Transcriptome-wide functions of the RNA-binding protein PURA in health and disease

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Biowissenschaften (FB15)
der Johann Wolfang Goethe-Universität
in Frankfurt am Main

von

Melina Klostermann

geboren in Heidelberg

Frankfurt 2023

D30

Vom Fachbereich Biowissenschaften (FB15)

der Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr. Sven Klimpel

1. Gutachter: Dr. Kathi Zarnack

2. Gutachter: Prof. Dr. Marcel Schulz

Datum der Disputation:

“After sleeping through a hundred million centuries
We have finally opened our eyes on a sumptuous planet
Sparkling with color, bountiful with life
Within decades we must close our eyes again
Isn’t it a noble, an enlightened way of spending our brief
Time in the sun, to work at understanding the universe
And how we have come to wake up in it?”

The Greatest Show on Earth, Nightwish & Richard Dawkins

Zusammenfassung

Das Interesse an dem Protein PURA hat in letzter Zeit durch die Entdeckung des PURA-Syndroms zugenommen. Diese seltene Krankheit wird durch De-novo-Mutationen im *PURA*-Gen ausgelöst und verursacht eine neurologische Entwicklungsstörung. Zusätzlich tauchte PURA in Studien zu der Amyotrophen Lateralsklerose und dem Fragilen-X-assoziierten-Tremor-Ataxie-Syndrom als neuroprotektiver Faktor auf. Das Protein PURA wurde *in vivo* zumeist als DNA-bindendes Protein beschrieben. Bindestudien *in vitro* haben jedoch gezeigt, dass PURA RNAs mit derselben Affinität bindet. In dieser Dissertation habe ich deshalb die zelluläre RNA-Bindungsaktivität von PURA *in vivo* untersucht. Weiterhin habe ich die Verbinding der PURA RNA-Bindung zu molekularen und zellulären Veränderungen in Zellen mit beeinträchtigter PURA Funktion betrachtet, wie sie auch in Patienten mit PURA-Syndrom auftritt.

Um die Bindung von PURA und die Auswirkungen des PURA-Mangels auf die zelluläre RNA- und Proteinexpression zu untersuchen, habe ich eine bioinformatische Analyse von Hochdurchsatz-Experimenten durchgeführt. Ein wesentlicher Bestandteil war die Untersuchung von UV-Crosslinking-und-Immunopräzipitations-Experimenten (CLIP), mit denen das globale RNA-Bindungsverhalten eines bestimmten Proteins im zellulären Kontext untersucht werden kann. Da die Verarbeitung und Analyse von CLIP-Daten recht komplex sind, habe ich im Rahmen dieser Dissertation eine automatisierte Kommandozeilen-Anwendung für die Verarbeitung von CLIP-Daten namens racoon_clip entwickelt.

Daher besteht diese Dissertation aus zwei Abschnitten. Erstens beschreibe ich den Aufbau und die Verwendung von racoon_clip für die CLIP-Datenanalyse. Zweitens zeige ich meine Forschungen über das Protein PURA, wobei ich seine globalen RNA-Bindungseigenschaften, die Auswirkungen eines PURA-Mangels auf die zelluläre RNA und Protein Expression und seine Assoziation mit neuronalen Kontexten und P-Bodies aufzeige.

Automatisierte Prozessierung von Daten aus UV-Crosslinking-und-Immunopräzipitations-Experimenten mit racoon_clip

racoon_clip ist eine Kommandozeilen-Anwendung, die ich mit dem Ziel entwickelt habe, die Daten von iCLIP (*individual-nucleotide resolution CLIP*) und eCLIP (*enhanced CLIP*) Experimenten - zwei der am häufigsten verwendeten Arten von CLIP-Experimenten - auf vergleichbare und benutzerfreundliche Weise zu prozessieren. Obwohl der Informationsgehalt von Daten dieser Experimente vergleichbar ist, werden für sie oft verschiedene Ansätze zur Prozessierung verwendet. Aus diesen Grund habe ich racoon_clip als automatisierten Arbeitsablauf entwickelt, der alle CLIP-Verarbeitungsschritte von den Rohdaten bis hin zu hochauflösenden Einzelnukleotid-Bindeereignissen umfasst. Diese beinhalten eine Qualitätskontrolle der Rohdaten, das Auftrennen gemeinsam sequenzierter Datensätze (Demultiplexen), das Entfernen von Barcode- und Adapter-Sequenzen, das Alinieren der Sequenzen an ein gegebenes Genom, das Deduplizieren der Sequenzen und schließlich das Bestimmen der *Crosslink*-Positionen.

racoon_clip kann vom Benutzer mit einem einzigen Befehl über ein Kommandozeilen-Interface ausgeführt werden. Unter dem Kommandozeilen-Interface sind die einzelnen Schritte mit Snakemake-Workflow-Management implementiert, was rechnerische Vorteile wie Parallelisierung, Skalierbarkeit und Transferierbarkeit des Arbeitsablaufs bietet. racoon_clip ist öffentlich zugänglich und kann von GitHub heruntergeladen werden. Außerdem gibt es eine ausführliche Dokumentation der Anwendung auf der Open-Source-Hosting-Plattform Read the Docs (https://racoon-clip.readthedocs.io/en/latest/).

Die Hauptaufgabe von racoon_clip besteht darin, RNA-Bindeereignisse mit Einzelnukleotid-Auflösung aus iCLIP-, iCLIP2-, eCLIP-Daten und ähnlichen Datentypen zu ermitteln. Um sicherzustellen, dass racoon_clip sowohl in hohem Maße anpassbar als auch einfach zu bedienen ist, bietet die Software voreingestellte Optionen für die gängigsten Experimenttypen. Zusätzlich haben die Benutzer die Möglichkeit, ein benutzerdefiniertes Setup von Barcode- und Adapterzusammensetzungen zu erstellen, um die Software für andere Arten von CLIP-Daten zu verwenden. Während die unterschiedlichen Architekturen in den Sequenzen berücksichtigt werden, bleiben die durchgeführten zentralen Verarbeitungsschritte die Gleichen. Dies führt zu einem hohen Maß an Vergleichbarkeit zwischen den verschiedenen Experimenttypen, was ich am Beispiel der Prozessierung von U2AF2 iCLIP- und eCLIP-Daten demonstriere. Insgesamt bin ich überzeugt, dass racoon_clip für zahlreiche Forscher, die sich für RNA-Protein-Interaktionen interessieren, von Nutzen sein wird, da es eine leicht zugängliche Prozessierung von CLIP-Daten erlaubt und die Vergleichbarkeit mehrerer CLIP-Datensätze über verschiedene Experimenttypen hinweg verbessert.

Die zelluläre Funktion von PURA und deren Bedeutung für PURA assoziierte Krankheiten

Im zweiten Teil dieser Dissertation zeige ich meine Ergebnisse zur zellulären Funktion des RNA-bindenden Proteins PURA. Durch eingehende bioinformatische Analysen von iCLIP Daten für endogenes und überexprimiertes PURA in HeLa Zellen habe ich eine globale Funktion von PURA als RNA-Bindeproteinfunktion bestätigen können. Es bindet die Transkripte von 4,391 Genen an über 50,000 Bindestellen. Diese liegen zumeist entweder in der kodierenden Region (*coding sequence*, CDS) oder in der 3' untranslatierten Region (3'UTR) protein-kodierender Transkripte. Dabei erkennt PURA ein Purin-reiches degeneriertes Sequenzmotiv in einem einzelsträngigen Bereich der Transkripte. Obwohl die Überexpression von PURA zu einem weniger spezifischen Bindungsverhalten führt, bleiben die gleichen allgemeinen Bindemuster wie bei endogenem PURA erhalten. Die allgemeinen Merkmale der PURA-Bindung sind daher in drei verschiedenen PURA-iCLIP-Datensätzen mit und ohne PURA-Überexpression ähnlich.

Um mehr über die molekularen Effekte eines partiellen Verlusts von funktionalem PURA

herauszufinden, habe ich eine 50-prozentige Reduktion von PURA in HeLa-Zellen als Modell für den heterozygoten Verlust von PURA beim PURA-Syndrom verwendet und dessen Auswirkungen auf die globale RNA- und Proteinexpression bewertet. Die Ergebnisse zeigen, dass die PURA-Reduktion die RNA- und Proteinexpression global beeinflusst. Weiterhin habe ich die PURA-RNA-Bindung mit den Expressionsniveaus von RNAs und Proteinen im Kontext der PURA-Reduktion integriert. Dabei ergaben sich keine eindeutigen Hinweise für eine Funktion von PURA in der Regulierung der RNA-Stabilität oder -Translation. Obwohl sich die Bindemuster in 3'UTR und CDS deutlich unterscheiden, konnte ich keinen Trend feststellen, dass die Transkriptregion, in der die Bindung stattfindet, einen Einfluss auf die Expressionveränderung von RNAs oder Proteinen hätte. Allerdings konnte ich 234 Transkripte herausarbeiten, die mit hoher Wahrscheinlichkeit in direkter Abhängigkeit von PURA stehen, da sie von PURA gebunden werden und sowohl auf RNA- als auch auf Proteinebene durch Reduktion der PURA Levels beeinflusst werden. Die in diesen Ziel-Transkripten kodierten Proteine bringen PURA unter anderem mit neuronaler Entwicklung, der Regulation von RNAs und mitochondrialen Funktionen in Verbindung.

Im Einklang mit einer möglichen Rolle von PURA beim neuronalen Transport gibt es erhebliche Überschneidungen zwischen PURA gebundenen Transkripten und Transkripten, die im Neuron in die Enden der Dendriten transportiert werden. Zusätzlich habe ich ein PURA-iCLIP-Experiment in neuronalen Vorläuferzellen untersucht, um die RNA-Bindung von PURA in physiologischen Bedingungen zu betrachten, die im Kontext einer neuronalen Entwicklungstörung relevant sind. Obwohl das iCLIP-Experiment in diesen Zellen eine limitierte Signalstärke hat und nicht zur eigenständigen Definition von Bindestellen verwendet werden konnte, konnte ich zeigen, dass die globale Verteilung der RNA-Bindeereignisse ähnlich der in HeLa Zellen ist. Daher kann näherungsweise ein ähnliches Bindeverhalten von PURA in beiden Zelltypen angenommen werden.

Desweiteren trat eine bisher unbekannte Verbindung von PURA zu Prozessierungskörperchen (*Processing Bodies*, P-Bodies) zu Tage. P-Bodies sind zytoplasmatische, membranlose, zelluläre Kompartimente, die durch Flüßig-Flüßig-Phasentrennung entstehen. Hier zeige ich, dass PURA-gebundene RNAs im Transkriptom von P-Bodies sowie von Stresskörperchen (*Stress Granules*) - einer weiteren Sorte von zytoplasmatischen, membranlosen, zellulären Kompartimenten - angereichert sind. Darüber hinaus konnten meine Kollaborationspartner zeigen, dass PURA in P-Bodies lokalisiert. Interessanterweise sind in Zellen, in denen die Menge an PURA Protein reduziert wird, deutlich weniger P-Bodies vorhanden. Diese Reduktion von P-Bodies könnte auf die Herunterregulierung der Proteine LSM14A (*LSM14A mRNA processing body assembly factor*) und DDX6 (*DEAD-Box Helicase 6*) zurückgeführt werden, beides Faktoren, die zuvor als wesentlich für die Bildung von P-Bodies identifiziert wurden. Weiterhin konnte ich feststellen, dass P-Body-assoziierte Transkripte bei einer Reduktion von PURA, und der damit einhergehenden Verminderung der P-Body Anzahl, vermehrt exprimiert werden. Insgesamt können die verminderte Anzahl der P-Bodies bei PURA-Reduktion, die neuronale Funktion von PURA und seine Verbindung zu Mitochondrien und der Regulation des RNA-Lebenszyklus für die zellulären Grundlagen sowohl des PURA-Syndroms als auch anderer von PURA beeinflusster, neuronaler Erkrankungen eine Rolle spielen.

Zusammenfassend konnte meine umfangreiche bioinformatische Analyse von iCLIP- und anderen Hochdurchsatzdaten wertvolle Beiträge zur Klärung der zellulären Funktion von PURA leisten. Diese Erkenntnisse tragen zum Verständnis der Auswirkungen des PURA-Verlusts beim PURA-Syndrom und anderen Krankheiten bei.

Insgesamt bietet diese Dissertation ein Beispiel für die erfolgreiche Verknüpfung von iCLIP und anderen Hochdurchsatz-Daten, um die zelluläre Funktion eines RNA-Bindeproteins zu verstehen. Um das Prozessieren von iCLIP-Daten zu erleichtern, stelle ich ein benutzerfreundliches Softwareprogram vor. Weiterhin konnte meine Arbeit wesentlich zum

Verständins der Funktion des Proteins PURA beitragen und liefert dadurch eine Grundlage für folgende Forschungsarbeiten an PURAs Role in diversen neuronalen Krankheiten.

Abstract

The attention on the protein PURA has increased recently following the discovery of the rare PURA Syndrome. This neurodevelopmental disorder is caused by *de novo* mutations in the *PURA* gene. Notably, our collaborators could show that the protein PURA can bind DNA and RNA *in vitro*. As a result, I was motivated to explore PURA's cellular RNA-binding activity. Furthermore, I inquired on the connection of PURA-RNA binding to the cellular effect of a reduction of functional PURA as present in PURA Syndrome patients. To investigate the binding of PURA and the impact of PURA deficiency on cellular RNA and protein expression, I performed an integrative computational analysis of multimodal data from complementary high-throughput experiments. An essential component was the examination of UV Crosslinking and immunoprecipitation (CLIP) experiments, which can query the global RNA-binding behaviour of a given protein in a cellular context. As the processing and analysis of CLIP data are rather complex, I introduce an automated command line tool for the processing of CLIP data named racoon_clip as part of this dissertation. Therefore, this dissertation comprises two major segments. Firstly, I describe the implementation and usage of racoon_clip for CLIP data analysis. Secondly, I discuss my research on the protein PURA, demonstrating its global RNA-binding properties, the effects of PURA depletion and its association with neuronal functions and P-bodies, among others.

racoon_clip is a command line application that I have developed for processing of individual-nucleotide resolution CLIP (iCLIP) and enhanced CLIP (eCLIP) experiments - two of the most commonly used types of CLIP experiments - in a comparable and user-friendly way. For this, I built racoon_clip as an automated workflow that encompasses all CLIP processing steps from raw data to single-nucleotide resolution crosslink events. racoon_clip is available as a command line tool that users can run with a single command. The workflow is implemented with Snakemake workflow management providing computational advantages including parallelisation, scalability and portability of the workflow. The main task of racoon_clip is to extract single-nucleotide crosslink events from iCLIP, iCLIP2, eCLIP and similar data types. To strike a balance between being highly customisable and easy to use, racoon_clip supplies pre-set options for the most common types of experiments. Additionally, it is possible for users to create a custom setup of barcode and adapter architectures, which allows them to use the software for other types of CLIP data. While accounting for the different architectures in the reads, the performed central processing steps remain the same. This leads to a high degree of comparability between the different experiment types, which I demonstrate in the exemplary processing of U2AF2 iCLIP and eCLIP data. Taken together, I am confident that racoon_clip will be beneficial to numerous researchers interested in RNA-Protein interactions as it offers easily accessible processing for CLIP data and enhances the comparability of multiple CLIP datasets across different experiment types.

