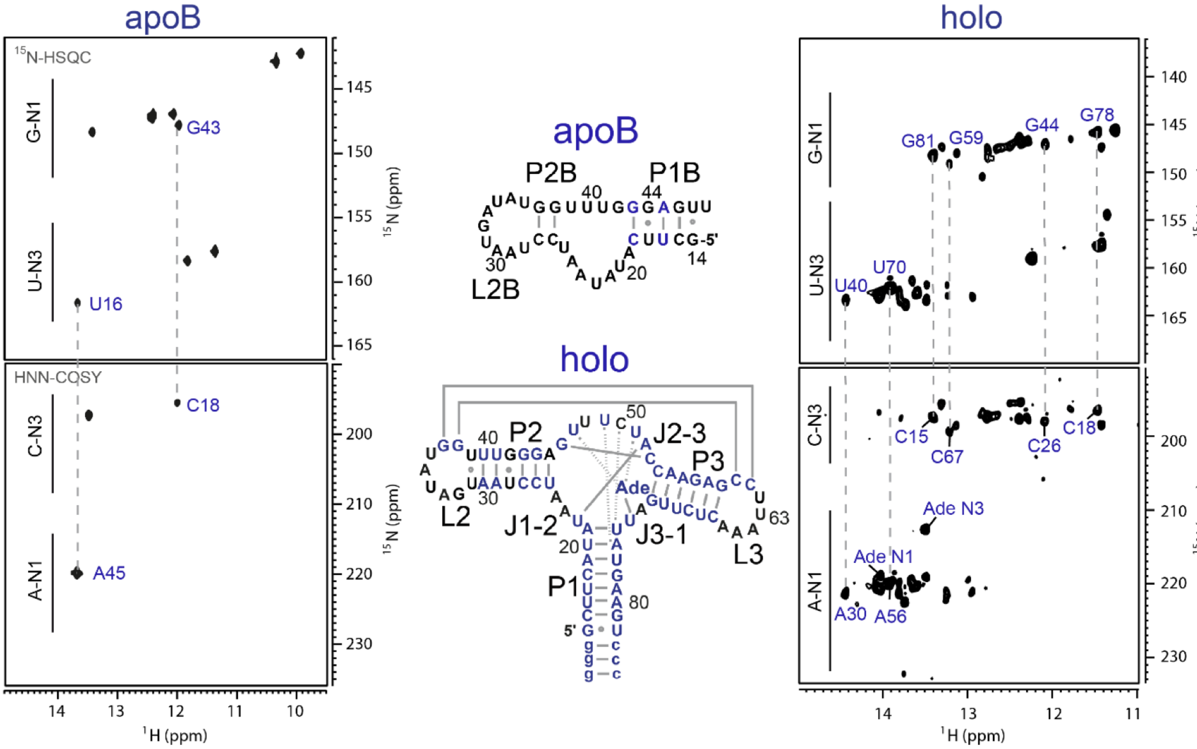


Fig. 9. 15N-HSQC and HNN-COSY spectra of the 35mer RNA and HNN-COSY spectra of the ASW aptamer domain. The HNN-COSY spectrum relates the 15N chemical shifts of the hydrogen-bond donor with that of the acceptor. The base pairs identified by HNN-COSY experiments are highlighted in blue. In apoB, two base pairs are highlighted (U16-A45 and G43-C18) to demonstrate the base pair assignment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

adjacent to a small reference lane containing formamide loading buffer with 0.01% [w/v] bromophenol blue and 0.01% [w/v] xylene cyanol. The gel was run in TBE buffer for 5 h–8 h at a constant power of 50 W under cooling with the integrated ventilator of the TVS1000 sequencer. RNA bands were visualized at the long edge of the gel by UV-shadowing (254 nm). The bands were excised into a 20 mL syringe avoiding the illuminated region. The syringe was filled with 0.3 M sodium acetate (pH 5.5) and eluted for 5 min. The gel was granulated by pushing the gel slices through the syringe and subsequently frozen at −80 °C for 15 min in order to facilitate the elution of the RNA. The RNA was eluted for 16 h at 4 °C under slow agitation. Then, the RNA was further eluted by fixing the tubes on top of an Eppendorf thermoshaker (700 rpm) for 1 h. The combined eluates were filtered through a sterile Corning® bottle top vacuum filter system with 0.2 µm pore size (Sigma-Aldrich), diluted with 2.3 vol abs. ethanol and cooled at −20 °C for 4 h to precipitate the RNA. The RNA was pelleted by centrifugation (10,000 g, 4 °C, 1 h). The pellet was reconstituted in water and then precipitated by addition of 5 vol 2% [w/v] lithium perchlorate in acetone and

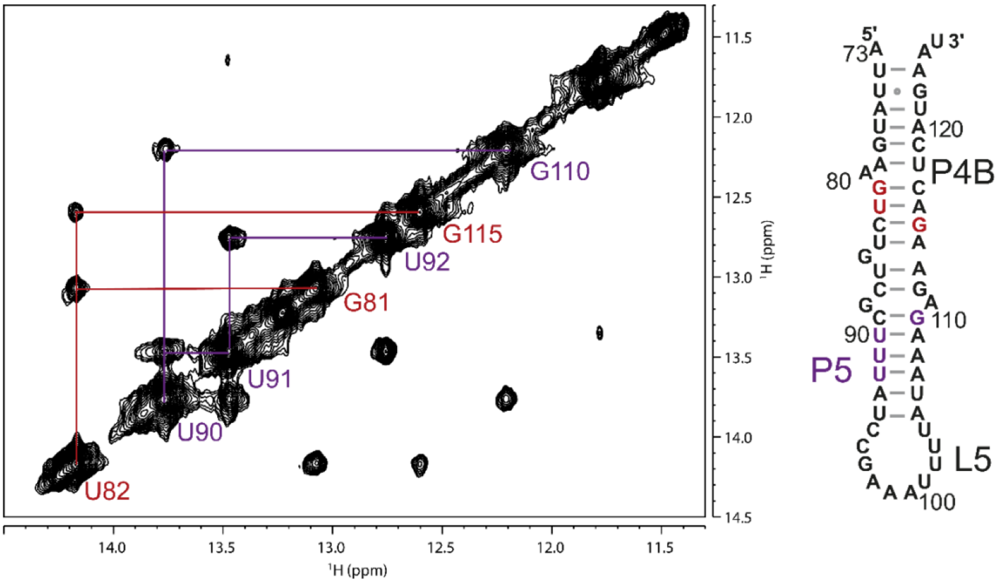


Fig. 10. [1H,1H]-NOESY spectrum of the 53mer ASW RNA. Two imino walks are highlighted in red (G81-U82-G115) and purple (G110-U90-U91-U92) for the P4B and P5 helix. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cooling at −20 °C for 4 h. The RNA was pelleted by centrifugation (10,000 g, 4 °C, 1 h), air-dried at room temperature for 15 min and reconstituted in 1 mL water. The lithium perchlorate precipitation was repeated to minimize contamination by low-molecular weight acrylamide. The RNA was desalted by changing the solvent into pure water via centrifugal concentrators. The solvent exchange was performed until a constant UV-absorbance ratio A_{260}/A_{200} is reached. After purification, all *in vitro* transcribed RNAs are folded by denaturation at 95 °C for 5 min following rapid 10-fold dilution ice cold water and incubation on ice for 1 h. The folded RNAs were then buffer exchanged into NMR buffer by repeated dilution cycles in centrifugal concentrators. The RNAs were then concentrated to concentrations between 0.1 mM and 0.5 mM and stored at 4 °C. The RNA concentrations was determined by UV/vis spectroscopy. Homogenous folding of the RNAs was verified by native polyacrylamide gel electrophoresis.

2.3.3. NMR sample preparation

All NMR samples were prepared in NMR buffer (25 mM K_2HPO_4/KH_2PO_4, 50 mM KCl, 5 mM magnesium chloride, pH 6.2) with 90% H_2O/10% D_2O and 100 µM DSS as internal ^{1}H chemical shift standard. For the deuterium lock, D_2O was added to each sample. The Shigemi tube was incubated with 0.1% Diethyl dicarbonate (DEPC) over night to inactivate RNases, washed with water, dried and filled with 280 µL of the RNA sample.

2.4. NMR methods

2.4.1. 1D ^{1}H and 2D ^{1}H, ^{15}N-BEST-TROSY experiments

For small to intermediated sized RNAs mapping of base paired nucleotides was achieved in 1D-^{1}H experiments. The jump-return-echo experiment [23] was applied as it achieves the best S/N ratio for the exchangeable imino protons in conjunction with an optimal water suppression. The excitation maximum was set to 11.5 ppm as this allows detection of imino signals with optimal sensitivity. The experiment was performed utilizing the determined hard pulses, a smoothed squared gradient of 1 ms length at 24 G/cm, an acquisition time of 200 ms at a ^{1}H spectral width of 25 ppm, with a GARP4 decoupling for ^{15}N applied at 153.5 ppm transmitter frequency and with a relaxation delay of 1 s. The number of transients was set to achieve the desired S/N value; the receiver should be set to maximum gain as water suppression is very efficient. The resulting FID was multiplied with a shifted sine bell function, Fourier transformed and phase corrected that all signal resonating downfield of the HDO-frequency have positive absorptive lines.