In the second part of this dissertation, I focus on the cellular function of the RNA-binding protein PURA. Through in-depth computational analysis of one iCLIP data set of endogenous PURA and two iCLIP data sets of overexpressed PURA in HeLa cells, I establish that PURA is a global RNA-binding protein. It preferentially binds RNAs in either the coding sequence (CDS) or the 3' untranslated region (3'UTR) of mature protein-coding transcripts by recognising a Purine-rich degenerated sequence motif. Even though overexpression of PURA results in less specific binding behaviour, the same overall binding patterns as from endogenous PURA persist. Overall characteristics of PURA binding remain similar in three distinct PURA iCLIP data sets with and without PURA overexpression.

To learn about the molecular consequences of a depletion of functional PURA in a cellular context, I used a 50% reduction of PURA in HeLa cells as a model for the heterozygous loss of PURA in PURA Syndrome and evaluated its impact on global RNA and protein expression. The results demonstrate that PURA depletion globally affects RNA and protein expression. Additionally, I integrate PURA RNA binding with the changes in expression of RNAs and proteins in the context of PURA depletion. This reveals 234 targets of PURA that are bound by PURA and are impacted at both RNA and protein levels by the PURA protein. RNAs that are bound by PURA or change in abundance upon PURA depletion are enriched in neuronal development factors, RNA lifecycle regulators, and mitochondrial factors, among others. Consistent with a possible role of PURA in neuronal transport, there is considerable overlap between PURA bound transcripts and transcripts, that are transported to the dendritic end of neurons.

Notably, there is a link between PURA and P-bodies, as documented by the enrichment of PURA-bound RNAs in both the P-body and stress granule transcriptome. Further, PURA was found by our collaborators to be localised within P-bodies and P-body numbers were strongly reduced in cells that are depleted of PURA. This absence might be attributed to the downregulation of the proteins encoded by the PURA targets *LSM14A* and *DDX6* as both of them were previously identified as essential for P-body formation. Overall, the reduction of P-body numbers in PURA depletion, the neuronal function of PURA, and its association with mitochondria and RNA lifecycle regulation may indicate the cellular foundation of both PURA Syndrome and related neuronal diseases.

In summary, I present a versatile and user-friendly computational tool for the analysis of CLIP data. Subsequently, I conduct a thorough computational analysis of CLIP and other high-throughput data in the context of the RNA-binding protein PURA, which offers valuable insights into the cellular functions of PURA. These insights advance our understanding of the impact of PURA loss in PURA Syndrome and other disease contexts.

Preface

The work on PURA function (**Chapter 4**) was done in close collaboration with the Niessing groups at the Helmholtz Center Munich and Ulm university, the Gene Centre Munich at the Ludwigs-Maximilians-Universität in Munich, the Core Facility Bioinformatics at the Institute of Molecular Biology (IMB) in Mainz and the Metabolomics and Proteomics Core at Helmholtz Center Munich. All analysed high-throughput experiments were performed by Lena Molitor in the Niessing group at Helmholtz Zentrum München. The immunofluorescence stainings shown in **Figure 4.25 A & B** were done by Sabrina Bacher in the same group. The iCLIP and RNAseq libraries were sequenced by Stephan Krebs at the Gene Centre Munich. The iCLIP data was preprocessed from raw reads to single nucleotide crosslinks by Anke Busch in the Bioinformatics Core Facility at IMB in Mainz. Shotgun proteomics and preprocessing of proteomic data was done by Juliane Merl-Pham in the Metabolomics and Proteomics Core at Helmholtz Center Munich. All further analyses of **Chapter 4** was done by me, aided by regular discussions with Kathi Zarnack, Lena Molitor and Dierk Niessing.

The results of this thesis are included in the following publications:

Depletion of the RNA-binding protein PURA triggers changes in posttranscriptional gene regulation and loss of P-bodies.
Molitor L*, Klostermann M*, Bacher S, Merl-Pham J, Spranger N, Burczyk S, Ketteler C, Rusha E, Tews D, Pertek A, Proske M, Busch A, Reschke S, Feederle R, Hauck SM, Blum H, Drukker M, Fischer-Posovszky P, König J, Zarnack K$^{\$}$, Niessing D$^{\$}$ (2023); Nucleic Acids Research, 51(3):12971316. DOI http://dx.doi.org/10.1093/nar/gkac1237. (** both authors contributed equally;* $^{\$}$ *shared correspondence*)

racoon_clip – a complete pipeline for single-nucleotide analyses of iCLIP and eCLIP data.
Klostermann M and Zarnack K; *Manuscript is currently under review at Bioinformatics.*

Acknowledgements

This dissertation would not have been possible without the support from many sites and I am grateful to everyone who helped me along the way. I enjoyed working with all of you!

First and foremost, I am absolutely thankful to my supervisor Kathi, who believed that I could learn coding in a master thesis without any prior knowledge. Your feedback was invaluable for all of my projects, talks and texts, including this dissertation, your set me up with nice projects and collaborations, sent me travelling around the world and with your aura of calmness you always had my back.

I'm grateful to Marcel Schulz, my second supervisor, and Andreas Schlund, a member of my thesis advisory committee, for their insightful contributions and guidance during our discussions in the TAC meetings.

I want to thank all 20 great collaborators on the PURA project and the great PURA Syndrome community. My thanks go especially to Lena, who already knew every paper on PURA there is and from whose experience I learn a lot and to Dierk who guided us in this project. I am also thankful to the PUR protein PhDs Sabrina, Sandra, Caro and Thomas from the Niessing groups in München and Ulm, who all contributed to this project and who shared with me many other ongoing research related to the PUR proteins.

Beyond the PURA project, I am thankful for the collaborations in my side projects. Although not explicitly mentioned in this dissertation, these collaborations offered me the opportunity to delve into diverse biological questions, broadening my knowledge and skills.

I enjoyed being a part of the Zarnack group and am thankful to all members who contributed to the relaxed and friendly surrounding and the many discussions on projects, computational methods and coding tricks. Among others these are Mirko who shared his iCLIP expertise with me, Mario who taught me many R coding tricks, You and Dominik who were test using and reading racoon_clip and its documentation and who together with Clara kept me company in the office. All of you made this a both productive and fun time.

Lastly, I'm deeply thankful to my family and friends who totally supported me during this time and who are the best in taking my head off work. Especially thanks to Livia who was prove reading the whole thesis and Julius who took care of everything else while I was writing.

Abbreviations

A

ACR Acrosin

ALS Amyotrophic Lateral Sclerosis

AGO2 Argonaute RISC Catalytic Component 2

B

bash Bourne-again shell

BC1 Brain Cytoplasmatic RNA 1

C

C9ORF72 Chromosome 9 Open Reading Frame 72

CAMK2 Calcium/Calmodulin Dependent Protein Kinase II

CDS coding sequence

ChIP-seq Chromatin Immunoprecipitation Sequencing

CLIP UV Crosslinking and Immunoprecipitation

- **CLIP-seq** CLIP coupled with high-throughput sequencing
- **HITS-CLIP** High-Throughput Sequencing of RNA isolated by CLIP
- **eCLIP** enhanced CLIP
- **iCLIP** individual-nucleotide resolution CLIP
- **PAR-CLIP** Photoactivable Ribonucleoside-enhanced CLIP

CNOT1 CCR4-NOT Transcription Complex Subunit 1

CPU Central Processing Unit

CTCF CCCTC-Binding Factor

CTNNA1 Catenin Alpha 1

CUX1 Cut Like Homeobox 1

D

DCP1A mRNA-Decapping Enzyme 1A

DCP2A mRNA-Decapping Enzyme 2A

DDD study Deciphering Developmental Disorders study

DDX3 DEAD-Box Helicase 3

DDX6 DEAD-Box Helicase 6

DNA Deoxyribonucleic Acid

- **cDNA** complementary DNA

E

ecdf empirical cumulative distribution function

EIF4ENIF1 Eukaryotic Translation Initiation Factor 4E Transporter, also known as 4E-T

ENCODE Encyclopedia of DNA Elements

F

FDR false discovery rate

FL-AB anti-FLAG antibody

FMR1 Fragile X Messenger Ribonucleoprotein 1

FTD Frontotemporal Dementia

FUS FUS (Fused in Sarcoma) RNA Binding Protein

FXTAS Fragile X-Associated Tremor/Ataxia syndrome

G

G3BP Ras GTPase-Activating Protein-Binding Protein

GB gigabyte

GO Gene Ontology

GTEx Genotype-Tissue Expression project

H

HIV-1 Human Immunodeficiency Virus 1

HNRNPA1 Heterogeneous Nuclear Ribonucleoprotein A1

HNRNPC Heterogeneous Nuclear Ribonucleoprotein C

HPC High-Performance Computing

I

IH-AB anti-PURA in-house antibody

iPSCs induced pluripotent stem cells

J

JCV Human Polyomavirus 2

K

KIF5 Kinesin Family Member 5

L

LLPS liquid-liquid phase separation

LSM14A LSM14A mRNA Processing Body Assembly Factor (LSM14 Homolog A)

M

MLOs membrane-less organelles

MYC V-Myc Avian Myelocytomatosis Viral Oncogene Homolog

N

NPC neuronal precursor cells

nt nucleotide

P

PABP PolyA-Binding Protein

PABPC1 PolyA-Binding Protein Cytoplasmic 1

P-body processing body

PCR polymerase chain reaction

PUM2 Pumilio Homolog 2

PURA Purine-rich Binding Element Alpha, also known as PUR-alpha or PUR-α

PURB Purine-rich Binding Element Beta

PURG Purine-rich Binding Element Gamma

PUR-protein Purine-rich binding protein

PWM position weight matrix

PyPi python package index

R

RAM Random access memory

RBP RNA-binding protein

RNA Ribonucleic acid

- **lncRNA** long non-coding RNA
- **miRNA** micro RNA
- **mRNA** messenger RNA
- **rRNA** ribosomal RNA
- **siRNA** small interfering RNA
- **snoRNA** small nucleolar RNA
- **snRNA** small nuclear RNA
- **tRNA** transfer RNA

pre-mRNA premature messenger RNA

RNAseq RNA sequencing

RNP ribonucleoprotein complex

RRM RNA recognition motif

RSS residual sum of squares

S

SEC61A1 SEC61 Translocon Subunit Alpha 1

SELEX systematic evolution of ligands by exponential enrichment

SLURM simple linux utility for resource management

SMInput size-matched input control

SRA sequence read archive

SSH secure socket shell

STAR spliced transcripts alignment to a reference algorithm

STARD7 StAR-Related Lipid Transfer Domain Protein 7

STAU1 Staufen Double-Stranded RNA Binding Protein 1

SOD1 Superoxid Dismutase-1

SOX2 SRY-Box Transcription Factor 2

T

TIA T-Cell-Restricted Intracellular Antigen

TIAL T-Cell-Restricted Intracellular Antigen Like

TPM transcripts per million

U

U2AF2 U2 Small Nuclear RNA Auxiliary Factor 2

UMI Unique molecular identifier

UPF1 Up-Frameshift Suppressor 1 Homolog

UTR Untranslated region

Y

YBX1 Y-Box Binding Protein 1

YWHAE Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Epsilon

Contents

Chapter 1

Introduction

In the last two decades, RNAs have become a major field of interest in biology. They are the central piece connecting genetic information in the DNA and cellular functionalities. RNAs undergo a complex lifecycle. The maturation of RNA includes multiple processes like splicing, editing and modification. Mature RNAs will be further regulated in their stability, their localisation and - in case of messenger RNAs - their translation. The regulation of these processes is performed by RNA-binding proteins (RBPs). Therefore, RBPs can be major regulators of the RNA lifecycle and cellular functions and, consequently, understanding of them becomes an essential compartment of understanding of RNA dynamics and their effects in cellular contexts [Dreyfuss *et al.*, 2002, Gerstberger *et al.*, 2014, Singh *et al.*, 2015]. Unsurprisingly, the dysfunction of RNA-RBP interactions has been shown to play a role in a vast number of disease contexts [Cooper *et al.*, 2009, Zipeto *et al.*, 2015].

Together with the interest in RNA, the usage of high-throughput sequencing experiments has risen and there is now a multitude of experimental protocols available that can be used to address different questions about RNA on a global scale [Reuter *et al.*, 2015]. The most common of these is RNA sequencing, which can be used to query RNA expression changes between different conditions. Furthermore, RBP binding to the complete transcriptome of a cell can be elucidated by UV crosslinking and immunoprecipitation (CLIP) experiments, where an RBP of interest is crosslinked to its bound RNAs in a cell and sequence libraries of these bound RNAs are obtained [Ule *et al.*, 2003, Lee and Ule, 2018, Hafner *et al.*, 2021]. These high-throughput sequencing experiments produce large amounts of sequencing data that need specially tailored bioinformatics analysis. The development and execution of bioinformatic approaches, has therefore become an important part in the research of RNAs and RBPs. Moreover, the growing complexity of the computational analysis of high-throughput sequencing data asks for standardised analysis methods that reduce the time and expertise needed for the analysis, while rendering experiments from different research groups more comparable.

High-throughput sequencing also advanced the diagnostics and discovery of genetic diseases via for example whole genome sequencing, which allows to query for alterations of genes in the complete genome of patients. The Deciphering Developmental Disorders (DDD) study (http://www.ddduk.org) [Wright *et al.*, 2014] was one large-scale approach to find previously unknown genetic implications in developmental disorders that could not previously be diagnosed via whole genome sequencing and complementary methods. This aided the description of various rare genetic diseases. A rare disease is defined as a disease with less than 2,000 patients world-wide. Although the small number may seem neglectable, the existence of over 7,000 rare diseases underlines the importance of studying them [Boycott *et al.*, 2017]. Among others, the DDD study found *de novo* mutations in the *PURA* gene. These cause the newly described neurodevelopmental disorder PURA Syndrome [Reijnders *et al.*, 2018], with until now more then 150 clinically described patients. Phenotypes of PURA Syndrome include intellectual disability, movement problems, hypotonia, epilepsy and gastrointestinal abnormalities. As little was known about the cellular function of the PURA protein, which seems to be heterozygously lost in

PURA Syndrome patients, I set out to delineate PURAs global cellular function from high-throughput experiments and could establish, that PURA has a global function as an RBP.

1.1 Outline of this dissertation

In this thesis, I present an advanced processing method for CLIP experiments. This method has potential applications in various biological contexts for the investigation of RBPs. Furthermore, I show the work of me and my collaborators on comprehending the cellular RNA-binding function of the Protein PURA, which provides the groundwork for understanding PURA-related diseases.

In the **Chapter 2 - Background** a comprehensive background on the subject matter is provided, which covers the biological concepts and disease contexts, along with the basics on the used statistical analyses, algorithms and computational methods.

Chapter 3 - racoon_clip: an automated pipeline for iCLIP and eCLIP data processing details the implementation of racoon_clip - a command-line application for the automated processing of CLIP data - including the exemplary processing of U2AF2 iCLIP and eCLIP data.

Chapter 4 - The cellular role of the RNA-binding protein PURA and its connection to cytoplasmic granules presents our findings on the RNA-binding protein PURA. Firstly, this study establishes the universal RNA-binding function of PURA and its binding characteristics. Secondly, PURA reduction was used to simulate the heterozygous loss of functional PURA in PURA Syndrome patients. Thirdly, these findings are connected to neuronal contexts and cytoplasmic granules.

Chapter 5 - Conclusion and Outlook provides a conclusion to the research presented.

Chapter 2

Background

2.1 RNAs and their regulation by RNA-binding proteins

Ribonucleic acid (RNA) is transcribed from genes stored in the deoxyribonucleic acid (DNA) of a cell. In the case of messenger RNAs (mRNA), they are then translated into proteins [Crick, 1958]. These proteins, often described as cellular machines or building blocks, execute cellular processes [Alberts, 1998]. In this chain of action, RNAs are the essential link between stored information and executed functions. Their lifecycle is heavily regulated by various RNA-binding proteins (RBPs) [Dreyfuss *et al.*, 2002, Gerstberger *et al.*, 2014, Singh *et al.*, 2015] and misregulation of RNAs is frequently linked to disease [Cooper *et al.*, 2009, Zipeto *et al.*, 2015].