BEST-TROSY (Band-selective Excitation Short-Transient – Transverse relaxation-optimized spectroscopy) experiments were recorded to observe ^{1}H, ^{15}N correlations of ^{15}N-isotope labeled RNA. The ^{15}N-correlation experiments were recorded with a nitrogen carrier frequency of 153 ppm, resembling the intermediate resonance frequency between guanine N1 and uracil N3. Spectra were recorded with a spectral width of 25 ppm in the direct dimension (^{1}H) and 30 ppm in the indirect dimension (^{15}N). TROSY experiments were recorded with up to 512 points. The selective ^{1}H pulses were centered at 12.3 ppm and applied with the following shapes and field strengths: polychromatic PC9 [24] with a field strength of 6.4 kHz (as 90° pulse in the proton to nitrogen transfer), REBURP [25] with a field strength of 8 kHz (as 180° pulse), E-BURP2tr and E-BURP2 with a field strength of 5.6 kHz (as 90° pulse in the nitrogen to proton back-transfer). All selective ^{15}N pulses were centered at 153.5 ppm and applied with the following shapes: bipolar BIP-720-50-20 [5] with a field strength of 21.6 kHz and EBURP [25] with a field strength of 6.8 kHz. The HN-transfer delays were set to $\Delta = 1/4 * {}^1J(NH) = 2.7$ ms. During nitrogen chemical evolution period adjacent carbons were decoupled by selective ^{13}C adiabatic composite smoothed chirp pulse [26], centered at 110 ppm with a field strength of 89 kHz.

The stability of RNA molecules can best be investigated by following temperature-induced changes in the 1D imino and 2D BEST-TROSY spectra.

2.4.2. HNN-COSY experiment

HNN-COSY spectra can be acquired on various different temperatures; nevertheless, most often it is advantageous to record the experiment at lower temperatures (283 K to 293 K). A standard pulse program was applied for the 2D J_{NN}HNN-COSY experiment employing water flip-back pulses and WATERGATE for water suppression. The following HNN-COSY parameters were used: In the direct dimension (^{1}H), 2048 data points were recorded at a spectral width of 22 ppm. For the indirect dimension (^{15}N), 128 data points were collected at a spectral width of 120 ppm with 1056 scans. For the NN-COSY transfer, a nitrogen carrier frequency of 183 ppm was used, resembling the intermediate resonance frequency between acceptor (A and C) and donor (G and U). Initially 2D test spectra were recorded with high sensitivity and a short N-N transfer period of 4 ms. If the processed spectra show peaks with good line shape, phase behaviours and good signal-to-noise ratio a new spectrum was recorded with N-N transfer period of 10 to 15 ms. For the INEPT-transfer, a nitrogen carrier frequency of 153 ppm was used.

The magnitude of the $^{2h}J_{NN}$ coupling from diagonal- and cross-peaks was determined by using the formula: $|^{2h}J_{NN}| = \arctan \frac{\left(-\frac{I_A}{I_D}\right)^{1/2}}{2\pi\Delta}$ [27].

2.4.3. HNCO experiment

HNCO experiment provided indirect evidence of hydrogen bonds. By evaluation of the chemical shift changes of the quaternary carbons, H-bonded and free carbonyls can be distinguished. In the HNCO experiment the imino protons of the attached nitrogen were correlated with the adjacent carbons, as shown for guanine and uracil in Fig. 2 [9].

HNCO spectra were recorded utilising a standard triple-resonance pulse program for the out-and back-correlation of H-N-C. The experiment was usually applied as a 2D experiment (^{13}CO, ^{1}H) but for larger RNAs it can also be recorded in its original 3D format with evolution of chemical shifts of CO during t_1, of N during t_2 and direct detection of H in t_3. It was recorded with Echo/Antiecho gradient selection in t2, where it also applied a constant time period. Magnetization transfer was achieved via INEPT transfer steps. The ^{1}H transmitter frequency was set to the water resonance frequency. For HNC-correlations starting from the imino-proton, the transmitter frequency for ^{15}N was set to 154 ppm and for ^{13}C to 158 ppm. For correlations starting from amino protons in cytosine the ^{15}N transmitter frequency was set at 80 ppm. Field strengths of 27.5 kHz and 7.5 kHz were applied for hard pulses of ^{1}H and ^{13}C. The GARP4 (globally optimized alternating phase rectangular pulses) sequence was required for pulse decoupling with a field strength of 1.3 kHz. Following frequency ranges can be applied for selective ^{13}C pulses: 90° Q5 pulse with 14 kHz and a 180° Q3 pulse with 12 kHz.

2.4.4. [^{1}H,^{1}H]-NOESY

The [^{1}H, ^{1}H]-NOESY spectra were recorded using a jump return echo water suppression scheme [23]. The proton carrier frequency was set to the resonance frequency of the solvent (4.7 ppm) in the direct dimension and switched to the frequency of 8.5 ppm in the indirect dimension. The spectra were recorded with a spectral width of 24 ppm in the direct dimension and 12 ppm in the indirect dimension. The 2D [^{1}H, ^{1}H]-NOESY spectra were recorded with different mixing times (between 40 ms and 250 ms depending on the size of the RNA molecule and the purpose of the experiment, a good starting value is 150 ms).

2.4.5. Diffusion ordered spectroscopy (DOSY)

DOSY experiments to evaluate the overall shape of the RNA were measured in buffered D_2O and in 5 mm diameter Shigemi tubes. The reduced sample height inhibits the onset of convection [28]. Since solution conditions (ionic strength, pH, temperature and buffer

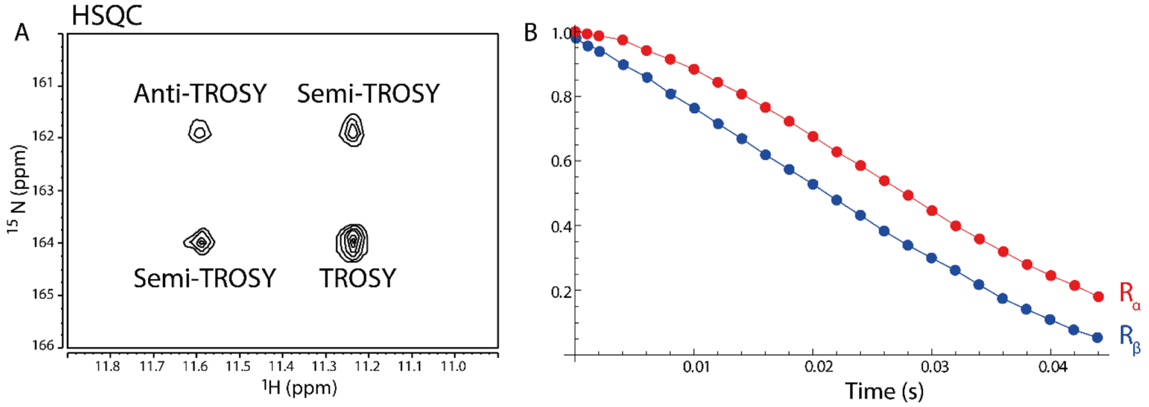


Fig. 11. Schematic representation of the selection of TROSY and semi-TROSY peaks. A series of experiments with increasing relaxation periods is recorded to obtain the relaxation rates R_α and R_β.

concentrations) have a great impact on DOSY measurements, the same buffer conditions were used for all of the different NMR samples [28]. A reference substance, e.g. dioxane, was added to all samples.

In a 1D version of the experiment the duration of the magnetic field pulse gradients (δ) as well as the diffusion time (Δ) was optimized in order to obtain ~5% residual signal with maximum gradient strength. In the actual 2D experiment, the pulse gradients were incremented from 2% to 95% of the maximum gradient strength in a linear ramp.

2.4.6. ^{1}H-^{15}N BEST-TRACT

The relaxation rates of the α- and the β-spin state (Fig. 11) were measured in two different experiments that select either the TROSY or SEMI-TROSY component of the NH-doublet. The two experiments are different in their phase cycle to select for the slowly relaxing α-spin state and the more rapidly relaxing β-spin state of the ^{15}N nucleus. A series of spectra with increasing relaxation periods (Fig. 12) were recorded. The signal intensity of the TROSY and semi-TROSY peak decreases with increasing relaxation delay τ_i. The relaxation rates R_α and R_β were obtained by fitting both of the signal decays. For the analysis of ASW the conventional [^{1}H, ^{15}N]-TRACT pulse sequence was implemented as a BEST-version, which is required to gain sensitivity and resolution (Fig. 3).

The integrals of the HN-correlation peaks along the relaxation time dimension were fitted by the following equation:

$$R = A\cos[\pi J(t_p + (\tau_i/2))]e^{-R_{\alpha/\beta}\tau_i}$$

with τ_i as the relaxation delay, t_p as the duration of the ^{13}C pulse and $R_{\alpha/\beta}$ the relaxation rates of the TROSY and semi-TROSY peak.

The rotational correlation time τ_C were extracted from the difference of the relaxation rates η_{xy}, by solving the following equation for τ_c:

$$\eta_{xy} = R_\beta - R_\alpha = 2p\delta_N\left(4\frac{2\tau_C}{5} + 3\frac{2\tau_C}{5(1+(\tau_C\omega)^2)}\right)(3cos^2\theta - 1)$$

with

$$p = \left(\frac{\mu_0}{4\pi}\right)\frac{\gamma_I\gamma_S\hbar}{r^3_{NH}\sqrt{8}}$$

and

$$\delta_N = \left(\frac{1}{\sqrt{6}}\frac{1}{\sqrt{3}}\right)\gamma_N B_0\Delta\sigma_N$$

where μ_0 is the vacuum permeability, γ_H and γ_N are the gyromagnetic ratios of ^{1}H and ^{15}N, ħ the Planck constant divided by 2π, r_{HN} represents the ^{1}H-^{15}N internuclear distance, B_0 as the magnetic field strength. $\Delta\sigma_N$ describes the difference of the two principal components of the axially symmetric ^{15}N chemical shift tensor.

2.4.7. Time resolved NMR initiated by in situ mixing

For each experiment 300 µL of the riboswitch sample, with a concentration of 300 µM, were initially placed in a conventional Shigemi NMR tube. This tube thus also served as the mixing chamber. For folding of the ASW a selectively ^{15}N-uridine-labelled RNA were used, the temperature was set to 20 °C and the Mg^{2+} concentration was set to $c_{Mg2+} = 2$ mM.

A coaxial insert containing 40 µL of the solution (5 mM Adenine to prepare an [RNA]:[ligand] ratio of 1:2) to be injected was placed so that the tip of the insert contacts the surface of the sample solution (Fig. 4). Spacers allowed the axial position of the insert and reduced its movement in the actual mixing process.

The insert was connected via a water-filled connection hose to a pneumatic punch outside the spectrometer, which was coupled to the

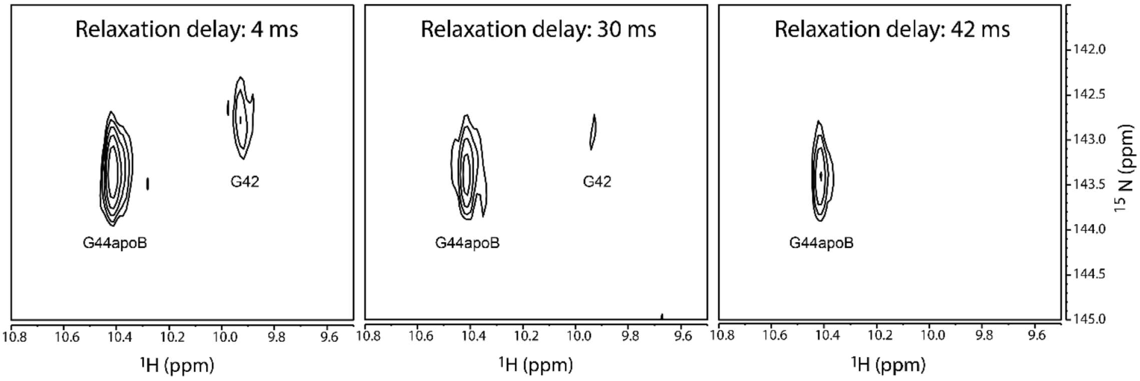


Fig. 12. [^{1}H,^{15}N]-BEST-TRACT spectra of the 127mer RNA with increasing relaxation delay τ_i.

console of the spectrometer. This device allowed the initiation of the mixing process via the pulse program utilizing a TTL-connection between the spectrometer and the electronic control of the pneumatic punch. This setup ensured a time-defined mixing tuned with the NMR experiment.

In order to avoid premature mixing, the solution for injection was separated by small air bubbles from the sample solution as well as from the water in the connection tube. The injection solution was finally injected at a pressure of approx. 2 bar. The time resolved experiments for the wild-type ASW were recorded as a pseudo-2D experiment consisting of ^{15}N-filtered/edited jump-and-return experiments. To achieve highest S/N per unit measurement time, Ernst-angle excitation [30] was applied. The experiment was conducted in a tuned fashion with respect to the injection of the ligand. First 32 data points (with eight scans per time points resulting in a resolution of 2.1 s) were recorded before injection of adenine. Then the injection was triggered (t_{dead} = d1 + t_{hom} = 0.2 s + 0.05 s) followed by another 128 data points.

The selection of the signals of the ^{15}N-bound imino-protons (U-residues) and those from ^{14}N-bound imino-protons (G-residues) was achieved using an X-filter pulse sequence element [31]. It consisted of two ^{15}N-90°-pulses, where the phase of the first one is incremented in an interleaved fashion by 180° in the subsequent 1D-spectra. This resulted in inverted signs of the different spin systems in the even data points, whereas both had the same sign in the odd data points. To extract the ^{1}H, ^{14}N-filtered kinetic out of the recorded spectrum odd and even data points had to be summed, while for the ^{1}H, ^{15}N-filtered kinetic the difference between odd and even data points had to be formed. The rate constants for folding of the complexes were determined by calculating the normalized integrals over the full width at half height of the imino-proton signal. In order to extract site-specific refolding rates only resonances that are well resolved were used. The obtained kinetic traces were then subsequently fitted with an appropriate function describing the anticipated underlying kinetic mechanism.

2.4.8. ZZ-exchange

As the exchange kinetics must be faster than the relaxation rates of the magnetizations or coherences present during the mixing time of the pulse sequence, a matching experiment had to be chosen at first. Early developed experiments used $2I_zS_z$ longitudinal two-spin order or S_z magnetization [32]. In most cases pulse sequence were appropriate that use S_z-magnetization during the mixing time and frequency labeling in an HSQC-manner [33,34]. If resonance overlap between signals of the two interconverting states hampered the analysis a heteronuclear zero quantum coherence Nzz-exchange experiment that resolves the resonance overlap could be applied [35]. For the analysis of the ASW's conformational exchange in the apo-state a pseudo 3D ^{1}H-^{15}N correlation experiment was used, that utilized $^{15}N_z$-magnetisation during mixing time. The correlation between proton and nitrogen was achieved via double INEPT transfer steps. Evolution of nitrogen chemical shifts was recorded in a semi-constant-time period. The experiment used flip-back pulses and Watergate-sequence for water suppression. Transmitter frequencies and spectral widths and resolution in the ^{1}H and ^{15}N dimensions were set to the same values as in the ^{1}H-^{15}N BEST-TROSY experiments. The relaxation delay was set to 2 s. Mixing times for the evolution ranged between 3 μs and 800 ms, 12 to 16 different time points were applied in the final pseudo-3D experiment. The experiment was multiplied with squared sine bell shaped window functions in both frequency dimensions, Fourier transformed, and phase corrected.

Integrals of diagonal- and cross-peaks were taken along the mixing time dimension. If resolution allowed integration of all four peaks was performed. Subsequently, these were simultaneously fitted to the appropriate functions [33] yielding the rates of interconversion.

3. Results

3.1. Construct design for NMR and smFRET

In order to obtain results that are comparable between both NMR and smFRET, several prerequisites have to be met. The length of the constructs of course has to be identical, and fluorophore placement should not interfere with RNA folding. The latter can most easily be achieved by placing the dye attachment sites into helical regions; for loop labeling, possible effects of the dye positions have to be taken into account. Furthermore, the dyes have to be placed in regions that supposedly alter their distance undergoing conformational dynamics. In the case of the adenine-sensing riboswitch from *Vibrio vulnificus*, dye positions were chosen for the L2, L3, and P5, respectively (Fig. 5). Whereas labeling of loops L2 and L3 had been used previously [36], we additionally placed one label into the P5 separating the aptamer domain from the expression platform.

3.2. FRET results

Characterization of the construct with both labels placed in the loop regions revealed that a significant number of molecules remain in a low-FRET state, corresponding to a conformation where the loops are open, regardless whether the adenine ligand is present (Fig. 6). Upon addition of ligand, the portion of molecules that adopt a closed conformation (as evidenced by an increase in FRET efficiency from ~0.2 to 0.9) about doubles. Under the conditions tested, this shows that our construct is indeed functional in that it shows a clear response to ligand binding.

Initial investigation of all labelling site combinations also revealed that the P5 helix does not come into closer contact with the aptamer domain. This is of special interest, as the mechanism of functional crosstalk between the aptamer domain and the expression platform could in principle also require structural interactions between these two domains. Our results however clearly demonstrate that this is not the case.