In addition to mRNAs, so-called non-coding RNAs are not translated into proteins themself but can interact with mRNAs and RBPs. Of these, ribosomal RNAs (rRNA) and transfer RNAs (tRNA) are major components of the translation machinery [Rodnina, 2023]. Other non-coding RNAs often exhibit regulatory functions by recognising certain sequence elements in mRNAs or DNA by sequence homology and sequestering specific RBPs to these locations. They include small nuclear RNAs (snRNA) that are for example part of the spliceosome, small nucleolar RNAs (snoRNAs) that play a role in the positioning of post-transcriptional modifications, micro RNAs (miRNAs) and small interfering RNAs (siRNAs) which both take part in post transcriptional gene silencing via RNA decay or translational repression and long non-coding RNAs (lncRNA) [Costa, 2005].

2.1.1 The lifecycle of mRNAs

In animal cells, RNA is transcribed from DNA in the nucleus of the cell. This process is initiated by a group of DNA-binding proteins known as transcription factors. During and after transcription, nascent mRNAs (pre-mRNAs) undergo processing and editing to produce mature mRNAs (**Figure 2.1**). These processing steps encompass 5' mRNA capping, post-transcriptional editing and modifications like pseudouridylation and m6A modification, splicing, 3'-end processing, and polyadenylation [Proudfoot *et al.*, 2002, Arzumanian *et al.*, 2022, Delaunay *et al.*, 2023].

The processed mature mRNA is then exported to the cytoplasm via dedicated export mechanisms [Köhler and Hurt, 2007, Björk and Wieslander, 2017] where mRNA translation takes place. To ensure accurate RNA production, several RNA quality control mechanisms are employed that initiate RNA degradation when errors in transcription or processing are detected [Doma and Parker, 2007, Lykke-Andersen and Jensen, 2015]. Even correctly processed RNA is degraded after a certain number of rounds of translations, and this regular turnover of mRNAs is thought to play a crucial role in regulating RNA expression [Houseley and Tollervey, 2009].

All these processes of the RNA lifecycle are executed by a vast number of RBPs that assemble with the processed RNAs into ribonucleoprotein complexes (RNP complexes). Therefore, RBPs play a crucial role in regulating gene expression by shaping and control-

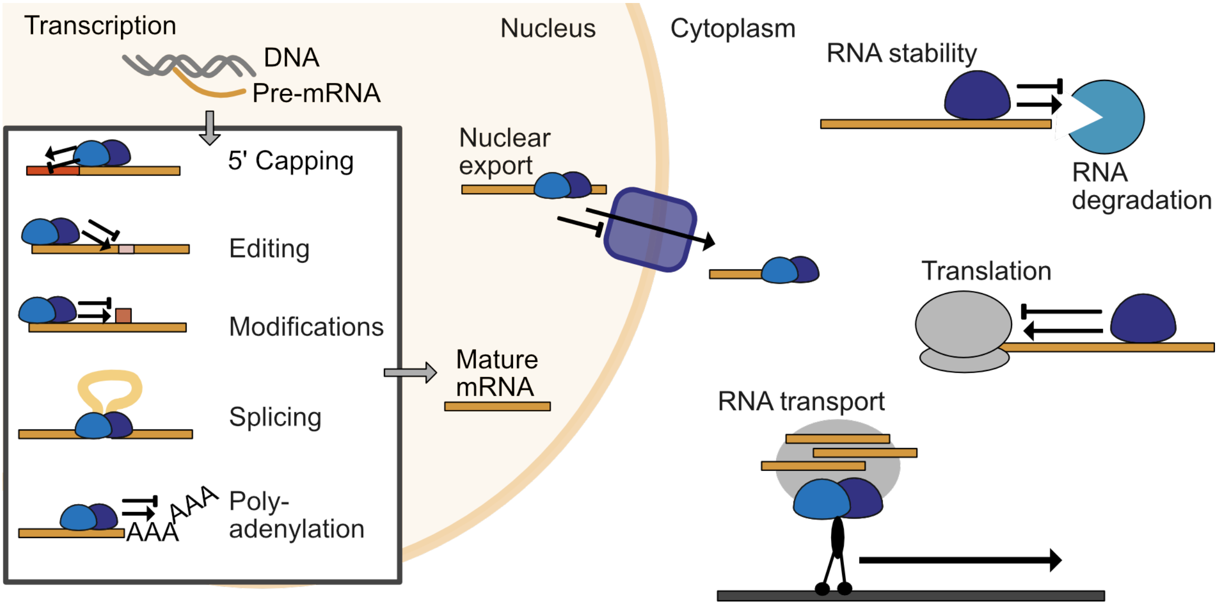


Figure 2.1: Overview of the lifecycle of RNAs. Pre-mRNAs are transcribed from DNA in the nucleus of the cell. Processing of pre-mRNA into mature RNA already starts during transcription and includes 5' capping, RNA editing, RNA modifications, splicing and polyadenylation. Mature mRNAs are then exported to the cytoplasm where they are translated and/or degraded. RNA degradation can be postponed by RNA-stabilising factors or promoted by RNA-destabilising factors. RNAs can also be transported through the cytoplasm to specific locations. This graphic is simplified from [Gerstberger *et al.*, 2014].

ling the number of mature mRNAs in the cell, based on the current cellular requirements [Dreyfuss *et al.*, 2002, Gerstberger *et al.*, 2014, Singh *et al.*, 2015].

In recent years, there has been increased attention towards an additional aspect of mRNA regulation, specifically the localisation of mRNA within the cell. An mRNA's cellular localisation will strongly impact which RBPs the RNA will interact with, thus introducing an extra layer of information to the mRNA lifecycle regulation [Das *et al.*, 2021] (see also **Chapter 2.1.3 & 2.2.3**).

In summary, the route from an immature mRNA to a protein is an intricate process with multiple stages. The entire process heavily relies on RBPs and their interactions in RNPs. This makes RBPs an intriguing subject to study.

2.1.2 RNA-binding proteins

Human cells contain numerous RBPs as exemplified by 1,542 human RBPs reported by a previous survey [Gerstberger *et al.*, 2014]. Often, they exhibit higher expression levels compared to average cellular proteins and they are evolutionarily highly conserved, particularly in their RNA-binding domains.

Classical RNA-binding domains include RNA recognition motif (RRM) domains which are found in 225 human RBPs, DEAD domains which are found in around 80 RBPs and KH domains which are found in around 40 RBPs [Gerstberger *et al.*, 2013, Gerstberger *et al.*, 2014]. These domains mainly recognise their RNA targets by sequence-specific binding to short sequence motifs (5-9 nt) in the RNA [Ray *et al.*, 2013], but domains recognising RNA secondary structures like the double-stranded RNA binding domain also exist [Masliah *et al.*, 2013].

As highlighted by [Lunde *et al.*, 2007], the low diversity of RNA-binding domains contrasts with the precise targeting necessary for RBP functionality. Therefore, RBPs employ multiple other strategies to achieve specificity, such as possessing more than one RNA-binding domain of the same or different types. For instance, U2AF2 (U2 Small Nuclear RNA Auxiliary Factor 2) contains an R/S domain followed by three RRM domains. In addition, RBPs can attain specificity through their partner proteins within an RNP complex, and the cellular localisation of the RBP will determine which RNAs it can bind.

Interestingly, an experimental approach called RNA Interactome Capture (RIC) uses UV crosslinking and enrichment of polyadenylated RNAs followed by mass spectrometry analysis developed to identify novel RBPs. RIC has been used by multiple research groups [Castello *et al.*, 2012, Baltz *et al.*, 2012], discovering that about half of the RNA-interacting proteins did not contain any known RBDs, and hundreds had not been described in the context RNA biology before [Hentze *et al.*, 2018]. These findings significantly expanded the range of RBPs and their functions and highlight the importance of studying RBPs without classical RNA-binding domains.

2.1.3 Subcellular compartments formed by liquid-liquid phase-separation

Membrane-less organelles (MLOs) are biomolecular condensates that form in the absence of a lipid membrane. They arise via a process known as liquid-liquid phase separation (LLPS), where the interactions of key proteins and RNAs lead to the formation of droplets that possess liquid or gel-like features. The creation of MLOs is facilitated by specific proteins, which typically use intrinsically disordered regions for LLPS. Other mechanisms for the formation of a particular MLO include the dimerisation of proteins or interactions with other proteins or nucleic acids, both of which can induce phase separation. LLPS-prone proteins are crucial to the formation of MLOs, but additional proteins that cannot undergo LLPS themselves are also incorporated into the scaffold, contributing to the final architecture or function of the MLO [Langdon and Gladfelter, 2018, Antifeeva *et al.*, 2022, Hirose *et al.*, 2023].

MLOs play an essential role in regulating several stages of gene expression, both within the nucleus, involving transcription, RNA processing, and export, and within the cytoplasm, influencing RNA stability and translation among others [Spector, 2006]. They achieve this by recruiting enzymes and their substrates to establish a biochemical reaction space with high concentrations of required components [Shin and Brangwynne, 2017, Ninomiya *et al.*, 2020]. In alternative settings, they operate as ‘sponges‘, constraining the function of specific processes [Hirose *et al.*, 2014, Imamura *et al.*, 2014].

I became specifically interested in two types of cytoplasmic MLOs, namely stress granules and processing bodies (P-bodies). Stress granules, measuring 0.1-2 μm, form in response to cellular stressors such as chemicals, UV light, heat and oxygen radicals, viral infection and nutrient deprivation [Riggs *et al.*, 2020, Kedersha *et al.*, 1999, Spector, 2006] and are of significant interest due to their role in the cellular response to these stressors. Stress granules can rapidly assemble and disassemble, which is suggested to be triggered by a global halt in protein translation. This leads to an accumulation of stalled pre-initiation complexes, which ultimately combine with untranslated mRNAs and various proteins to create stress granules. This process is believed to help maintain RNA integrity during stressful conditions [Panas *et al.*, 2016]. Stress granules contain stalled 48S preinitiation complexes consisting of translationally arrested mRNAs, small ribosomal subunits, translation initiation factors, and RNA-binding proteins like PABP (PolyA-binding protein), G3BP (Ras GTPase-Activating Protein-Binding Protein) and TIA1 (T-Cell-Restricted Intracellular Antigen 1) [Guzikowski *et al.*, 2019]. Recent studies have linked the dysregulation of stress granule formation and the consolidation of stress granules to numerous disease contexts, including cancer, ageing, viral infection, and neuronal disorders such as amyotrophic lateral sclerosis (ALS)/frontotemporal dementia (FTD), Huntington's, and Alzheimer's [Panas *et al.*, 2012, Alberti and Hyman, 2021, Liu-Yesucevitz *et al.*, 2010].

P-bodies are 0.1-1.0 μm granules that are constitutive in most cells and can increase in size under stress conditions [Kedersha *et al.*, 2005, Spector, 2006, Kedersha and Anderson, 2007]. Although they contain RNA decay factors, including DCP1A (mRNA-decapping enzyme 1A) and DCP2A (mRNA-decapping enzyme 2A), P-bodies are not required for mRNA decay [Eulalio *et al.*, 2007, Horvathova *et al.*, 2017]. At the same time, they are enriched in factors associated with RNA stabilisation or mRNA silencing [Hubstenberger *et al.*, 2017]. Therefore, it remains controversial whether RNA decay takes place inside P-bodies or whether they act as RNA storages [Decker and Parker, 2012, Hubstenberger *et al.*, 2017]. Of note, only very few proteins like DDX6 (DEAD-Box Helicase 6), LSM14A (LSM14A mRNA Processing Body Assembly Factor) and EIF4ENIF1 (Eukaryotic Translation Initiation Factor 4E Transporter; also known as 4E-T) appear to be essential for P-body assembly [Ohn *et al.*, 2008, Ayache *et al.*, 2015]. Similar to stress granules, P-bodies have been linked to viral infection and cancer [Alberti and Hyman, 2021, Kanakamani *et al.*, 2021].

P-bodies and stress granules share a considerable amount of proteins and RNAs [Kedersha and Anderson, 2007, Youn *et al.*, 2019], while at the same time some proteins are exclusively found in P-bodies or stress granules. Interestingly, it was proposed that both granule types are dynamically linked and that a docking mechanism of those granules might facilitate the exchange of RNAs [Kedersha *et al.*, 2005, Kedersha and Anderson, 2007]. Indeed, the stress granules and P-bodies physically associate with one another *in vivo*, for example in arsenite treated human U2OS cells [Sanders *et al.*, 2020]. The authors of this study performed complex network analyses on the formation of stress granules and their interaction with P-bodies and actually suggest to consider P-bodies and stress granules rather as a dynamic spectrum of condensates instead of using the binary classification of stress granules versus P-bodies. However, there will be more studies needed to confirm direct interactions of both granule types. In conclusion, two closely related MLOs formed by LLPS in the cytoplasm are P-bodies and stress granules. They organise the processing of mature mRNAs, including translation, degradation and RNA silencing in the cytoplasm, although their role in these processes is not fully understood mechanistically.

2.2 The protein PURA and its connection to neuronal diseases

2.2.1 The protein PURA

The protein PURA (Purine-rich binding element alpha, also known as PUR-alpha or PUR-α) is a multifaceted DNA- and RNA-binding protein [Molitor *et al.*, 2021]. In human tissues, the expression of *PURA* RNA is ubiquitous but relatively low (TPM <

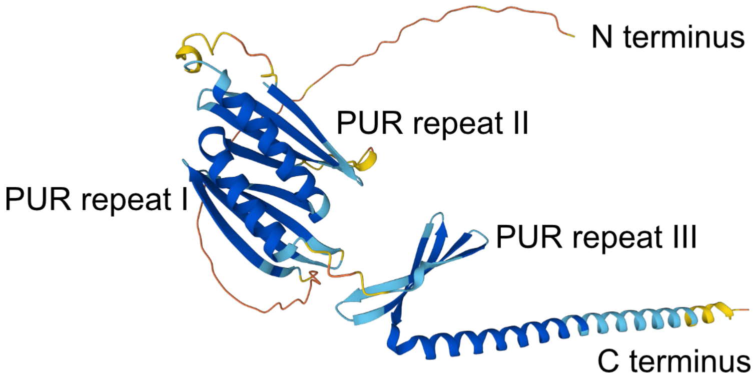


Figure 2.2: Structure of the PURA protein as predicted by AlphaFold [Jumper *et al.*, 2021]. PUR repeat I and II form a DNA or RNA-binding domain. PUR repeat III and the C-terminal alpha helix are proposed to mediate dimerisation with a second PURA protein [Weber *et al.*, 2016]

.

15, TPM - transcript per million) across all adult tissues, as evident from data provided by the Genotype-Tissue Expression (GTEx) project [Lonsdale *et al.*, 2013].

PURA orthologs are prevalent in various domains of life, including human [Bergemann *et al.*, 1992], mouse (*Mus musculus*) [Ma *et al.*, 1994], fruit fly (*Drosophila melanogaster*) [Graebsch *et al.*, 2009], zebrafish (*Danio rerio*) [Penberthy *et al.*, 2004], and silkworm (*Caenorhabditis elegans*) [Witze *et al.*, 2009]. Notably, the sequence of the *PURA* gene exhibits remarkable conservation among mammals, with the mouse *PURA* ortholog sharing a sequence identity of over 99% with human *PURA* [Ma *et al.*, 1994].