For the loop-labeled construct (L2/L3 in Fig. 6), additional time-dependent FRET traces were recorded. This served two purposes: first, the single-step photobleaching event (indicated by an arrow in Fig. 7A) verifies that we are observing single molecules. Second, as the dyes over time show anticorrelated intensity profiles, the riboswitch constructs show intramolecular dynamics by adopting two states with distinct FRET efficiencies. Hidden Markov modeling and statistical analysis are shown in Fig. 7B, and further demonstrate that there are two major populations of molecules: e which remains constant at low FRET, and one which interconverts between low and high FRET. Further kinetic analysis also shows first-order kinetics for both docking and undocking, and allows for determining these rates (Fig. 7C).

3.3. NMR results

3.3.1. Mapping base paired nucleotides

3.3.1.1. 1D ^{1}H and 2D [^{1}H,^{15}N]-BEST-TROSY experiments. The measurement of RNA molecules by NMR spectroscopy gives information about the standard and non-standard Watson-Crick-type base pairs and allows the determination of secondary structure elements (1D spectrum in Fig. 8). A 1D ^{1}H spectrum contains valuable information about the base pairing pattern, especially the imino proton resonances of guanines (H1) and uracils (H3) between 9 ppm and 15 ppm. The resonances of standard Watson-Crick base tend to be found in the region of 12 ppm to 15 ppm. In contrast, non-standard Watson-Crick base pairs like U-U and G-U wobble base pairs are often shifted upfield between 9 ppm and 12 ppm. The signals are only observable when the specific nucleotides are engaged in stable base pairing interactions protected from solvent exchange. It is possible to count the number of imino proton resonances, which correspond to the number of base pairs. The BEST-TROSY spectrum of the 127mer ASW RNA shows the assigned imino proton resonances. In the spectrum the characteristic chemical shifts for G-C Watson-Crick base pairs are highlighted in red and for A-U base pairs in blue. In addition, the wobble base pairs of G-U are shown in green and U-U in purple (Fig. 8).

3.3.2. Characterization of base pair types and their spatial arrangement

3.3.2.1. HNN-COSY experiment. The HNN-COSY (correlation spectroscopy) experiment can be used to elucidate the hydrogen bond formation and the base pairing by space scalar $^{2h}J_{NN}$ coupling constant across the N-H···N-type hydrogen bond between the hydrogen bond and of ^{15}N donor nuclei and ^{15}N acceptor nuclei in RNA and DNA. [9] HNN-COSY experiments can be performed to investigate the aptamer and full-length riboswitch as shown in Fig. 9. On the left-hand side of Fig. 9, ^{15}N-HSQC and HNN-COSY spectra of a 35mer RNA of apoB ASW are shown. The HNN-COSY spectrum relates to the ^{15}N chemical shifts of ^{15}N-HSQC, the assigned base pairs (U16-A45 and G43-C18) are highlighted in blue and indicate the Watson-Crick G-C and A-U base pairing between the hydrogen-bond donor with that of the acceptor. On the right-hand side, the HNN-COSY spectrum for the ligand-bound holo-state of the ASW is shown. All base pairs identified by HNN-COSY experiments are highlighted in blue in the secondary structure of the holo-state of ASW. Besides the canonical base pairs of the helices identified by HNN-COSY experiments, two canonical base pairs from the core-region (G46-C53, U22-A52) and three $^{2h}J_{NN}$ couplings of loop-loop interactions (G37-C60, G38-61 and U34-A65) were detected. Further two signals of the ^{15}N-labeled ligand adenine (AdeN1 and AdeN3) were identified (Fig. 9).

3.3.2.2. [^1H,^1H]-NOESY. The NOESY (Nuclear Overhauser and Exchange Spectroscopy) provides through-space information of protons. This information can be utilized to determine the secondary and tertiary structure of RNAs. The exemplary spectrum of a structural module of the ASW (Fig. 10) shows several cross peaks, each of which corresponds to a pair of protons which are separated by less than 5 Å. Two imino walks are highlighted in red (G81-U82-G115) and purple (G110-U90-U91-U92) for the P4B and P5 helix of the ASW RNA.

The application of the above described experiments on the full length ASW (Fig. 5) and on six additional fragment structural reference modules (5′-P1-P2-P3-3′, 5′-P1-P2-3′, 5′-P4-P5-3′, 5′-P1-P2-P3-P5-3′, 5′-P5-3′, 5′- P3-3′) lead to a full structural description of all stable and persistent secondary structures adopted by the riboswitch in the apo- and holo-state [4,13]. Further, application of these experiments to the isolated aptamer of this particular riboswitch yielded in a detailed description of the ligand recognition [37].

Also for other riboswitch systems these methodology was successfully applied. In case of the I-A type 2′dG binding riboswitch from *Mesoplasma fluorum* transcriptional intermediates were structurally characterized and showed that the regulation of transcription is actually achieved by metastable RNA conformations [38]. In combination with further advanced labelling schemes these experiments led to the description of the structural dynamics needed for ligand recognition and gene-regulatory function in other multi-state riboswitches such as the SAM-II riboswitch [39] or the pseudoknot forming preQ1 riboswitch [40]. Penultimately, this set of experiments is also the basis for modelling the three-dimensional structure and dynamics of small riboswitch systems such as the Fluoride riboswitch from *Bacillus cereus* [41].

3.3.3. Characterization of global structural parameters

3.3.3.1. Diffusion ordered spectroscopy (DOSY). DOSY is a powerful technique to gain information about diffusion rates and molecular sizes with the help of pulsed magnetic field gradients [42]. DOSY separates NMR signals according to their diffusion coefficient. It is widely used for component analysis of complex mixtures [43].

The method relies on two bipolar gradient pulses [44] separated by a centered 180° pulse. The first gradient pulse dephases the transverse magnetization in a spatially dependent manner along the z-axis which is then refocused by the second bipolar gradient pulse in the opposite direction. If the molecule moves along the z-axis during the time between both gradient pulses, the magnetization will not refocus completely. The faster the molecule moves, the larger the attenuation of the resonance will be.

3.3.3.2. [^1H-^{15}N]-BEST-TRACT. As DOSY experiments lead to a description of the translation diffusion of molecules, TRACT (TROSY for rotational correlation times) is a novel approach to determine rotational correlation times τ_c. It is based on TROSY effect. In TROSY experiments (transverse relaxation-optimized spectroscopy) the interference of dipole-dipole (DD) and chemical shift anisotropy relaxation leads to a reduction of effective transverse relaxation. TRACT uses the cross-correlation of chemical shift anisotropy and dipole-dipole relaxation to determine τ_c [45]. Typically, spectra are measured with increasing relaxation delays (Fig. 12) and fitted to obtain the relaxation rates R_α and R_β (Fig. 11). As outlined above, from these differential relaxation rates the rotational diffusion constant can be extracted.

The diffusion behavior of an RNA is dependent on its overall shape. Therefore, all conformational changes will inevitably lead to a modulation of the diffusion of the RNA. Exemplary, the diffusion behavior of the full length ASW is slightly modulated upon ligand binding. In the apo-state diffusion of ASW is characterized by a diffusion coefficient of $D_{ASW}{}^{apo} = 4.3 \times 10^{-7}\,cm^2/s$. A compaction of the riboswitch in the holo-state can be inferred as the diffusion coefficient is increased to $D_{ASW}{}^{holo} = 4.8 \times 10^{-7}\,cm^2/s$. Similarly, for other riboswitches also compaction of the overall structure upon ligand binding is revealed by other biophysical or biochemical techniques such as size exclusion chromatography (SEC) [46] and small angle X-ray scattering (SAXS) [47,48].

3.3.4. Characterization of secondary structure transitions

3.3.4.1. Time resolved NMR initiated by in situ mixing. Non-equilibrium RNA folding kinetics characterized by rates lower than $1\,s^{-1}$ can be analyzed by time resolved 1D- or 2D-NMR spectroscopy. Initiation of the reaction can be achieved by either temperature jump, in situ laser illumination of photo-caged reaction inductors or by rapid mixing of reactants [10]. The exemplified ligand induced refolding reactions from the apo to the holo-state of adenine sensing riboswitch was induced by mixing the RNA with adenine. A special mixing apparatus originally developed by Hore and co-workers was used. [49] It allows two liquids to be mixed together directly in the NMR spectrometer. The dead time for complete mixing is 50 ms. In Fig. 4, the mixing chamber which is placed in the active volume of the NMR spectrometer's probe, is schematically shown. With this approach the tertiary complex formation of was characterized and it could be shown that the structure of the holo-state is adopted with rates of $k_F = (4.0 \pm 0.1) \times 10^{-2}\,s^{-1}$. Therefore, folding into the holo-state conformation occurs on the same timescale in the full length ASW as in the isolated aptamer domain [50]. As NMR has the intrinsic power of atomic resolution, with these experiments differential structural kinetics reporting on the detailed folding trajectory of the riboswitch can be measured [51]. This unique feature of time-resolved NMR is lately challenged by real-time crystallography using X-ray free-electron lasers [52].