PURA is a member of the PUR-protein family, alongside the equally highly conserved paralogs PURB (Purine-rich binding element beta) and PURG (Purine-rich binding element gamma) [Johnson *et al.*, 2013, Janowski and Niessing, 2020]. A distinctive feature of the PUR-proteins is the presence of three PUR repeats, which are sequence-conserved across all PUR-proteins [Graebsch *et al.*, 2009]. The PURA protein additionally possesses a distinctive N-terminus and an intrinsically disordered C-terminus, which can be visualized using protein structure models like AlphaFold (**Figure 2.2**). The PUR repeats I and II together form one PUR domain, which has been shown to bind DNA and RNA with equal affinity *in vitro* [Weber *et al.*, 2016], while the third PUR repeat mediates dimerisation with a second PUR-protein. When interacting with double-stranded DNA, PURA exhibits the capacity to unwind it [Weber *et al.*, 2016].

In terms of cellular functionality, PURA was initially identified as a transcription factor binding to the purine-rich region upstream of the *MYC* (*V-Myc Avian Myelocytomatosis Viral Oncogene Homolog*) gene [Bergemann *et al.*, 1992]. More recent studies have expanded its role by describing its involvement in neuronal transport complexes [Ohashi *et al.*, 2002, Kanai *et al.*, 2004] and stress granules [Daigle *et al.*, 2016, Markmiller *et al.*, 2018]. Additionally, PURA has been implicated in various disease contexts, as will be explored in detail below. In summary, PURA is a conserved DNA- and RNA-binding protein with three PUR repeats. There have been reports of PURA binding RNA in very specific contexts but a comprehensive understanding of PURA's global RNA-binding capabilities is missing and remains a compelling area of investigation.

2.2.2 PURA Syndrome

PURA Syndrome was discovered as part of the Deciphering Developmental Disorders (DDD) study (http://www.ddduk.org) [Wright *et al.*, 2014]. This study included genome-wide approaches such as whole-exome sequencing and extensive phenotyping of 12,000 family trios of children with undiagnosed developmental disorders. In total, the DDD study found more than 215 likely diagnostic, non-inherited (*de novo*) mutations. These included *de novo* mutations in the *PURA* gene [Hunt *et al.*, 2014] that were subsequently associated with PURA Syndrome (**Figure 2.3 A**) [Reijnders *et al.*, 2018]. PURA Syndrome is a rare disease. Its phenotypes include abnormal neuronal development, developmental delay in movement and language, with many patients remaining nonverbal, muscle atrophy, regular epileptic, neonatal hypotonia, respiratory and feeding difficulties, and mild facial dysplasia (**Figure 2.3 B**). To date, more than 150 cases of PURA

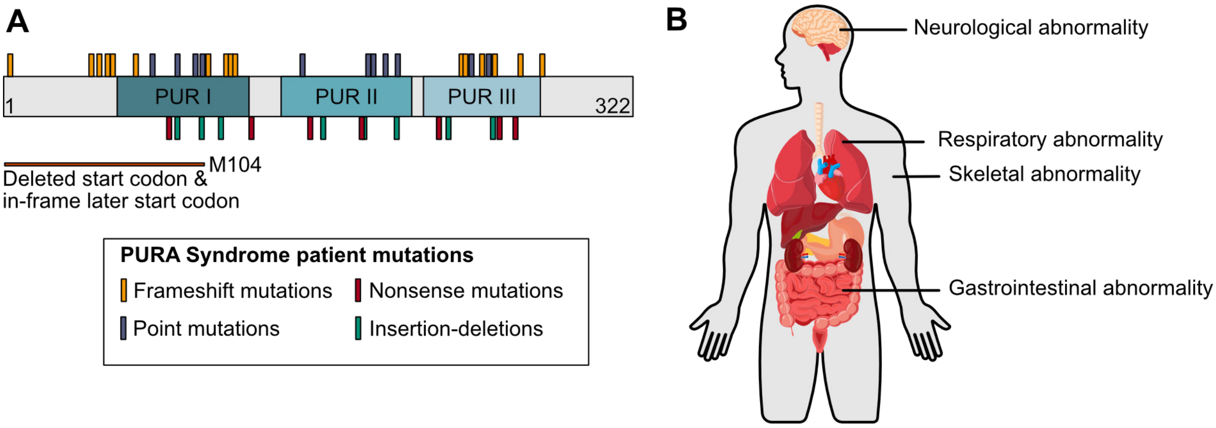


Figure 2.3: Overview of PURA Syndrome.
(A) Types and locations of *de novo* PURA patient mutations are depicted on the PURA protein. Numbers give the first and last amino acid positions. Orange - Frameshift mutations, red - nonsense mutations, blue - point mutations, green - insertion-deletions. **(B)** PURA Syndrome is a neurodevelopmental disorder, but also includes respiratory, skeletal and gastrointestinal abnormalities.
This figure was inspired by [Molitor *et al.*, 2021].

Syndrome have been described clinically [Reijnders *et al.*, 2018, Dai *et al.*, 2023].

In PURA Syndrome, each *PURA* mutation occurs *de novo*, resulting in a wide variety of *PURA* mutations distributed along the entire PURA protein (**Figure 2.3 A**). These include frameshift, nonsense and missense mutations that are predicted to disrupt the PURA folding structure or DNA/RNA interaction sites. The different mutations are unrelated to the severity of the phenotype, with no strong genotype-phenotype correlation [Reijnders *et al.*, 2018]. For example, five patients with the same mutation had a high variance in phenotypic symptoms, with one in five being able to walk and two in five without epilepsy. It was therefore proposed that all *PURA* mutations lead to a dysfunctional PURA protein and that PURA Syndrome is caused by a heterozygous loss of functional PURA protein (haploinsufficiency).

A related condition is the 5q31.3 Microdeletion Syndrome, which is caused by a microdeletion involving the *PURA* gene. The shortest region (101 kb) of the 5q31.3 microdeletion has been reported to involve only three genes: *PURA*, *IGIP* and *CYSTM1* [Brown *et al.*, 2013]. In a comparison of 157 PURA Syndrome patients with 11 5q31.3 Microdeletion patients, no phenotypic differences were found except for slightly more brain abnormalities in the 5q31.3 Microdeletion patients [Dai *et al.*, 2023], strengthening the idea that both syndromes are caused by PURA haploinsufficiency. Therefore, PURA Syndrome and 5q31.3 microdeletion syndrome are summarised under the term PURA-related neurodevelopmental disorders (https://purasyndrome.org/research/).

There have been several approaches to create mice with heterozygous [Khalili *et al.*, 2003, Johnson *et al.*, 2006, Barbe *et al.*, 2016] and homozygous [Khalili *et al.*, 2003, Hokkanen *et al.*, 2011] PURA deficiency that are reviewed in [Molitor *et al.*, 2021]. In short, mice with a homozygous loss of PURA were normal at birth, but developed severe tremor and mobility problems as well as spontaneous seizures and died before half a year after birth. Mice with a heterozygous loss of PURA showed only mild phenotypes but displayed a reduced number of neurons and dendrites. In conclusion, the impact of PURA loss on mouse and human development suggests an important cellular role for PURA.

2.2.3 RNA transport along neurons

As described above, PURA has a strong connection to neuronal functionalities. Here, I briefly want to introduce the concept of RNA transport in neurons and the role of PURA in it.

Neurons are the central building blocks of the nervous system, transmitting and processing electrical and chemical signals to integrate, store, and transmit information, enabling sensory perception, motor control, and complex cognitive functions. To serve this purpose, a neuron has a very specific shape, consisting of a cell body from which a single axon and multiple dendrites emerge [Craig and Banker, 1994]. The contacts between neurons are established by synapses, specific junctions that connect dendrites of one neuron to the axons of other neurons, allowing for a chemical transmission of the information signals

[Maynard *et al.*, 2023]. The amount and strength of synapses undergo constant change due to experiences through a process called synaptic plasticity.

With axons of up to hundreds of centimetres in length and dendrites that can measure more than 10 millimetres [Ishizuka *et al.*, 1995], neurons are very large in comparison to other eukaryotic cells. However, this size also leads to a special problem for neurons which is the maintenance and modification of the proteins needed in close proximity to the synapses in both axons and dendrites. Neurons solve this with local translation of proteins, requiring ribosomes and mRNAs to be transported to axon and dendrite ends [Nirschl *et al.*, 2017]. Here, I will focus on RNA transport in dendrites.

Multiple studies described the dendritic transcriptome by dissecting neurons into dendrites and cell bodies and performing RNA sequencing (RNAseq) on both parts separately [Wright *et al.*, 2014]. These approaches identified thousands of dendritic mRNAs including mRNAs of translation factors, cytoskeleton components, different receptor types and ion channels [Holt *et al.*, 2019]. Dendritic transport of mRNAs has been shown to depend on motor proteins, especially kinesins, transporting cargo along the microtubular cytoskeleton [Bassell *et al.*, 1998, Nirschl *et al.*, 2017]. Interestingly, PURA and its paralog PURB have been found to link *Bc1* RNA (Brain cytoplasmatic RNA 1) to mictrotubules by binding to its 5'UTR in mouse neurons and PURA controls i*Bc1* distribution within dendrites [Ohashi *et al.*, 2000]. Furthermore, PURA and PURB were found in mouse dentritic RNA transport granules containing the RNAs *Camk2* (Calcium/calmodulin dependent protein kinase II, also known as CaMKII) and *Arc* (Acrosin) that were moved along microtubuli by KIF5 (Kinesin Family Member 5) together with DDX3 (DEAD-Box Helicase 3), STAU1 (Staufen double-stranded RNA binding protein 1), FMR1 (Fragile X Messenger Ribonucleoprotein 1) and FXR1 (FMR1 Autosomal Homolog 1) among others. The distribution of the *Camk2* RNA was suppressed by the loss of PURA and STAU1, but not DDX3 and SYNCRIP [Kanai *et al.*, 2004]. Taken together, these studies suggest a role of PURA in the microtubuli depended dendritic RNA transport.

2.2.4 Neurodegenerative diseases connected to PURA

In addition to the described PURA-related neurodevelopmental disorders, PURA has been linked to the repeat expansion disorders Amyotrophic Lateral Sclerosis (ALS), Frontotemporal Dementia (FTD), and Fragile X-Associated Tremor/Ataxia Syndrome (FXTAS), all of which lead to the development of late-onset neurological disorders [Molitor *et al.*, 2021]. Both ALS and FTD are phenotypically distinct diseases that arise from a common genotype of mutations in one of several genes including *C9ORF72* (*Chromosome 9 Open Reading Frame 72*), *SOD1* (*Superoxid Dismutase-1*), *FUS* (*FUS RNA Binding Protein*) and *TARDBP* (*TAR DNA-Binding Protein*). [DeJesus-Hernandez *et al.*, 2011, Renton *et al.*, 2011, Goutman *et al.*, 2022]. FTD is the second most common cause of presenile dementia, characterised by degeneration of the frontal and temporal lobes of the brain [Graff-Radford and Woodruff, 2007]. ALS, the third most common neurodegenerative disease in the Western world [Hirtz *et al.*, 2007], leads to progressive paralysis and fatal respiratory failure [Mead *et al.*, 2023]. FXTAS is a late-onset disease caused by 55-200 CGG-repeats in the 5'UTR of the *FMR1* gene [Hagerman and Hagerman, 2016, Salcedo-Arellano *et al.*, 2020], that causes tremor, ataxia and dementia.

A hallmark of C9ORF72 ALS/FTD are RNA foci containing G4C2-repeat RNAs together with several proteins. It was shown that PURA is recruited into these foci [Donnelly *et al.*, 2013, Mizielinska *et al.*, 2013]. Similarly, in FXTAS CGG-repeat expansion RNAs form neuronal foci that also contain PURA among other proteins [Sofola *et al.*, 2007]. In both cases it remains to be shown, whether PURA localisation to these foci is caused by PURA binding to the repeat-expansion RNAs. Interestingly, in model organisms, neurotoxicity induced by overexpression of G4C2-repeat RNAs or CGG-repeat RNAs can be overcome by simultaneous ectopic overexpression of PURA [Jin *et al.*, 2007, Xu *et al.*, 2013, Boivin *et al.*, 2018, Swinnen *et al.*, 2020], stressing the importance of PURA for neuronal functioning.

2.3 High-throughput experiments as a lens on RNAs and their interaction with RNA-binding proteins

The development of high-throughput experiments like high-throughput sequencing and proteomics has added a new dimension to how biological questions can be asked. Instead of studying a few specific factors in a given context, high-throughput methods allow the observation of a specific effect on thousands of RNAs or proteins in parallel (**Figure 2.4**). This also gave rise to the need for in-depth computational analyses for the high amounts of data being produced. Here, I will shortly introduce several high-throughput methods that are used or mentioned throughout this thesis and then focus on UV crosslinking and immunoprecipitation (CLIP) experiments.

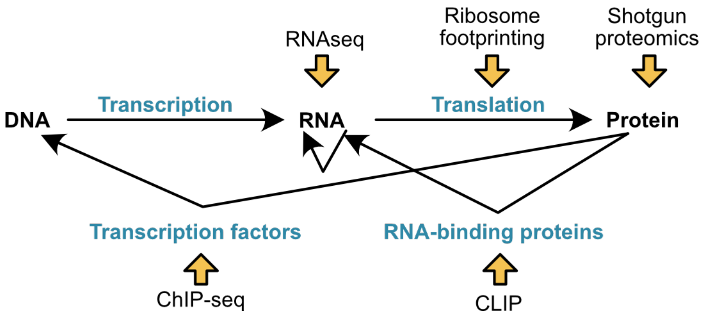


Figure 2.4: High-throughput experiments allow to globally elucidate different stages in the interactions between DNA, RNA and proteins. RNAseq queries the RNA expression level and shotgun proteomics the protein expression levels. Ribosome footprinting reveals RNAs that are actively translated. CLIP is used to study the binding of proteins to RNA (RNA-binding proteins) and chromatin immunoprecipitation sequencing (ChIP-seq) is the binding of proteins to DNA (transcription factors).

2.3.1 An overview of high-throughput experiments used to query the RNA lifecycle

Sequencing of DNA was first developed by Frederick Sanger [Sanger *et al.*, 1977]. In brief, the sequence of a single-stranded DNA can be detected by subsequent pairing of the complementary nucleotide. This complementary nucleotide is marked radioactively or fluorescently for detection. Next-generation sequencing is a further development of this method that allows the sequencing of millions of DNA fragments in parallel [Bentley *et al.*, 2008]. For this, the sequence fragments are immobilized on a flow cell, and the pairing of complementary nucleotides is performed on all fragments simultaneously. High-resolution imaging methods then allow the detection of all sequences by the spatial distinctness of fluorescent signals.

Next-generation sequencing can be used for various types of sequencing experiments, like RNA sequencing (RNAseq) [Wang *et al.*, 2009], ribosome footprinting [Ingolia, 2016], chromatin immunoprecipitation sequencing (ChIP-seq) [Mikkelsen *et al.*, 2007] and UV crosslinking and immunoprecipitation (CLIP) [Ule *et al.*, 2003]. RNAseq can be utilized to quantify the abundance of all present RNAs under a specific condition and to query differences in RNA abundance between two or more conditions. Moreover, it can be used to analyse for example alternative splicing behaviour or alternative polyadenylation between conditions. In ribosome footprinting, only RNA fragments that were shielded from RNA digestion by the ribosome are sequenced, which allows the mapping of ribosome positions on mRNAs to decipher active translation sites. ChIP-seq is a method used to analyse interactions between specific proteins and DNA. In short, the proteins are crosslinked to their DNA-binding positions and then immunoprecipitated, which allows the sequencing of bound DNA fragments providing information on the protein's binding sites across the genome.

In contrast, shotgun proteomics is based on high-throughput mass spectrometry and involves the digestion of complex protein mixtures into peptides that are then identified by mass spectrometry. The global protein abundances can then be inferred from the abundances of peptides belonging to the proteins [Zhang *et al.*, 2013]. Taken together, these techniques collectively contribute to our understanding of gene expression, translation, and chromatin interactions, offering critical tools for unravelling the complexities of different stages in the lifecycle of RNAs.

2.3.2 UV crosslinking and immunoprecipitation (CLIP) experiments

UV crosslinking and immunoprecipitation (CLIP) is the method of choice when querying global RNA-binding of an RBP of interest. In this method, RBPs are covalently crosslinked to their bound RNAs by UV irradiation *in vivo*. Afterwards, RNAs are fragmented and the RBP of interest is immunoprecipitated with a specific antibody. This will extract both the RBP and the fragment of the bound RNA. The crosslinked proteins are then digested and the fragmented RNAs are subjected to reverse transcription and polymerase chain reaction (PCR) amplification. The so-obtained complementary DNA (cDNA) library is sequenced by Sanger sequencing in the original CLIP protocol [Ule *et al.*, 2003, Ule *et al.*, 2005]. Then, the pileup of the RNA sequence fragments is used to extract the region that is bound and crosslinked to the RBP of interest.

From this original method over 28 CLIP experiment types have been evolved [Lee and Ule, 2018]. As a first important improvement high-throughput sequencing was introduced by 'High-throughput sequencing of RNA isolated by CLIP' (HITS-CLIP) [Licatalosi *et al.*, 2008, Chi *et al.*, 2009] and 'CLIP coupled with high-throughput sequencing' (CLIP-seq) [Yeo *et al.*, 2009]. An interesting characteristic of CLIP protocols is that some residual amino acids remain covalently attached to the RNA at the crosslink position [Lee and Ule, 2018]. This will lead on the one hand to a termination of the reverse transcrip-

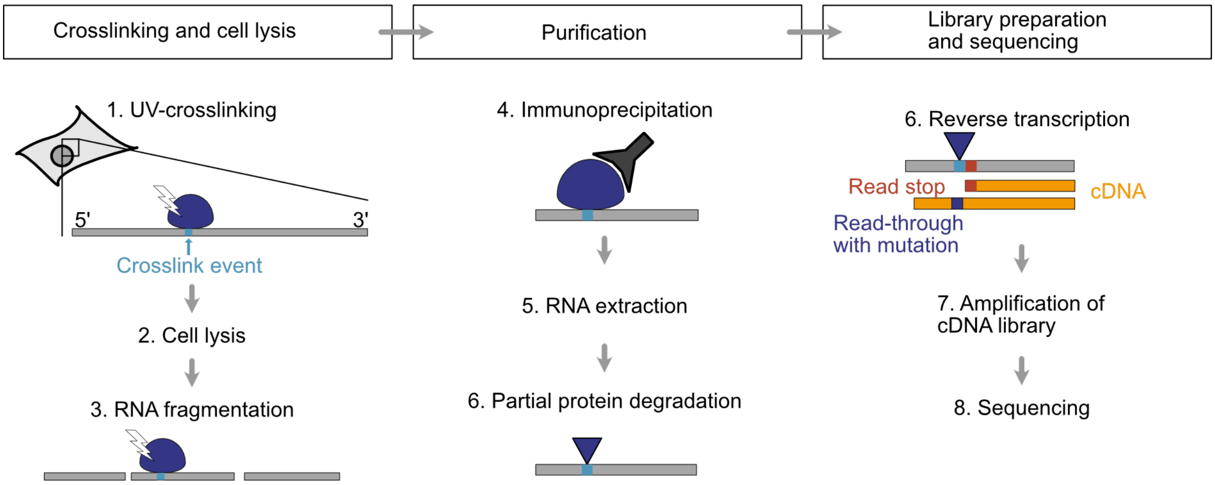


Figure 2.5: Schema of CLIP experiments. CLIP experiments consist of three major steps (Top), namely crosslinking and cell lysis, purification and library preparation and sequencing. In the first part, proteins are cross-linked to their bound RNAs by UV irradiation (1.). Then, cells are lysed (2.) and RNAs are fragmented (3.). In the second part, the RNA-binding protein is immunoprecipitated (4.), then RNAs are extracted (5.) and the crosslinked proteins are partially degraded (6.). In the third part, RNA fragments are reverse-transcribed into cDNA. The reverse transcription will stop one nucleotide before the crosslinked position (red) or might read through the crosslink incorporating a mutation at the crosslinked position (blue). The cDNA will then be amplified (7.) and sequenced (8.). Blue - protein or peptide, grey - RNA, light blue crosslinked position, red - read stop, orange - cDNA. This graphic is inspired by [Hafner *et al.*, 2021].

tase one nucleotide before the UV-crosslink (read stop). On the other hand, when the reverse transcriptase reads through the crosslinked nucleotide (read-through) mutations in the cDNA can occur (**Figure 2.5**). Following HITS-CLIP and CLIP-seq, two experiment types were developed that exploit either the read-through mutation or the read stops to enhance the resolution of the obtained binding patterns. One is ‘Photoactivable ribonucleoside-enhanced CLIP‘ (PAR-CLIP), which introduces the treatment of cells with 4-thiouridine before UV crosslinking to enhance the crosslinking efficiency and uses C-to-T transition analysis on read-through sequences to determine the exact position of crosslinks [Hafner *et al.*, 2010]. While the mutations can point directly to the crosslinked nucleotide, they only occur on uridines and might therefore bias the analysis towards uridine crosslinks.

The CLIP type ‘individual-nucleotide resolution CLIP‘ (iCLIP) adds the second adapter for PCR amplification only after reverse transcription at the 5’-end of the cDNA, which allows the extraction of the read stop position and consequently the crosslink position one nucleotide upstream of the read stop position [König *et al.*, 2010]. An advantage of iCLIP in contrast to PAR-CLIP is that read stops are by far more prevalent than read-through [Urlaub *et al.*, 2002]. This leads to a higher signal depth of iCLIP in contrast to PAR-CLIP at the same read depth. As UV crosslinking in general works best on uridines, iCLIP like all CLIP methods retains some bias towards uridines [Sugimoto *et al.*, 2012]. However, iCLIP in contrast to PAR-CLIP is not limited to the detection of crosslinks on uridines, reducing the bias of the method.

In the following years, various CLIP protocols have been evolved that reduce the complexity of the experiment and improve library generation [Lee and Ule, 2018, Hafner *et al.*, 2021] namely ‘enhanced CLIP‘ (eCLIP) [Van Nostrand *et al.*, 2016] and iCLIP2 [Buchbender *et al.*, 2020]. Both experiment types reduce the complexity of the initial iCLIP experiment as can be seen from the reduction of the experimental time from six days for iCLIP to four days for eCLIP and iCLIP2. At the same time, they improve library generation leading to a higher signal depth. The experimental differences behind this are described in several publications [Hafner *et al.*, 2010, Buchbender *et al.*, 2020].

From the data analysis point of view, a main difference between eCLIP to iCLIP is that it uses a paired-end library strategy, meaning that the cDNA is sequenced both from 3’-end to 5’-end (read #1) and from 5’-end to 3’-end (read #2). In this setup, read #2 contains the read stop information. iCLIP2 keeps the single-read sequencing approach from the original iCLIP experiment. Furthermore, eCLIP introduced the usage of size-matched input control (SMInput) samples as a control for unspecific background signal. SMInput samples are treated exactly like the eCLIP samples but no immunoprecipitation is performed. Therefore it will contain crosslinks of all RBPs with a similar size as the RBP of interest, including also crosslinks of the RBP of interest itself. The idea of the authors was that the usage of SMInput samples improves the quantification of binding sites from the eCLIP data [Van Nostrand *et al.*, 2016] and many computational approaches to analyse CLIP data are based on the usage of an SMInput sample or other background reads as discussed later in this dissertation (**Chapter 3.7.4**). However, whether the SMInput samples improves the analysis of binding sites is under debate in the field [Chakrabarti *et al.*, 2018, Lee and Ule, 2018]. iCLIP2 like iCLIP is usually performed without SMInput samples.

In summary, while all CLIP experiments in theory produce both read stops and read-through mutations, the main difference in CLIP experiments is whether they strive to capture one or the other. In this thesis, I analyse iCLIP, iCLIP2 and eCLIP experiments based on their read stop information. These experiment types differ in their execution [Hafner *et al.*, 2021], but the main differences lie in the library preparation, while the first steps are very similar. Therefore the underlying signal in the obtained data should be very comparable when accounting for the different library architectures.

2.4 Computational methods for high-throughput experiments

2.4.1 Preprocessing and alignment of next-generation sequencing data

All next-generation sequencing data will have to undergo some basic steps of processing namely demultiplexing, quality control, trimming of adapter and barcode sequences and alignment to the respective genome (**Figure 2.6 A**). After these steps, the data will be further processed according to the requirements of the given experiment type (**Figure 2.6 B, C**).

The preprocessing should start with quality filtering of the reads. Next-generation sequencing usually provides the obtained sequences (reads) in FASTQ format [Cock *et al.*, 2010]. This format contains both the nucleotide sequence of the read and an associated quality score for each base which can be used for quality filtering. Next, the reads of different samples that were sequenced together are separated (demultiplexing). To do this, specific barcode sequences are fused to the cDNA fragments of each sample before sequencing. The barcodes are later detected in the reads to split up the samples and are subsequently trimmed off the reads. Reads where no barcode can be detected are usually discarded. Similarly, certain adapters are fused to the cDNA fragments for the sequencing process. These will also be trimmed off during adapter trimming.

The steps of quality control, demultiplexing, adapter and barcode trimming can be performed in parallel by for example the software FLEXBAR [Dodt *et al.*, 2012]. After these steps, reads are aligned to a genome assembly of the respective genome. In this thesis, gene alignment is done with the ‘spliced transcripts alignment to a reference‘ (STAR) algorithm [Dobin *et al.*, 2013]. As the name implies, STAR is a splice-aware aligner, meaning that it can align discontinuous reads to the genome. The alignment process of STAR consists of two major steps, a seed search step and a clustering, stitching and scoring step. In the seed search step, STAR searches sequentially for the maximum mappable prefix. This means it will find the longest mappable fragment of the read starting at the beginning of the read. Then, it will remove the mappable part from the beginning of the read and again search for the longest mappable fragment starting from the first base that was previously unmapped. In the clustering, stitching and scoring step, the aligned sequences will be clustered around a selected set of anchor seeds and then stitched together. The stitched read will be scored according to the number of matches, mismatches, insertions, deletions, and junction gaps. In the end, the stitched sequence alignment with the highest score is reported as the alignment of the read, which is provided in BAM format.

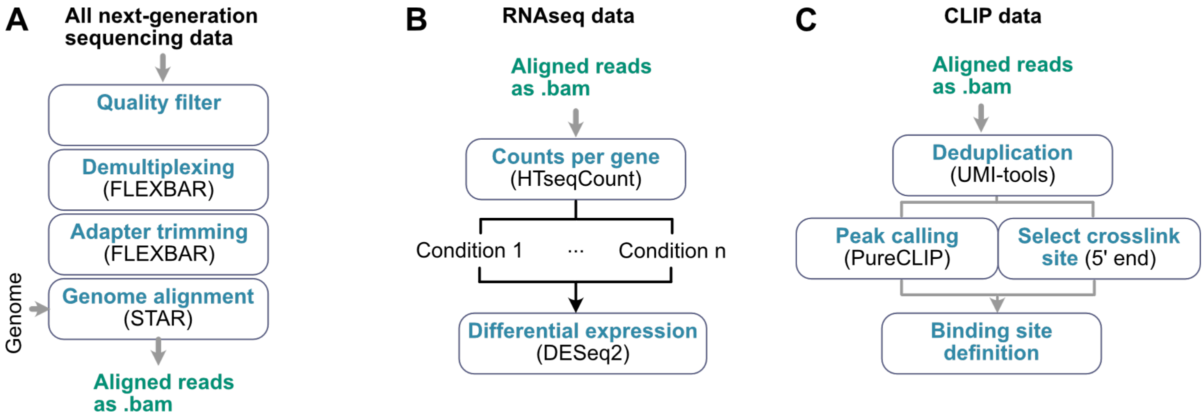


Figure 2.6: Workflows for the processing of high-throughput sequencing data.
(A) Preprocessing steps are generally performed for next-generation sequencing data.
(B) Further processing and analysis for RNAseq data.
(C) Further processing and analysis for CLIP data. Tools used in this thesis are denoted in brackets below the steps.

For RNAseq data, gene alignment will usually be followed by the quantification of reads per gene for example with HTSeq [Anders *et al.*, 2014] and differential analysis between multiple conditions for example with DESeq2 [Anders and Huber, 2010] (**Figure 2.6 B**, see **Chapter 2.4.2**). For CLIP data, genome alignment is followed by deduplication and, depending on the analysis method, crosslink extraction, peak calling and binding site definition (**Figure 2.6 C**, see **Chapter 2.4.3**).

2.4.2 Differential analysis of read count data and of protein abundance

When analysing high-throughput data, one often wants to quantify differences in expression levels between two or more conditions, like the difference between cell lines, the response to stimuli or the difference that a loss or increase of a specific factor makes in comparison to a control. In this thesis, I used DESeq2 [Anders and Huber, 2010] for differential RNA expression analysis, BindingSiteFinder, which is based on DESeq2, for differential binding analysis (https://doi.org/doi:10.18129/B9.bioc.BindingSiteFinder) and DEqMS [Zhu *et al.*, 2020] for differential proteomics analysis. In the following, I will briefly explain the steps performed by these tools.

In a nutshell, DESeq2 first performs sample-wise library normalisation to account for differences in the library depth of the samples. Next, the dispersion of read counts is calculated for genes with a similar number of read counts. This step is performed group-wise because the dispersion of lowly expressed genes will be much higher than the dispersion of highly expressed genes. Then, a generalized linear model is fitted for each group of genes with similar expression assuming a negative binomial distribution. To test for differences between the conditions, the fold changes between the given conditions are calculated and a Wald test is performed against the null hypothesis that the fold change belongs to the modelled background distribution. Finally, P values are corrected to reduce false positive results. This is done by default using by the Benjamini-Hochberg method, but other methods can be chosen.

In addition to differential gene expression analysis, the same approach can also be used for differential binding analysis as implemented in BindingSiteFinder (https://doi.org/doi:10.18129/B9.bioc.BindingSiteFinder). Instead of read counts per gene, BindingSiteFinder uses the crosslink number per binding site. Like DESeq2 BindingSiteFinder assumes a negative binomial distribution of the crosslinks. Furthermore, it uses the crosslink counts on a gene but outside of all binding sites as a background to account for gene expression. Both the crosslink number per binding site and the crosslink number in the background are included in the same generalised linear model, by adjusting the design formula used by DESeq2. BindingSiteFinder will return both the fold change per binding site and the fold change per gene-wise background, depicting both binding changes and gene expression changes between experiments.

In principle, a very similar approach is applied to proteomics data by DEqMS [Zhu *et al.*, 2020]. From the data analysis perspective, the biggest differences between proteomics and sequencing data are the following: Firstly, protein abundances are measured as continuous numbers in contrast to count data from sequencing. Secondly, proteins are detected by unique peptides, which are peptides that can be uniquely attributed to a specific protein. However, different proteins vary in their number of unique peptides, that can be used for their detection. Therefore, DEqMS assumes a normal distribution of protein abundances and uses a intensity-based moderated t-statistic which was first developed for microarray data in the limma package [Sartor *et al.*, 2006]. Moreover, DEqMS implements a method of scaling P values according to the number of unique peptides for each protein. In summary, differential analysis tools compute fold changes and test the significance of the fold changes based on the fit of the appropriate distribution function.