3.3.4.2. ZZ-exchange. Equilibrium RNA folding kinetics characterized by rates between approx. $1\,s^{-1}$ and $100\,s^{-1}$ can be analyzed by ^{15}N-ZZ-exchange experiments [33]. In general, these experiments are appropriate for the analysis of equilibrium conformational changes that occur in slow chemical exchange (Fig. 13). To successfully apply these experiments, it is necessary that resolved resonances are observed for the individual chemical states. For the adenine sensing riboswitch

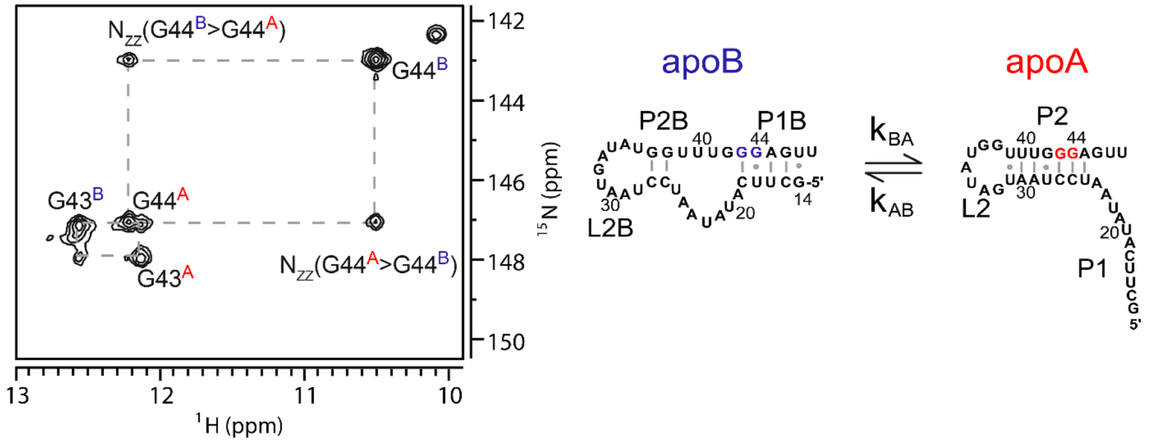


Fig. 13. ^{15}Nzz exchange spectrum of the 35mer RNA, left panel 2D ^{1}H-^{15}N spectral plane (300 ms) of a pseudo 3D ^{15}N-ZZ-exchange experiment highlighting the G43/G44 peak pairs, right panel interconverting conformations of the 35 nt long ASW temperature regulation element.

two distinct conformational states can be observed in the absence of ligand. Their interconversion, that relies on the bistable behavior of the first 35 nt of the full-length riboswitch, can be characterized by ZZ-exchange experiments. The rates of interconversion between both apo-states were $k_{AB} = 0.91\,s^{-1}$ and $k_{BA} = 0.40\,s^{-1}$ at 25 °C. In line with RNA transition-state stabilization, increasing the Mg^{2+}concentrations accelerated these rates up to $k_{AB} = 1.99\,s^{-1}$ and $k_{BA} = 0.68\,s^{-1}$. The finding that the 5′-terminal first 35 nucleotides of the riboswitch were responsible for the conformational heterogeneity in the apo state was corroborated by analysis of a 35-nucleotide RNA fragment (G14-U49) that showed the same conformational heterogeneity with identical base pairing in the two underlying secondary structure elements (Fig. 13).

4. Conclusion

Fig. 14 shows how the information obtained from both single molecule FRET and NMR can be combined into a picture that is by far more detailed than obtainable from either technique alone: The detailed structural information from NMR includes e.g. information on the state of different secondary structures by means of their basepairing patterns, whereas smFRET yields additional information on Mg2+-dependent states, refolding kinetics, and heterogeneity of the observed population.

To be more general, the experiments such as the ones mentioned above result in information about conformational dynamics of the riboswitch at different timescales: whereas NMR covers a wide range of available time regimen [53], the standard lower limit for smFRET of immobilized molecules is usually not significantly shorter than 100 ms. Both methods however easily afford to monitor molecules over minutes or even hours [2]. The information gain by combining these two methods therefore is twofold: The combination of knowledge on base-pairing states of nucleotides combined with the long-range distance information allows to assign i.e. relative helical arrangements. In addition, the coexistence of two conformations can be further characterized, either confirming heterogeneous populations of dynamic and static molecules, or timescales of the interconversions between the two conformational states. This is particularly true for the analysis of riboswitches, where a small molecule ligand can be easily included into both smFRET and NMR measurements. Thus, it is not despite, but rather because of the apparent technical differences between those two techniques that with these additional insights, the combination of NMR

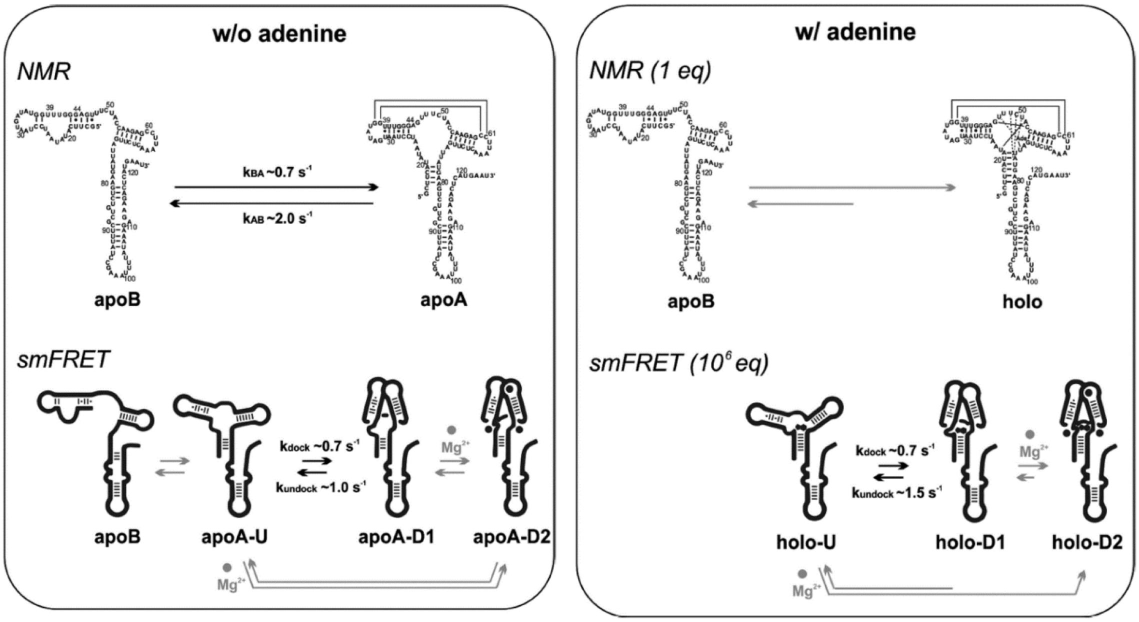


Fig. 14. Model combining information from smFRET and NMR for conformational dynamics of the adenine responsive riboswitch both in the presence and absence of the ligand adenine (adapted from [13].

and smFRET allows for significantly deeper insights into the structure and dynamics of riboswitches than either technique alone.

Acknowledgements

The authors acknowledge funding by the State of Hesse (BMRZ), Deutsche Forschungsgemeinschaft, CRC “Molecular mechanisms of RNA-mediated regulation”, GRK-1986 CLiC, and SPP1157. H.S. and M.H. are members of the Cluster of Excellence “Macromolecular complexes in action”.