2.4.3 Processing and analysis of CLIP data

Analysis of CLIP data includes - in addition to the before described preprocessing steps - deduplication, crosslink extraction, peak calling and binding site definition. Deduplication removes quantitative biases introduced during cDNA amplification (**Figure 2.6 C**). To distinguish PCR duplicates from real duplicates, a unique molecular identifier (UMI) is introduced to the cDNA fragments together with the barcode sequence before PCR amplification. It marks each cDNA fragment with a unique random sequence of 5 - 11 nucleotides (nt) length depending on the protocol. The UMI sequence can be extracted from the sequenced reads together with the barcodes and is then used to remove duplicated reads arising from the same cDNA. This process is called deduplication and can be executed for example with UMI-tools [Smith *et al.*, 2017].

In this thesis, I use PureCLIP [Krakau *et al.*, 2017] to call peaks from the aligned reads and then use the called peaks and integrate them with the crosslink profiles to define RBP binding sites as proposed by [Busch *et al.*, 2020]. In brief, PureCLIP uses a non-homogeneous hidden Markov model (HMM) to define bound nucleotides while incorporating the non-specific background signal into the model to reduce the number of false positives. The HMM has a single-nucleotide resolution categorising each position either as non-enriched or enriched, indicating whether the position is enriched in read coverage. In addition, each position is categorised as non-crosslink or crosslink, indicating whether it is the position of a single-nucleotide crosslink site or not. This results in four hidden states, where positions with the hidden state ‘enriched and crosslinked‘ are defined as peaks.

In addition to peak calling, the single-nucleotide crosslink events are extracted. Then, both the PureCLIP peaks and the crosslink events are used to define equal-sized narrow RBP binding sites as described in [Busch *et al.*, 2020]. Firstly PureCLIP peaks in close proximity are merged into regions. These regions are then iteratively split up into binding sites of a predefined width (e.g. 5 nt) centred at the position with the highest score from PureCLIP. This is then repeated with the next highest position until the region is smaller than a defined minimum width (e.g 2 nt). Finally, binding sites lacking sufficient support by crosslinks are filtered out. This results in equal-sized narrow binding sites, depicting the binding behaviour of the RBP.

2.4.4 Workflow management

Due to the recent increase in the amount of biological data in need of processing and analysis, the use of standardised workflows and workflow managers have become more and more popular. A workflow consists of multiple steps that need to be executed in a given order (**Figure 2.6**) [Reiter *et al.*, 2021]. For most of the individual processing steps of biological data, specialised tools already exist that were optimised to perform their given tasks and have often been benchmarked by the scientific community. A good workflow therefore does not reimplement all the needed steps but uses these highly sophisticated tools.

Still, a typical workflow for biological data consists of five to twenty steps that need to

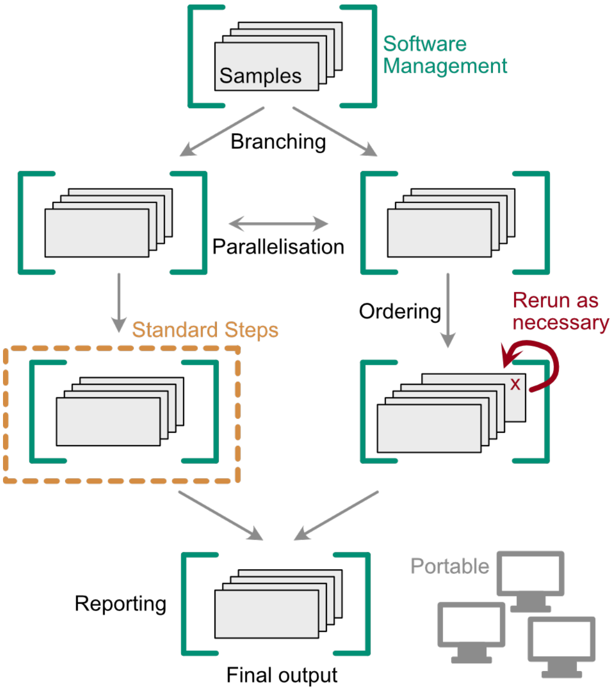


Figure 2.7: Characteristics of workflow management. The advantages of a managed workflow are shown. **Samples** - the workflow can be executed at any number of samples. **Software Management** - the needed software for each step can be managed separately. **Branching** - the output from one step can be used as input for several new steps. **Parallelisation** - the execution of multiple steps or the same step for multiple samples is automatically parallelised according to the provided resources. **Standard Steps** - the steps are performed in a standardised way. **Ordering** - the executed steps are ordered according to their dependencies on each other. **Rerun as necessary** - single steps can be rerun individually if needed. **Reporting** - statistics and command line messages for all executed steps are reported. **Portable** - the workflow is robust to be used in different computing environments. **Final output** - for the final output, the results from multiple branches can be reintegrated. This figure is inspired by [Reiter *et al.*, 2021].

be performed with different tools. Having all these steps performed by hand for each data set is an error-prone and time-consuming practice. Furthermore it can reduce the comparability between different data sets if different data sets are processed by different people with different computational environments in a slightly different way. A workflow manager can solve this by implementing an automation of all needed steps. In the field of bioinformatics, common workflow managers for the development of new research pipelines are Snakemake, [Mölder *et al.*, 2021] and Nextflow [Tommaso *et al.*, 2017].

A workflow manager performs multiple functions (**Figure 2.7**) [Reiter *et al.*, 2021, Wratten *et al.*, 2021]. To start with, it implements standardised steps for the workflow. The code for every step is specified inside a so-called rule. Each rule can be performed multiple times for multiple inputs e.g. samples. Each execution of a rule will be performed in a separate job. Workflow managers will order the execution of the jobs by tracking the input and output files of each job. Jobs can linearly depend on each other but one job might also branch into multiple downstream jobs or one job might fuse outputs from multiple other jobs. In addition, the execution of some jobs might be conditional, depending on the input or the specifications of the user.

Splitting up the workflow into jobs allows the parallelisation of multiple independent steps. The number of jobs executed in parallel is automatically scaled to the allocated resources making the whole process faster and more efficient. Another useful characteristic of managed workflows is the possibility to rerun only specific jobs of the workflow if necessary.

Each rule can be executed in a specific computational environment that contains all needed software. This resolves clashes of different software tools with different dependencies e.g. if one tool is implemented in Python 2 but others in Python 3. In addition, it makes the workflows independent of the computational environment and allows them to be used across different computing systems without changes in the code.

Taken together, the usage of workflow managers for commonly used bioinformatic workflows makes the analysis reproducible, comparable to other data sets and scalable, while reducing the possibilities for errors and the time needed for the analysis. In order to improve the usability of a given workflow, it can be implemented behind a simple command line interface.

Chapter 3

racoon_clip: an automated pipeline for iCLIP and eCLIP data processing

3.1 Motivation

In the last years, the interaction of RBPs with their target RNAs came into focus in the context of various biological research topics. To understand these interactions, many CLIP experiments are performed and a streamlined analysis of these experiments is needed. However, as researchers use different experimental versions of CLIP, they most of the time also use different computational analyses for the different experiments. Here, I hypothesised that iCLIP and eCLIP data are interchangeable - although the experiments differ - as long as they are processed the same way. I chose to process both data types to single-nucleotide crosslinks, which display an extremely high resolution of binding events in CLIP data. Furthermore, I wanted the processing to be easily usable for other researchers. To achieve these goals, I built racoon_clip an automated pipeline that can process both iCLIP and eCLIP data to single-nucleotide resolution from a single command.

3.2 Aims of this chapter

1. **Provide the same processing for iCLIP and eCLIP data:** Although there are currently some tools available to process either iCLIP or eCLIP data, most tools focus on one of the two experiment types. Therefore, I wanted to provide a tool that processes both iCLIP and eCLIP data in a comparable manner to allow downstream comparisons of experiments from both types.

2. **Provide a method to obtain single-nucleotide resolution crosslinks for eCLIP data:** While the advantages of single-nucleotide resolution of CLIP data have been clearly shown [Chakrabarti *et al.*, 2018], most eCLIP data are not processed to single-nucleotide resolution, as can be seen for example in the eCLIP processing of the ENCODE database [Van Nostrand *et al.*, 2020b]. I, therefore, wanted to provide an option to process or reprocess eCLIP data to single-nucleotide resolution. This should help improve the overall quality of CLIP data processing and the following analyses.

3. **Automate and optimise the workflow:** I aimed to provide the workflow in an automatised way to make its execution easy and reproducible. Furthermore, I wanted to provide a flexible execution structure that could accommodate variable sample numbers and compute resources, together with robust management of tool dependencies to make the workflow independent of the computational surroundings.

4. **Simplify the usage of complex CLIP processing for a broad user base:** Tools streamlining the processing of high-throughput data for a broader audience exist for many types of experiments. Easy-to-use solutions are still missing for CLIP experiments. Therefore, I aimed to build a workflow that could be run by a single command from a simple command line interface.

In concordance with the aims I described above, I built a command line tool called racoon_clip that can process iCLIP and eCLIP data to single-nucleotide resolution. In the next sections, I will first describe the methods used for the implementation of racoon_clip, then I will describe the workflow executed by racoon_clip and finally, I will show racoon_clip processing on the examples of U2AF2 iCLIP and eCLIP data.

3.3 Publication

This project has been submitted for publication as a software release manuscript as Klostermann & Zarnack - 'racoon_clip – a complete pipeline for single-nucleotide analyses of iCLIP and eCLIP data'. The manuscript is undergoing review at Bioinformatics at the time of writing this thesis.

3.4 racoon_clip - Implementation and availability

3.4.1 Tool availability and documentation

racoon_clip can be downloaded from GitHub (https://github.com/ZarnackGroup/racoon_clip) and installed via pip. It requires mamba (version ≥ 1.3.1) and Python (version ≥ 3.9). Detailed documentation of racoon_clip is also provided in the Read the Docs format at https://racoon-clip.readthedocs.io/en/latest/.

3.4.2 Used example data

I am showcasing the usage of racoon_clip on the example of U2AF2 iCLIP [Sutandy *et al.*, 2018] (SRA: SRR5646576, SRR5646577, SRR5646578) and U2AF2 eCLIP [Luo *et al.*, 2019] (ENCODE: ENCSR202BFN) data.

3.4.3 Implementation

racoon_clip is implemented as a command line tool. It runs a complete Snakemake workflow from raw read data to single-nucleotide crosslinks in one single command.

Command line interface

The command line user interface was built using a template from [Roach *et al.*, 2022]. This template is based on the Python click package to pass command line options to the Snakemake workflow and provides help messages in the command line for every option. It also passes down all Snakemake options to Snakemake, handles default options and writes a configuration file containing all user-specified and default options for later reproducibility.

Snakemake workflow

The core of racoon_clip is a Snakemake workflow (version 7.22) [Mölder *et al.*, 2021]. Most rules in the Snakemake workflow execute common command line tools that are listed here:

- FastQC (version 0.12) [Andrews, 2010]
- MultiQC (version 1.14) [Ewels *et al.*, 2016]
- FLEXBAR (version 3.5.0) [Dodt *et al.*, 2012]
- STAR (version 2.7.10) [Dobin *et al.*, 2013]
- UMItools (version 1.1.1) [Smith *et al.*, 2017]
- SAMtools (version 1.11) [Li *et al.*, 2009]
- BEDTools (version 2.30.0) [Quinlan and Hall, 2010]
- kentUtils (version 377) (https://github.com/ucscGenomeBrowser/kent)

In addition, some steps include awk (mawk version 1.3.3) implementations when no tool was available for this task.

The tools are included into racoon_clip through multiple mamba environments stored in YAML files, to prevent version clashes. All racoon_clip environments can be found at https://github.com/ZarnackGroup/racoon_clip/tree/main/racoon_clip/workflow/envs.

racoon_clip also provides an HTML report summarising the processing statistics, which is implemented as Rmarkdown (version 1.1). The necessary R packages are also provided in a mamba environment.

Documentation

The racoon_clip documentation was made with the documentation tool Sphinx and is hosted at the documentation hosting platform Read the Docs (https://racoon-clip.readthedocs.io/en/latest/). The documentation is written as reStructuredText and connected to the GitHub repository of racoon_clip. This provides searchable documentation and a new version of the documentation whenever a new version of racoon_clip is released on GitHub.

Used machines and run-time measurement

racoon_clip was implemented on the König group's high-performance computing (HPC) cluster at the IMB in Mainz, named koenighpc.imb.uni-mainz.de (x86_64 GNU/Linux, CPU: Intel(R) Xeon(R) Silver 4215 CPU @ 2.50GHz). The operation system was 'Debian GNU/Linux 10' (buster) and the shell used was the Bourne-again shell (bash, version 5.0.3). The cluster was accessed via secure socket shell (SSH). Command line tools were installed using anaconda (conda version 22.11.1) with mamba drop-in extension (version 1.3.1).

The racoon_clip tool was then tested on the Ebersberger group's compute cluster at the

Goethe University Frankfurt, named ebersberger-46-150.biologie.uni-frankfurt.de (Kernelx86_64 Linux 5.4.0-165-generic, CPU: Intel Xeon E5-2609 v4 @ 8x 1.7GHz [40.0°C]). The operating system was ‘Ubuntu 20.04 focal‘ and the shell used was bash (version 5.0.17). Access was via SSH and job management was via the simple linux utility for resource management (SLURM) workload manager. racoon_clip inherits its SLURM compatibility from Snakemake. The following additional parameters were passed to racoon_clip for SLURM execution: −−profile <path/to/slurm/profile> −−wait-for-files −−latency-wait 60.

Reported run times and RAM usage were measured with the /usr/bin/time -v utility.

3.5 Workflow

racoon_clip is a command line tool designed for complete CLIP data processing in an automated and parallelised manner through a single command. To accommodate different input data arising from diverse experimental protocols and sources, racoon_clip offers pre-set experiment types (’iCLIP’, ’iCLIP2’, ’eCLIP_5ntUMI’, ’eCLIP_10ntUMI’, ’eCLIP_ENCODE_5ntUMI’, ’eCLIP_ENCODE_10ntUMI’, or ’noBarcode_noUMI’) with the experiment_type parameter. Alternatively, a custom barcode and UMI setup can be used for data sets not covered by the pre-set options. racoon_clip requires sequencing reads in multiplexed or demultiplexed FASTQ files, along with a genome assembly in FASTA format and gene annotation in GTF format. Its customizable steps can be configured via a configuration (config) file or directly in the command line. Outputs from racoon_clip include aligned reads in BAM format, single-nucleotide crosslink events in BED and BIGWIG formats, and a detailed processing report in HTML format.

3.5.1 All steps

The racoon_clip pipeline consists of three major steps (**Figure 3.1**): (1) preprocessing of sequencing reads, (2) genomic alignment, and (3) extraction of crosslink events. In the preprocessing step, racoon_clip manages iCLIP or eCLIP experiment barcodes, accommodating various barcode formats. It demultiplexes FASTQ files, trims barcodes, and stores UMIs in read names. racoon_clip performs barcode and/or adapter trimming and/or demultiplexing using FLEXBAR [Dodt *et al.*, 2012], based on input data requirements. The UMIs are appended to the read names during barcode trimming for subsequent deduplication. If provided, barcodes need to be in a FASTA file. To ensure accurate barcode and UMI assignment, racoon_clip can perform optional quality filtering on barcode regions (quality_filter_barcodes) using awk. Custom adapters in a FASTA file or standard sequencing adapters from Illumina and eCLIP are used for adapter trimming. This preprocessing step produces demultiplexed FASTQ files containing sequencing reads for individual samples, free of adapters and barcodes but with UMIs in the read names. By default, one round of adapter trimming is performed but multiple cycles can be chosen. For example, two cycles are recommended for ENCODE eCLIP data (https://github.com/yeolab/eclip, accessed 03.12.2023). Later, the HTML report can be used to check that adapters have been trimmed off correctly.

In the second step, demultiplexed reads undergo genome alignment followed by deduplication. Reads are aligned by STAR [Dobin *et al.*, 2013] to the genome assembly (provided as FASTA) using corresponding gene annotation (provided as GTF file). STAR applies soft-clipping to the 3' end of reads (−−alignEndsType Extend5pOfRead1) to retain exact crosslink positions upstream of the 5' end. All other STAR parameters can be customised, e.g., for handling multi-mapping reads. By default, STAR outputs only unique alignments (−−outFilterMultimapNmax 1) with up to 4% mismatches (−−outFilterMismatchNoverReadLmax 0.04).

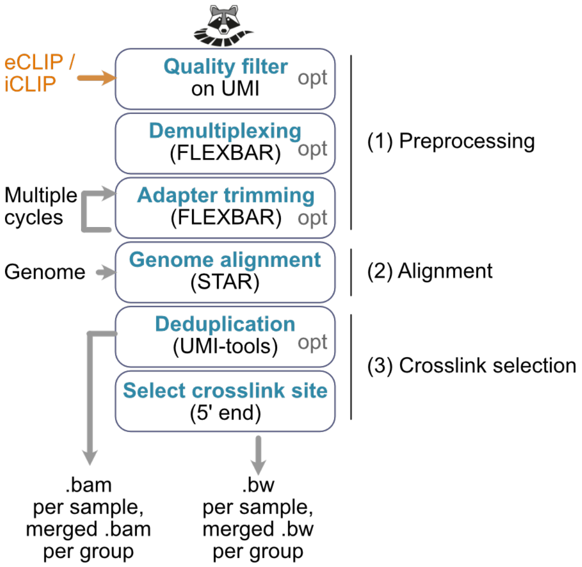

eCLIP / iCLIP
Quality filter on UMI opt
Demultiplexing (FLEXBAR) opt
Multiple cycles
Adapter trimming (FLEXBAR) opt
(1) Preprocessing
Genome
Genome alignment (STAR)
(2) Alignment
Deduplication (UMI-tools) opt
Select crosslink site (5' end)
(3) Crosslink selection
.bam per sample, merged .bam per group
.bw per sample, merged .bw per group

Figure 3.1: Schematic overview of the racoon_clip workflow to process CLIP data. Used command line tools are given in brackets. opt - optional step.

Aligned reads are then deduplicated based on identical locations and UMIs, using UMI-tools [Smith *et al.*, 2017]. While there is an option to skip deduplication (deduplicate: False), it is recommended for data quality. Finally, BAM files are indexed using SAMtools [Li *et al.*, 2009], producing aligned, deduplicated reads in BAM format.

The third step of racoon_clip involves extracting crosslink sites, selecting the position 1 nt upstream of the 5' end of aligned reads using BEDTools [Quinlan and Hall, 2010] and kentUtils (https://github.com/ucscGenomeBrowser/kent). Extracted crosslink events are provided in BED and BIGWIG formats. racoon_clip enables merging files of samples from the same condition by specifying sample groups and provides additional merged files for each sample group. In the absence of specified groups, all samples are merged by default. At the end, racoon_clip generates an HTML processing report summarising data quality using FastQC [Andrews, 2010] and MultiQC [Ewels *et al.*, 2016], along with relevant statistics for each step, facilitating user assessment of processed data set quality and the success of each workflow step.

3.5.2 Dealing with different types of input through pre-set options or customisation

To perform analyses on data from different protocols, racoon_clip offers pre-designed settings for iCLIP, iCLIP2, eCLIP with either a 5 nt or 10 nt UMI, eCLIP downloaded from the encyclopedia of DNA elements (ENCODE) and data without UMI or barcode as is the case when downloaded from the sequence read archive (SRA). The main difference for the data processing of the experiments (**Figure 3.2**) is that iCLIP uses single-end reads and, therefore, an interspaced experimental barcode that consists of the first half of the UMI, then the sample barcode and then the second half of the UMI. In contrast, eCLIP uses paired-end reads where the UMI can be found in read #2 and the barcode in read #1. Of the paired-end reads, racoon_clip only uses the read #2, which contains the read stop position. The newer versions of iCLIP (iCLIP2) and eCLIP use longer UMIs (iCLIP2 11 nt, eCLIP 10 nt) than the older protocols (iCLIP and eCLIP 5 nt). The type of the experiment can be chosen via the racoon_clip parameter experiment_type which includes the options ‘iCLIP‘, ‘iCLIP2‘, ‘eCLIP_5ntUMI‘ ‘eCLIP_10ntUMI‘, and ‘eCLIP_ENCODE_5ntUMI,‘ ‘eCLIP_ENCODE_10ntUMI‘, or ‘noBarcode_noUMI‘.

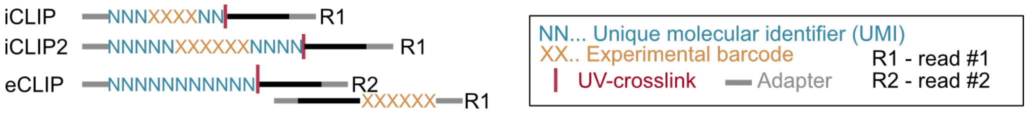


Figure 3.2: Barcode and UMI architecture used by the different CLIP experiments. Top - iCLIP; middle - iCLIP2; bottom - eCLIP; N, blue - UMI; X, orange - experimental barcode; red bar - read stop position; grey bar - adapter; black bar - read; R1 - read #1; R2 - read #2.

These pre-set options are based on a flexible architecture of experimental barcodes that can also be fully customised instead of using the pre-settings. The basic structure of experimental barcodes that racoon_clip assumes is

$$N_{umi1\_len} + X_{barcode\_len} + N_{umi2\_len} \tag{3.1}$$

where N are bases from the UMI, X are bases from the barcode and umi1_len, umi2_len and barcode_len are the racoon_clip parameters that specify the number of bases in each part. **Table 3.1** shows the racoon_clip parameters behind the different pre-settings.

	iCLIP	**iCLIP2**	**eCLIP**	**eCLIP from ENCODE**	**Removed barcode and UMI**
umi1_len	3	5	10 (5)	0	0
umi2_len	3	5	0	0	0
barcode_len	4	6	0	0	0
encode	False	False	False	True	False

Table 3.1: racoon_clip parameters behind the different pre-set options.

Of note, there are specialised options for publicly available data which are usually preprocessed. ENCODE provides preprocessed sequencing reads of eCLIP experiments without barcodes and with UMIs already stored in the read name. This specific setting is taken care of in racoon_clip with the ‘eCLIP_ENCODE_10ntUMI‘ and ‘eCLIP_ENCODE_5ntUMI‘ settings for the experiment_type parameter or by setting the parameter encode to ‘True‘. Moreover, in the SRA database iCLIP or eCLIP data are stored already demultiplexed and often with the barcode and UMI already trimmed off. In this case, the ‘noBarcode_noUMI‘ option can be used. This will also turn off deduplication, which depends on the UMI sequence.

In addition to the different setups for the experimental barcode, racoon_clip will include different processing steps in its default behaviour for each of the pre-settings (**Figure**

3.3). Demultiplexing is by default always turned off. For iCLIP and iCLIP2 data all other steps will be executed, while for eCLIP data the barcode trimming step is skipped, as the barcodes are located on read #1. For already preprocessed data as from ENCODE or SRA, racoon_clip directly starts with the genome alignment. The steps performed by racoon_clip can also be customised by the user.

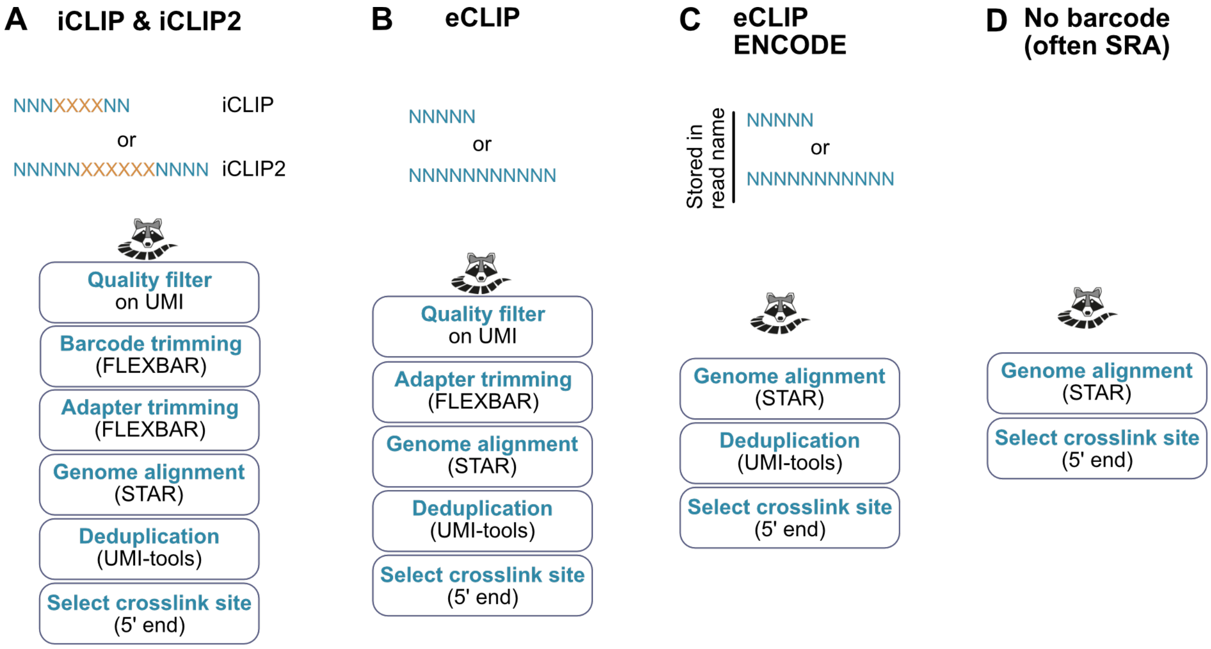


Figure 3.3: Default executed steps of racoon_clip for the different experiment pre-settings. Top - barcode and UMI architecture; bottom - racoon_clip workflow.

3.6 Case study

To demonstrate the comparability of iCLIP and eCLIP data after racoon_clip processing I reprocessed two publicly available U2AF2 eCLIP and iCLIP data sets. The U2AF2 eCLIP has been performed in two replicates in HepG2 cells [Van Nostrand *et al.*, 2020b], contains 18,760,220 reads in total and is available on the ENCODE database [Luo *et al.*, 2019]. The U2AF2 iCLIP has been performed in three replicates in HeLa cells [Sutandy *et al.*, 2018], contains 66,041,912 reads in total and is available in the SRA database.

3.6.1 Execution

For the U2AF2 iCLIP, the data stored at SRA is already demultiplexed, deduplicated and adapters and barcodes are trimmed off. Therefore, I used the following settings for the processing: experiment_type: iCLIP, demultiplex: False, quality_filter_barcodes: False, adapter_trimming: False, deduplicate: False. As the barcodes including the UMIs were not present in the data anymore, no deduplication could be performed. As data stored at SRA often already has barcodes trimmed off, I implemented the additional experiment type experiment_type 'noBarcode_noUMI' (**Figure 3.3**), that contains these settings. However, it should be noted that skipping deduplication might introduce PCR biases into the crosslink numbers. For the eCLIP data set, the following settings were used: experiment_type: eCLIP_ENCODE, demuliplex: False, quality_filter_barcodes: False, adapter_trimming: True, adapter_cycles: 2, deduplicate: True.

In terms of computational resources, racoon_clip needed 1 hour and 54 minutes for the iCLIP data set at a peak RAM usage of 40.5 GB on 6 cores. For the eCLIP data set, 2 hours and 13 minutes at a peak RAM usage of 35.4 GB were needed on 6 cores. In the following, I will describe the resulting processing report and then compare the obtained crosslinks of both experiments.

3.6.2 Reported processing steps

racoon_clip provides a detailed HTML report about the performed steps (**Figure 3.4 A**). The full reports of the U2AF2 data sets are stored at https://github.com/MelinaKlostermann/phD_thesis_additional_code/tree/main/racoon_clip_U2AF2. The report includes a list of all chosen or default parameters used. Then, graphical representations of the read quality before and after quality filtering and the adapter content before and after adapter trimming as generated by MultiQC are shown. Furthermore, the mapping statistics of all samples as provided by STAR and the number of crosslinks and crosslinked positions per sample are given. Here, I will give some examples from the processed U2AF2 data.

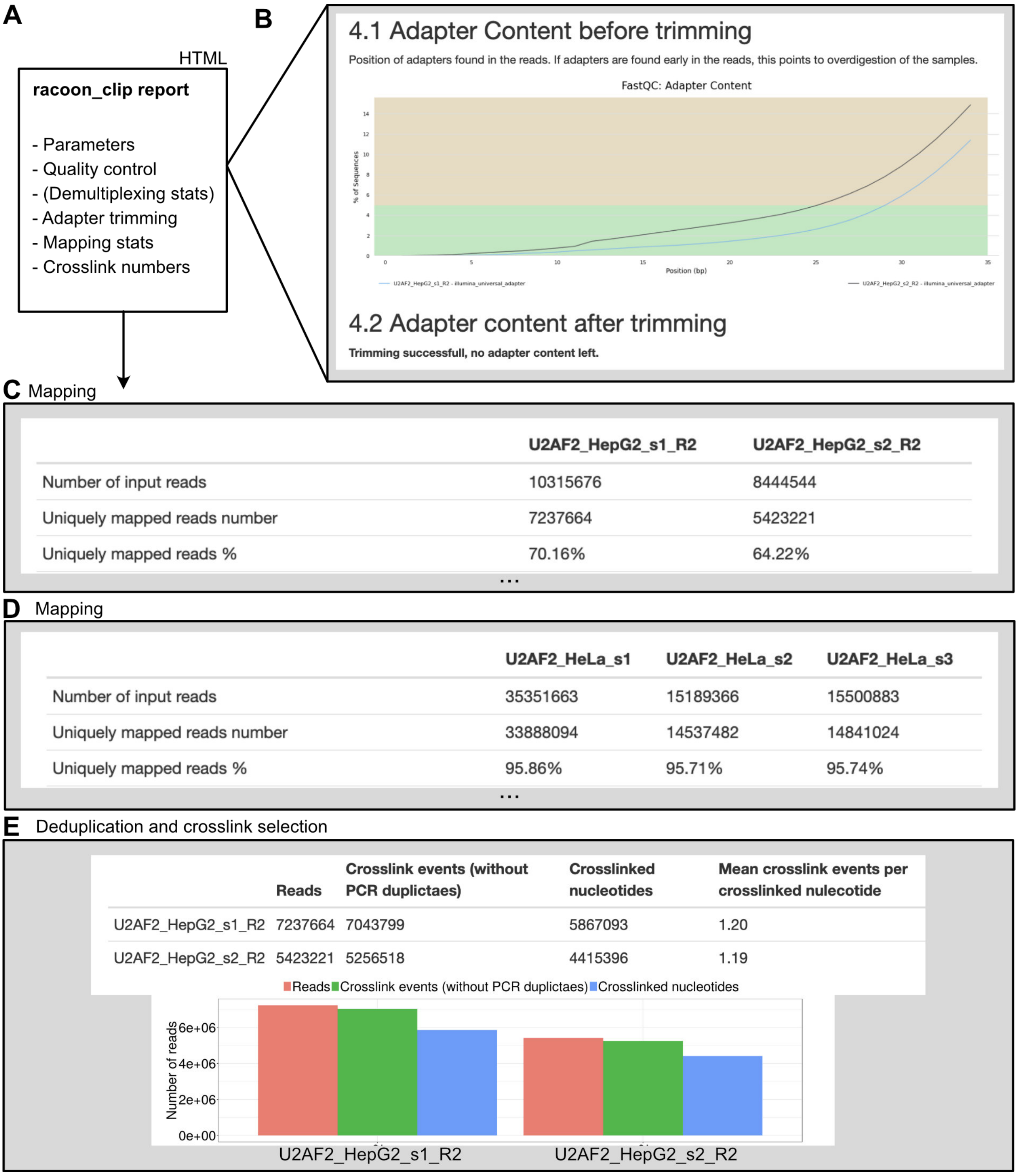
The mean read quality is above a Phred score of 28 for all positions in the eCLIP and iCLIP data, except for a drop in read quality at the last position in the iCLIP reads. In the eCLIP data, an effective adapter trimming can be confirmed by the report, which shows that adapter content was detected before but not after adapter trimming (**Figure 3.4 B**). The iCLIP data does not contain adapters and no adapter trimming was performed.