References

[1] D. Kowerko, et al., Cation-induced kinetic heterogeneity of the intron-exon recognition in single group II introns, Proc. Natl. Acad. Sci. U.S.A. 112 (11) (2015) 3403–3408.
[2] M. Hengesbach, et al., Single-molecule FRET reveals the folding dynamics of the human telomerase RNA pseudoknot domain, Angew. Chem. Int. Ed. Engl. 51 (24) (2012) 5876–5879.
[3] R. Lamichhane, et al., RNA looping by PTB: Evidence using FRET and NMR spectroscopy for a role in splicing repression, Proc. Natl. Acad. Sci. U.S.A. 107 (9) (2010) 4105–4110.
[4] A. Reining, et al., Three-state mechanism couples ligand and temperature sensing in riboswitches, Nature 499 (7458) (2013) 355–359.
[5] S. Warhaut, et al., Ligand-modulated folding of the full-length adenine riboswitch probed by NMR and single-molecule FRET spectroscopy, Nucl. Acids Res. 45 (9) (2017) 5512–5522.
[6] M. Helm, A.Y. Kobitski, G.U. Nienhaus, Single-molecule Forster resonance energy transfer studies of RNA structure, dynamics and function, Biophys. Rev. 1 (4) (2009) 161.
[7] R. Roy, S. Hohng, T. Ha, A practical guide to single-molecule FRET, Nat. Methods 5 (6) (2008) 507–516.
[8] M. Hengesbach, et al., RNA intramolecular dynamics by single-molecule FRET, Curr. Protoc. Nucl. Acid Chem. (2008) 11–12.
[9] B. Furtig, et al., NMR spectroscopy of RNA, Chembiochem 4 (10) (2003) 936–962.
[10] B. Furtig, et al., Time-resolved NMR studies of RNA folding, Biopolymers 86 (5–6) (2007) 360–383.
[11] A.G. Palmer 3rd, Chemical exchange in biomacromolecules: past, present, and future, J. Magn. Reson. 241 (2014) 3–17.
[12] I. Bertini, K.S. McGreevy, G. Parigi, NMR of Biomolecules: Towards Mechanistic Systems Biology, Wiley-VCH, Weinheim, Germany, 2012, p. 612 xxviii.
[13] S. Warhaut, et al., Ligand-modulated folding of the full-length adenine riboswitch probed by NMR and single-molecule FRET spectroscopy, Nucl. Acids Res. 45 (2017) 5512–5522.
[14] P.R. Selvin, T. Ha, Single-molecule techniques: a laboratory manual (2008) 507.
[15] J. Larson, et al., Design and construction of a multiwavelength, micromirror total internal reflectance fluorescence microscope, Nat. Protocols 9 (2014) 2317–2328.
[16] B. Paudel, D. Rueda, RNA folding dynamics using laser-assisted single-molecule refolding, Methods Mol. Biol. (Clifton, N.J.) 1086 (2014) 289–307.
[17] S. Warhaut, Biophysical studies of the translation-regulating add adenine riboswitch from Vibrio vulnificus, 2017.
[18] M. Blanco, N.G. Walter, Analysis of complex single-molecule FRET time trajectories, Methods Enzymol 472 (2010) 153–178.
[19] S.A. McKinney, C. Joo, T. Ha, Analysis of single-molecule FRET trajectories using hidden Markov modeling, Biophys. J. 91 (5) (2006) 1941–1951.
[20] R.K. Hartmann, A.J. Andrews, Handbook of RNA biochemistry, 2015, Available from: http://site.ebrary.com/id/11041410.
[21] T.K. Stage-Zimmermann, O.C. Uhlenbeck, Hammerhead ribozyme kinetics, RNA 4 (8) (1998) 875–889.
[22] J.M. Avis, G.L. Conn, S.C. Walker, Cis-acting ribozymes for the production of RNA in vitro transcripts with defined 5′ and 3′ ends, Methods Mol. Biol. 941 (2012) 83–98.
[23] V. Sklenář, A. Bax, Spin-echo water suppression for the generation of pure-phase two-dimensional NMR spectra, J. Magn. Reson. (1969) 74 (1987) 469–479.
[24] E. Kupce, R. Freeman, Wideband excitation with plychromatic pulses, J. Magn. Reson. (1994) 268–273.
[25] H. Geen, R. Freeman, Band-selective radiofrequency pulses, J. Magn. Reson. (1969) 93 (1991) 93–141.
[26] R. Fu, G. Bodenhausen, Broadband decoupling in NMR with frequency-modulated ‘chirp’ pulses, Chem. Phys. Lett. 245 (1995) 415–420.
[27] A.J. Dingley, et al., Direct detection of N-H[…]N hydrogen bonds in biomolecules by NMR spectroscopy, Nat. Protoc. 3 (2) (2008) 242–248.
[28] S.H.S. Chan, et al., Increasing the sensitivity of NMR diffusion measurements by paramagnetic longitudinal relaxation enhancement, with application to ribosome-nascent chain complexes, J. Biomol. NMR 63 (2) (2015) 151–163.
[29] M.A. Smith, H. Hu, A.J. Shaka, Improved broadband inversion performance for NMR in liquids, J. Magn. Reson. 151 (2001) 269–283.
[30] R.R. Ernst, G. Bodenhausen, A. Wokaun, Principles of Nuclear Magnetic Resonance in One and Two Dimensions, Clarendon Press, Oxford, 2004.
[31] E. Woergoetter, G. Wagner, K. Wuethrich, Simplification of two-dimensional proton NMR spectra using an X-filter, J. Am. Chem. Soc. 108 (20) (1986) 6162–6167.
[32] G.T. Montelione, G. Wagner, 2D Chemical exchange NMR spectroscopy by proton-detected heteronuclear correlation, J. Am. Chem. Soc. 111 (8) (1989) 3096–3098.
[33] N.A. Farrow, et al., A heteronuclear correlation experiment for simultaneous determination of 15N longitudinal decay and chemical exchange rates of systems in slow equilibrium, J. Biomol. NMR 4 (5) (1994) 727–734.
[34] G. Wider, D. Neri, K. Wüthrich, Studies of slow conformational equilibria in macromolecules by exchange of heteronuclear longitudinal 2-spin-order in a 2D difference correlation experiment, J. Biomol. NMR 1 (1) (1991) 93–98.
[35] S.A. Robson, et al., A heteronuclear zero quantum coherence Nz-exchange experiment that resolves resonance overlap and its application to measure the rates of heme binding to the IsdC protein, J. Am. Chem. Soc. 132 (28) (2010) 9522–9523.
[36] J.F. Lemay, et al., Folding of the adenine riboswitch, Chem. Biol. 13 (8) (2006) 857–868.
[37] J. Noeske, et al., An intermolecular base triple as the basis of ligand specificity and affinity in the guanine- and adenine-sensing riboswitch RNAs, Proc. Natl. Acad. Sci. U.S.A. 102 (5) (2005) 1372–1377.
[38] C. Helmling, et al., NMR structural profiling of transcriptional intermediates reveals riboswitch regulation by metastable RNA conformations, J. Am. Chem. Soc. 139 (7) (2017) 2647–2656.
[39] B. Chen, R. LeBlanc, T.K. Dayie, SAM-II riboswitch samples at least two conformations in solution in the absence of ligand: implications for recognition, Angew. Chem. Int. Ed. Engl. 55 (8) (2016) 2724–2727.
[40] T. Santner, et al., Pseudoknot preorganization of the preQ1 class I riboswitch, J. Am. Chem. Soc. 134 (29) (2012) 11928–11931.
[41] B. Zhao, et al., An excited state underlies gene regulation of a transcriptional riboswitch, Nat. Chem. Biol. 13 (9) (2017) 968–974.
[42] C.S. Johnson, Diffusion ordered nuclear magnetic resonance spectroscopy: principles and applications, Prog. Nucl. Magn. Res. Spectrosc. 34 (3) (1999) 203–256.
[43] R. Evans, et al., Quantitative interpretation of diffusion-ordered NMR spectra: can we rationalize small molecule diffusion coefficients? Angew. Chem. Int. Ed. Engl. 52 (11) (2013) 3199–3202.
[44] J. Lapham, et al., Measurement of diffusion constants for nucleic acids by NMR, J. Biomol. NMR 10 (3) (1997) 255–262.
[45] D. Lee, et al., Effective rotational correlation times of proteins from NMR relaxation interference, J. Magn. Reson. 178 (1) (2006) 72–76.
[46] B. Chen, et al., Multiple conformations of SAM-II riboswitch detected with SAXS and NMR spectroscopy, Nucl. Acids Res. 40 (7) (2012) 3117–3130.
[47] S. Roy, et al., A magnesium-induced triplex pre-organizes the SAM-II riboswitch, PLoS Comput. Biol. 13 (3) (2017) e1005406.
[48] J. Zhang, C.P. Jones, A.R. Ferre-D'Amare, Global analysis of riboswitches by small-angle X-ray scattering and calorimetry, Biochim. Biophys. Acta 1839 (10) (2014) 1020–1029.
[49] K.H. Mok, et al., Rapid sample-mixing technique for transient NMR and photo-CIDNP spectroscopy: applications to real-time protein folding, J. Am. Chem. Soc. 125 (41) (2003) 12484–12492.
[50] M.K. Lee, et al., Real-time multidimensional NMR follows RNA folding with second resolution, Proc. Natl. Acad. Sci. U.S.A. 107 (20) (2010) 9192–9197.
[51] J. Buck, et al., Time-resolved NMR methods resolving ligand-induced RNA folding at atomic resolution, Proc. Natl. Acad. Sci. U.S.A. 104 (40) (2007) 15699–15704.
[52] J.R. Stagno, et al., Real-time crystallographic studies of the adenine riboswitch using an X-ray free-electron laser, FEBS J. 284 (20) (2017) 3374–3380.
[53] J. Rinnenthal, et al., Mapping the landscape of RNA dynamics with NMR spectroscopy, Acc. Chem. Res. 44 (12) (2011) 1292–1301.

Publications

1. **Bains, J. K.**; Qureshi, N. S.; Ceylan, B.; Wacker, A., Schwalbe, H. Cell-free transcription-translation system – A dual read-out assay to characterize riboswitch function. *Nucleic Acids Res* **2022***,* in revision

2. Ali, T.; Rogala, S.; Krause, N.M.; **Bains, J.K.**; Melissari, ;.T.; Währisch, S.; Schwalbe, H.; Hermann, B.G.; Grote, P. *Fendrr* synergizes with Wnt signalling to regulate fibrosis related genes during lung development via its RNA:dsDNA Triplex Element. *Nucleic Acids Res* **2022**, in revision

3. Leisegang, M. S.; **Bains, J. K**.; Seredinski, S.; Oo, J. A.; Krause, N. M.; Kuo, C.-C.; Günther, S.; Sentürk Cetin, N.; Warwick, T.; Cao, C.; Boos, F.; Izquierdo Ponce, J.; Haydar, S.; Bednarz, R.; Valasarajan, C.; Fuhrmann, D. C.; Preussner, J.; Looso, M.; Pullamsetti, S. S.; Schulz, M. H.; Jonker, H. R. A.; Richter, C.; Rezende, F.; Gilsbach, R.; Pflüger-Müller, B.; Wittig, I.; Grummt, I.; Ribarska, T.; Costa, I. G.; Schwalbe, H.; Brandes, R. P. HIF1α-AS1 Is a DNA:DNA:RNA Triplex-Forming LncRNA Interacting with the HUSH Complex. *Nat Commun* **2022**, *13* (1), 6563. https://doi.org/10.1038/s41467-022-34252-2.

4. Warwick, T.; Seredinski, S.; Krause, N. M.; **Bains, J. K.**; Althaus, L.; Oo, J. A.; Bonetti, A.; Dueck, A.; Engelhardt, S.; Schwalbe, H.; Leisegang, M. S.; Schulz, M. H.; Brandes, R. P. A Universal Model of RNA.DNA:DNA Triplex Formation Accurately Predicts Genome-Wide RNA-DNA Interactions. *Brief Bioinform* **2022**, *23* (6). https://doi.org/10.1093/bib/bbac445.

5. Berg, H.; Wirtz Martin, M. A.; Altincekic, N.; Alshamleh, I.; **Kaur Bains, J**.; Blechar, J.; Ceylan, B.; de Jesus, V.; Dhamotharan, K.; Fuks, C.; Gande, S. L.; Hargittay, B.; Hohmann, K. F.; Hutchison, M. T.; Marianne Korn, S.; Krishnathas, R.; Kutz, F.; Linhard, V.; Matzel, T.; Meiser, N.; Niesteruk, A.; Pyper, D. J.; Schulte, L.; Trucks, S.; Azzaoui, K.; Blommers, M. J. J.; Gadiya, Y.; Karki, R.; Zaliani, A.; Gribbon, P.; da Silva Almeida, M.; Dinis Anobom, C.; Bula, A. L.; Bütikofer, M.; Putinhon Caruso, Í.; Caterina Felli, I.; da Poian, A. T.; Cardoso de Amorim, G.; Fourkiotis, N. K.; Gallo, A.; Ghosh, D.; Gomes-Neto, F.; Gorbatyuk, O.; Hao, B.; Kurauskas, V.; Lecoq, L.; Li, Y.; Cunha Mebus-Antunes, N.; Mompeán, M.; Cristtina Neves-Martins, T.; Ninot-Pedrosa, M.; Pinheiro, A. S.; Pontoriero, L.; Pustovalova, Y.; Riek, R.; Robertson, A. J.; Jose Abi Saad, M.; Treviño, M. Á.; Tsika, A. C.; Almeida, F. C. L.; Bax, A.; Henzler-Wildman, K.; Hoch, J. C.; Jaudzems, K.; Laurents, D. v; Orts, J.; Pierattelli, R.; Spyroulias, G. A.; Duchardt-Ferner, E.; Ferner, J.; Fürtig, B.; Hengesbach, M.; Löhr, F.; Qureshi, N.; Richter, C.; Saxena, K.; Schlundt, A.; Sreeramulu, S.; Wacker, A.; Weigand, J. E.; Wirmer-Bartoschek, J.; Wöhnert, J.; Schwalbe, H. Comprehensive Fragment Screening of the SARS-CoV-2 Proteome Explores Novel Chemical Space for Drug Development. *Angew Chem Int Ed Engl* **2022**, *61* (46), e202205858. https://doi.org/10.1002/anie.202205858.

6. Mertinkus, K. R.; Tassilo Grün, · J; Altincekic, Nadide; **Bains, J. K.**; Ceylan, B.; Ferner, J.-P.; Frydman, L.; Fürtig, B.; Hengesbach, M.; Hohmann, K. F.; Hymon, D.; Kim, J.; Knezic, B.; Novakovic, M.; Oxenfarth, A.; Peter, S. A.; Nusrat, ·; Qureshi, S.; Richter, · Christian; Scherf, T.; Schlundt, A.; Schnieders, R.; Schwalbe, H.; Stirnal, E.; Sudakov, A.; Vögele, J.; Wacker, A.; Weigand, J. E.; Wirmer-Bartoschek, J.; Martin, M. A. W.; Wöhnert, J. H, 13 C and 15 N Chemical Shift Assignment

of the Stem-Loops 5b + c from the 5′-UTR of SARS-CoV-2 Biological Context. **2022**, *16*, 17–25. https://doi.org/10.1007/s12104-021-10053-4.

7. Richter, C.; Hohmann, K. F.; Toews, S.; Mathieu, D.; Altincekic, N.; **Bains, J. K.**; Binas, O.; Ceylan, B.; Duchardt-Ferner, E.; Ferner, J.; Fürtig, B.; Tassilo Grün, J.; Hengesbach, M.; Hymon, D.; A Jonker, H. R.; Knezic, B.; Korn, S. M.; Landgraf, T.; Löhr, F.; Peter, S. A.; Pyper, D. J.; Qureshi, N. S.; Schlundt, A.; Schnieders, R.; Stirnal, E.; Sudakov, A.; Vögele, J.; Weigand, J. E.; Wirmer-Bartoschek, J.; Witt, K.; Wöhnert, J.; Schwalbe, H.; Wacker, A. H, 13 C and 15 N Assignment of Stem-Loop SL1 from the 5′-UTR of SARS-CoV-2 Biological Context. *Biomol NMR Assign* **2021**, *15*, 467–474. https://doi.org/10.1007/s12104-021-10047-2.

8. Vögele, J.; Ferner, J. P.; Altincekic, N.; **Bains, J. K.**; Ceylan, B.; Fürtig, B.; Grün, J. T.; Hengesbach, M.; Hohmann, K. F.; Hymon, D.; Knezic, B.; Löhr, F.; Peter, S. A.; Pyper, D.; Qureshi, N. S.; Richter, C.; Schlundt, A.; Schwalbe, H.; Stirnal, E.; Sudakov, A.; Wacker, A.; Weigand, J. E.; Wirmer-Bartoschek, J.; Wöhnert, J.; Duchardt-Ferner, E. 1H, 13C, 15N and 31P Chemical Shift Assignment for Stem-Loop 4 from the 5′-UTR of SARS-CoV-2. *Biomol NMR Assign* **2021**, *1*, 1. https://doi.org/10.1007/s12104-021-10026-7.

9. de Jesus, V.; Qureshi, N. S.; Warhaut, S.; **Bains, J. K.**; Dietz, M. S.; Heilemann, M.; Schwalbe, H.; Fürtig, B. Switching at the Ribosome: Riboswitches Need RProteins as Modulators to Regulate Translation. *Nat Commun* **2021**, *12* (1), 4723. https://doi.org/10.1038/s41467-021-25024-5.

10. Elamri, I.; Abdellaoui, C.; **Bains, J. K.**; Hohmann, K. F.; Gande, S. L.; Stirnal, E.; Wachtveitl, J.; Schwalbe, H. Wavelength-Selective Uncaging of Two Different Photoresponsive Groups on One Effector Molecule for Light-Controlled Activation and Deactivation. *J Am Chem Soc* **2021**, *143* (28), 10596–10603. https://doi.org/10.1021/jacs.1c02817.

11. Sreeramulu, S.; Richter, C.; Berg, H.; Wirtz Martin, M. A.; Ceylan, B.; Matzel, T.; Adam, J.; Altincekic, N.; Azzaoui, K.; **Bains, J. K.**; Blommers, M. J. J.; Ferner, J.; Fürtig, B.; Göbel, M.; Grün, J. T.; Hengesbach, M.; Hohmann, K. F.; Hymon, D.; Knezic, B.; Martins, J. N.; Mertinkus, K. R.; Niesteruk, A.; Peter, S. A.; Pyper, D. J.; Qureshi, N. S.; Scheffer, U.; Schlundt, A.; Schnieders, R.; Stirnal, E.; Sudakov, A.; Tröster, A.; Vögele, J.; Wacker, A.; Weigand, J. E.; Wirmer-Bartoschek, J.; Wöhnert, J.; Schwalbe, H. Exploring the Druggability of Conserved RNA Regulatory Elements in the SARS-CoV-2 Genome. *Angewandte Chemie International Edition* **2021**, *60* (35), 19191–19200. https://doi.org/10.1002/anie.202103693.

12. Altincekic, N.; Korn, S. M.; Qureshi, N. S.; Dujardin, M.; Ninot-Pedrosa, M.; Abele, R.; Abi Saad, M. J.; Alfano, C.; Almeida, F. C. L.; Alshamleh, I.; de Amorim, G. C.; Anderson, T. K.; Anobom, C. D.; Anorma, C.; **Bains, J. K.**; Bax, A.; Blackledge, M.; Blechar, J.; Böckmann, A.; Brigandat, L.; Bula, A.; Bütikofer, M.; Camacho-Zarco, A. R.; Carlomagno, T.; Caruso, I. P.; Ceylan, B.; Chaikuad, A.; Chu, F.; Cole, L.; Crosby, M. G.; de Jesus, V.; Dhamotharan, K.; Felli, I. C.; Ferner, J.; Fleischmann, Y.; Fogeron, M.-L.; Fourkiotis, N. K.; Fuks, C.; Fürtig, B.; Gallo, A.; Gande, S. L.; Gerez, J. A.; Ghosh, D.; Gomes-Neto, F.; Gorbatyuk, O.; Guseva, S.; Hacker, C.; Häfner, S.; Hao, B.; Hargittay, B.; Henzler-

Wildman, K.; Hoch, J. C.; Hohmann, K. F.; Hutchison, M. T.; Jaudzems, K.; Jović, K.; Kaderli, J.; Kalniņš, G.; Kaņepe, I.; Kirchdoerfer, R. N.; Kirkpatrick, J.; Knapp, S.; Krishnathas, R.; Kutz, F.; zur Lage, S.; Lambertz, R.; Lang, A.; Laurents, D.; Lecoq, L.; Linhard, V.; Löhr, F.; Malki, A.; Bessa, L. M.; Martin, R. W.; Matzel, T.; Maurin, D.; McNutt, S. W.; Mebus-Antunes, N. C.; Meier, B. H.; Meiser, N.; Mompeán, M.; Monaca, E.; Montserret, R.; Mariño Perez, L.; Moser, C.; Muhle-Goll, C.; Neves-Martins, T. C.; Ni, X.; Norton-Baker, B.; Pierattelli, R.; Pontoriero, L.; Pustovalova, Y.; Ohlenschläger, O.; Orts, J.; da Poian, A. T.; Pyper, D. J.; Richter, C.; Riek, R.; Rienstra, C. M.; Robertson, A.; Pinheiro, A. S.; Sabbatella, R.; Salvi, N.; Saxena, K.; Schulte, L.; Schiavina, M.; Schwalbe, H.; Silber, M.; Almeida, M. da S.; Sprague-Piercy, M. A.; Spyroulias, G. A.; Sreeramulu, S.; Tants, J.-N.; Tārs, K.; Torres, F.; Töws, S.; Treviño, M. Á.; Trucks, S.; Tsika, A. C.; Varga, K.; Wang, Y.; Weber, M. E.; Weigand, J. E.; Wiedemann, C.; Wirmer-Bartoschek, J.; Wirtz Martin, M. A.; Zehnder, J.; Hengesbach, M.; Schlundt, A. Large-Scale Recombinant Production of the SARS-CoV-2 Proteome for High-Throughput and Structural Biology Applications. *Front Mol Biosci* **2021**, *8*. https://doi.org/10.3389/fmolb.2021.653148.

13. Schnieders, R.; Peter, S. A.; Banijamali, E.; Riad, M.; Altincekic, N.; **Bains, J. K.**; Ceylan, B.; Fürtig, B.; Grün, J. T.; Hengesbach, M.; Hohmann, K. F.; Hymon, D.; Knezic, B.; Oxenfarth, A.; Petzold, K.; Qureshi, N. S.; Richter, C.; Schlagnitweit, J.; Schlundt, A.; Schwalbe, H.; Stirnal, E.; Sudakov, A.; Vögele, J.; Wacker, A.; Weigand, J. E.; Wirmer-Bartoschek, J.; Wöhnert, J. 1H, 13C and 15N Chemical Shift Assignment of the Stem-Loop 5a from the 5′-UTR of SARS-CoV-2. *Biomol NMR Assign* **2021**, *15* (1), 203–211. https://doi.org/10.1007/s12104-021-10007-w.

14. Binas, O.; de Jesus, V.; Landgraf, T.; Völklein, A. E.; Martins, J.; Hymon, D.; **Bains, J. K.**; Berg, H.; Biedenbänder, T.; Fürtig, B.; Lakshmi Gande, S.; Niesteruk, A.; Oxenfarth, A.; Shahin Qureshi, N.; Schamber, T.; Schnieders, R.; Tröster, A.; Wacker, A.; Wirmer-Bartoschek, J.; Wirtz Martin, M. A.; Stirnal, E.; Azzaoui, K.; Richter, C.; Sreeramulu, S.; José Blommers, M. J.; Schwalbe, H. 19 F NMR-Based Fragment Screening for 14 Different Biologically Active RNAs and 10 DNA and Protein Counter-Screens. *Chembiochem* **2021**, *22* (2), 423–433. https://doi.org/10.1002/cbic.202000476.

15. Kubatova, N.; Qureshi, N. S.; Altincekic, N.; Abele, R.; **Bains, J. K.**; Ceylan, B.; Ferner, J.; Fuks, C.; Hargittay, B.; Hutchison, M. T.; de Jesus, V.; Kutz, F.; Wirtz Martin, M. A.; Meiser, N.; Linhard, V.; Pyper, D. J.; Trucks, S.; Fürtig, B.; Hengesbach, M.; Löhr, F.; Richter, C.; Saxena, K.; Schlundt, A.; Schwalbe, H.; Sreeramulu, S.; Wacker, A.; Weigand, J. E.; Wirmer-Bartoschek, J.; Wöhnert, J. 1H, 13C, and 15N Backbone Chemical Shift Assignments of Coronavirus-2 Non-Structural Protein Nsp10. *Biomol NMR Assign* **2021**, *15* (1), 65–71. https://doi.org/10.1007/s12104-020-09984-1.

16. Cantini, F.; Banci, L.; Altincekic, N.; **Bains, J. K.**; Dhamotharan, K.; Fuks, C.; Fürtig, B.; Gande, S. L.; Hargittay, B.; Hengesbach, M.; Hutchison, M. T.; Korn, S. M.; Kubatova, N.; Kutz, F.; Linhard, V.; Löhr, F.; Meiser, N.; Pyper, D. J.; Qureshi, N. S.; Richter, C.; Saxena, K.; Schlundt, A.; Schwalbe, H.; Sreeramulu, S.; Tants, J. N.; Wacker, A.; Weigand, J. E.; Wöhnert, J.; Tsika, A. C.; Fourkiotis, N. K.; Spyroulias, G. A. 1H, 13C, and 15N Backbone Chemical Shift Assignments of the Apo and the ADP-Ribose Bound Forms of the Macrodomain of SARS-CoV-2 Non-Structural Protein 3b. *Biomol NMR Assign* **2020**, *14* (2), 339–346. https://doi.org/10.1007/s12104-020-09973-4.

17. Wacker, A.; Weigand, J. E.; Akabayov, S. R.; Altincekic, N.; **Bains, J. K.**; Banijamali, E.; Binas, O.; Castillo-Martinez, J.; Cetiner, E.; Ceylan, B.; Chiu, L. Y.; Davila-Calderon, J.; Dhamotharan, K.; Duchardt-Ferner, E.; Ferner, J.; Frydman, L.; Fürtig, B.; Gallego, J.; Grün, J. T.; Hacker, C.; Haddad, C.; Hähnke, M.; Hengesbach, M.; Hiller, F.; Hohmann, K. F.; Hymon, D.; de Jesus, V.; Jonker, H.; Keller, H.; Knezic, B.; Landgraf, T.; Löhr, F.; Luo, L.; Mertinkus, K. R.; Muhs, C.; Novakovic, M.; Oxenfarth, A.; Palomino-Schätzlein, M.; Petzold, K.; Peter, S. A.; Pyper, D. J.; Qureshi, N. S.; Riad, M.; Richter, C.; Saxena, K.; Schamber, T.; Scherf, T.; Schlagnitweit, J.; Schlundt, A.; Schnieders, R.; Schwalbe, H.; Simba-Lahuasi, A.; Sreeramulu, S.; Stirnal, E.; Sudakov, A.; Tants, J. N.; Tolbert, B. S.; Vögele, J.; Weiß, L.; Wirmer-Bartoschek, J.; Wirtz Martin, M. A.; Wöhnert, J.; Zetzsche, H. Secondary Structure Determination of Conserved SARS-CoV-2 RNA Elements by NMR Spectroscopy. *Nucleic Acids Res* **2020**, *48* (22), 12415–12435. https://doi.org/10.1093/nar/gkaa1013.

18. **Bains, J. K.**; Blechar, J.; de Jesus, V.; Meiser, N.; Zetzsche, H.; Fürtig, B.; Schwalbe, H.; Hengesbach, M. Combined smFRET and NMR Analysis of Riboswitch Structural Dynamics. *Methods* **2019**, *153*, 22–34. https://doi.org/10.1016/j.ymeth.2018.10.004.

19. Qureshi, N. S.; **Bains, J. K.**; Sreeramulu, S.; Schwalbe, H.; Fürtig, B. Conformational Switch in the Ribosomal Protein S1 Guides Unfolding of Structured RNAs for Translation Initiation. *Nucleic Acids Res* **2018**, *46* (20), 10917–10929. https://doi.org/10.1093/nar/gky746.