Over all samples, 67.2% and 95.8% of the sequencing reads are uniquely aligned for the eCLIP or iCLIP data, respectively (**Figure 3.4 C, D**). For the eCLIP data, 81.2% of the uniquely mapped reads pass the deduplication step (**Figure 3.4 E**), while no deduplication is performed for the iCLIP data.

As a result, racoon_clip processing of the U2AF2 eCLIP data set yields in sum of 10,282,489 crosslink events from both samples. Processing of the U2AF2 iCLIP data set resulted in a total of 63,266,600 crosslink events for the three iCLIP replicates.

3.6.3 Comparison of the resulting crosslink profiles

To test the comparability of iCLIP and eCLIP data subjected to the same processing, I first manually inspected the crosslink patterns of both data sets in a genome browser.


A
HTML
B
C Mapping
D Mapping
E Deduplication and crosslink selection

racoon_clip report

- Parameters
- Quality control
- (Demultiplexing stats)
- Adapter trimming
- Mapping stats
- Crosslink numbers

4.1 Adapter Content before trimming

Position of adapters found in the reads. If adapters are found early in the reads, this points to overdigestion of the samples.

FastQC: Adapter Content
% of Sequences
0
2
4
6
8
10
12
14
Position (bp)
0
5
10
15
20
25
30
35
U2AF2_HepG2_s1_R2 - illumina_universal_adapter
U2AF2_HepG2_s2_R2 - illumina_universal_adapter

4.2 Adapter content after trimming

Trimming successfull, no adapter content left.

	U2AF2_HepG2_s1_R2	U2AF2_HepG2_s2_R2
Number of input reads	10315676	8444544
Uniquely mapped reads number	7237664	5423221
Uniquely mapped reads %	70.16%	64.22%

...

	U2AF2_HeLa_s1	U2AF2_HeLa_s2	U2AF2_HeLa_s3
Number of input reads	35351663	15189366	15500883
Uniquely mapped reads number	33888094	14537482	14841024
Uniquely mapped reads %	95.86%	95.71%	95.74%

...

	Reads	Crosslink events (without PCR duplictaes)	Crosslinked nucleotides	Mean crosslink events per crosslinked nulecotide
U2AF2_HepG2_s1_R2	7237664	7043799	5867093	1.20
U2AF2_HepG2_s2_R2	5423221	5256518	4415396	1.19

🟥Reads🟩Crosslink events (without PCR duplictaes)🟦Crosslinked nucleotides
Number of reads
0e+00
2e+06
4e+06
6e+06
U2AF2_HepG2_s1_R2
U2AF2_HepG2_s2_R2

Figure 3.4: Overview of racoon_clips processing reports.
(A) Topics covered in the report.
(B, C, D, E) Screen shots from the reports.
(B) Adapter content in U2AF2 eCLIP data before and after adapter trimming.
(C, D) First rows of mapping statistics for U2AF2 eCLIP (C) and iCLIP (D) data.
(E) Numbers of reads before and numbers of crosslinks after deduplication of U2AF2 eCLIP data (top) and graphical representation (bottom).

Overall, the crosslink patterns of both data sets are similar as exemplified by the crosslink profile in the *AOPEP* (*Aminopeptidase O*) gene (**Figure 3.5**). Furthermore, I made metaprofiles of the iCLIP and eCLIP crosslinks in proximity to the 3' splice site. It has been shown that U2AF2, which is part of the splicing machinery, binds its target RNAs upstream of the 3' splice site [Zarnack *et al.*, 2013]. The metaprofiles of the iCLIP and eCLIP crosslinks demonstrate that this behaviour of U2AF2 can indeed be monitored in both data sets when using single-nucleotide resolution crosslinks from racoon_clip.

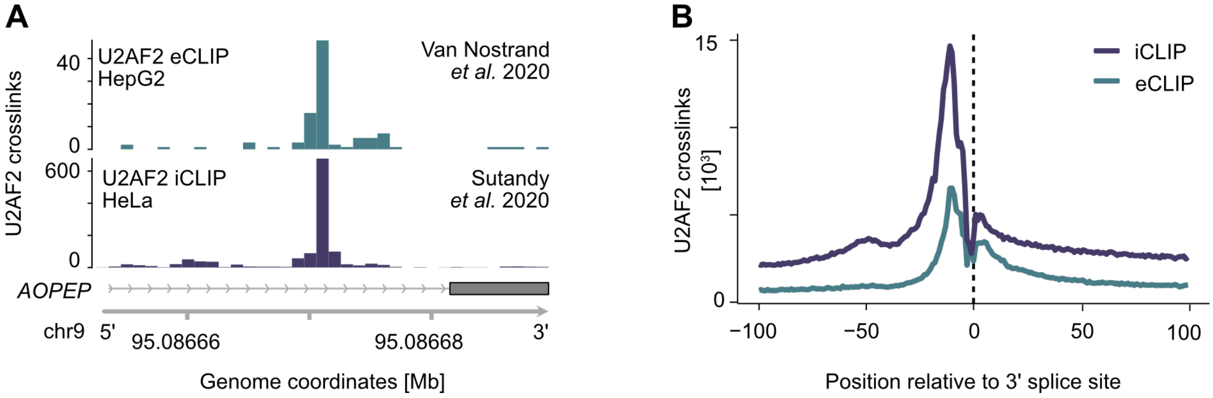


Figure 3.5: Comparison of U2AF2 crosslink profiles from eCLIP and iCLIP data.
(A) Exemplary genome browser view of the eCLIP (top) and iCLIP (bottom) U2AF2 crosslinks in the *AOPEP* gene. Crosslinks of all replicates are added. A model of the *AOPEP* transcript and the genomic coordinates are displayed below.
(B) Metaprofiles of U2AF2 crosslinks around the 3' splice site. Used are all annotated splice sites in the genome annotation. Dark blue - iCLIP, light blue - eCLIP.

In summary, racoon_clip can efficiently process both iCLIP and eCLIP data to single nucleotide resolution. The example of the U2AF2 data sets demonstrates the comparability of both experiment types when processed in the same way.

3.7 Discussion

The interest in RNA biology has increased immensely during the last two decades and with it the interest in RBPs which orchestrate the lifecycle of RNAs. CLIP experiments are an intriguing approach to query the interactions of RNAs and RBPs *in vivo*. Therefore, more and more CLIP experiments are performed both in the contexts of large-scale databases and of individual research questions. With racoon_clip I provide an automated and reproducible way to process iCLIP and eCLIP data, which I believe will be useful for a large user community. I facilitated the usage of racoon_clip by providing a command line interface, pre-set experiment types and a detailed documentation.

3.7.1 The usage of workflow managers

I aimed to optimise racoon_clip for two different use cases. On the one hand, I wanted to facilitate the processing of CLIP data for individual researchers analysing their own data. On the other hand, I wanted to provide an option to process large numbers of both iCLIP and eCLIP data sets in exactly the same way as a basis for large-scale computational CLIP analyses. I chose to implement racoon_clip based on workflow management because both use cases can gain from this approach. Workflow management will make the execution of a workflow easy to use because it allows a whole pipeline to be run in one command. Furthermore, it facilitates tool installation by providing the option to manage software dependencies in specific environments. For projects working with larger numbers of CLIP data sets, a workflow manager ensures that all data are processed in exactly the same way and that the workflows can easily be upscaled to process multiple samples and multiple data sets in parallel.

The most commonly used workflow managers for biological data analysis are Snakemake and Nextflow (**Chapter 2.4.4**). I chose to use Snakemake for the racoon_clip implementation because it is relatively flexible for development. Firstly, Snakemake allows the integration of Python code around its workflow syntax. This especially facilitated the implementation of optional steps by integrating conditional Python statements into the general steps of the workflow. Moreover, Snakemake organises the pipeline's structure based on the output files, in contrast to Nextflow which organises the pipeline's structure based on the input files. The output-based structure simplifies the development of new options or pipeline steps because Snakemake will check that all rules generating the necessary intermediate output files will be generated by some rule before starting. This has the main advantage that errors in new code will be found prior to the execution of all upstream steps. As a result, the Snakemake workflow that racoon_clip is based on can in the future easily be modified to include new options, different data types or additional processing steps.

While analyses implemented with workflow management certainly have many advantages, the usability for other users, who are unfamiliar with the syntax of the given workflow manager, might be very limited. Therefore, providing software as a native workflow script might be problematic. To resolve this problem [Roach *et al.*, 2022] proposed to hide the workflow back-end behind a simple command line interface to make it accessible to a larger group of users. This can reduce the execution of the workflow to two steps: install and run. I adopted their approach and added a command line interface to racoon_clip. Overall, I am confident that the combination of a Snakemake workflow management and a command line interface provides a sophisticated backbone to CLIP analyses that supplies high usability and reproducibility while processing iCLIP and eCLIP data in the same way.

3.7.2 Distributing software

An important question in the development of scientific tools is how to make the tool accessible to the scientific community. A guideline approach for scientific data management set up the so-called FAIR Data Principle, requesting scientific data to be findable, accessible, interoperable and reproducible [Wilkinson *et al.*, 2016]. While these rules primarily were made for scientific data, the same principles should be considered for scientific tools. Other scientists should be able to find, access and adjust software tools. A common practice that provides basics of FAIR principles for software is to store tools on GitHub together with an installation manual and documentation. The advantages of GitHub are the integration with git version control, which allows to make new releases of newer versions of the tool, while still keeping the older versions accessible. Moreover, it allows the possibility to share non-commercial code publicly and other developers can branch the original software to adjust it to their own needs.

However, a problem that can arise from providing a tool purely on GitHub is that the users have to install all tool dependencies by themself. Especially for tools with a high number of dependencies, this will easily lead to clashes of dependencies in different computing environments. This in turn will drastically increase the time needed by users to make a tool usable on their own system. A possibility to avoid complex installations is to provide tools via a package manager. Commonly used package managers for Python-based tools are Anaconda or the Python Package Index (PyPI). Both allow installing a tool by a single command line (conda install tool or pip install tool) while managing dependencies for the user. This improves the accessibility of a given tool. Uploading a tool to both PyPI or Anaconda requires an underlying GitHub repository of the given tool and a certain folder structure of the tool's code, also called skeleton. At the time of writing this thesis, I did not upload racoon_clip to PyPI or Anaconda but made it solely installable from GitHub. As racoon_clip uses Snakemake environments to manage most of its dependencies, its installation from GitHub is relatively easy. An advantage of providing racoon_clip only on GitHub at the moment is that changes to the code can be implemented faster. Still, I built racoon_clip already using the required Anaconda skeleton to make a later upload easier and plan to upload racoon_clip to both Anaconda and PyPI later after integrating the first feedback from the first users.

As most tool developers in academic groups will not stay in the group long-term to maintain their tools, the computational tools will actually ‘age‘. Over time, dependencies of tools will become outdated, as for example a tool might depend on old versions of programming languages, and a user might not want to downgrade his computational system to an older state just to be able to use a certain tool. This problem is not fully solved by using package managers like Anaconda and PyPI, because they expect software developers to maintain their tools in a way that is up-to-date with their surrounding computing systems. A solution to avoid both complicated installations and ageing of tool dependencies is the usage of containers [Reiter *et al.*, 2021, Wratten *et al.*, 2021]. The idea of a container is to completely separate a given process from the surrounding computational system. Therefore, the container will only contain the computational tool and its dependencies, with the advantage that old dependencies can be used inside the container forever without leading to clashes with newer computational systems. At the time of writing this thesis, the most popular container engine is Docker. To make it easier for users to run a given software tool in a container, one can provide a container image for the tool. This contains the full installation commands for the tool and all dependencies. Users can directly clone this image and start a container from it. A high-level example of this is the Bioconductor package manager, which provides R packages for biological use cases [Gentleman *et al.*, 2004]. Bioconductor uploads a docker image of each of its new versions (https://bioconductor.org/help/docker/, accessed 03.12.2023). Therefore, packages that are no longer maintained will still be easily usable in their docker container. Following this example, I plan to provide a racoon_clip docker image in the future.

3.7.3 Managing RAM requirements in workflows

A Snakemake workflow like racoon_clip inherits a major part of its RAM and run-time requirements from the command line tools used in the workflow. Moreover, it amplifies the needed RAM by executing multiple jobs in parallel. To optimise the RAM usage of Snakemake workflows, one can locate the steps with the highest requirements and possibly optimise the step or exchange the underlying tool used in the step by a tool with fewer requirements. For racoon_clip, the major part of the RAM usage can be traced back to the STAR aligner. The STAR developers already list its high RAM usage as a limitation of STAR, stating that 'mammalian genomes require at least 16 GB of RAM, ideally 32 GB' (https://github.com/alexdobin/STAR, accessed 02.12.2023). As a result, the STAR usage in racoon_clip has the same high RAM requirement, which is exemplified by the peak RAM usage for the U2AF2 data sets (iCLIP - 40.5 GB, eCLIP - 35.4 GB). While RAM requirements between 20 GB and 100 GB should not be problematic for a compute cluster, it rules out the usage on a local machine. This might reduce the usability of racoon_clip for researchers with little computational resources. I, therefore, want to discuss different approaches that could resolve this issue.

Firstly, to keep the RAM usage from multiplying too much on multiple data sets, Snakemake allows users to pre-set the maximum usable RAM. It then organises the parallelisation of the steps in a way not exceeding the RAM maximum. This temporary solution can be important to limit RAM during racoon_clip cluster execution to the available resources, especially when providing a high number of cores and many samples.

Secondly, a solution might be to exchange the STAR aligner in the racoon_clip workflow with another tool that has less RAM requirements. However, exchanging the aligner should not be done at the expense of the alignment quality. An important feature of an aligner for CLIP analysis splice-awareness, as RBPs can bind both to pre-mRNA and mature mRNA. Furthermore, in contrast to the alignment of RNAseq data, CLIP reads will not be distributed evenly over the whole transcript but accumulate at the crosslink sites. Therefore, splice-aware aligners that model splicing depending on the distribution of the alignments of other reads like TopHat2 [Kim *et al.*, 2013] are not suitable for CLIP data. The STAR approach of a maximum mappable prefix in combination with stitching and scoring of the aligned fragments is independent of the alignment of other reads and, therefore, more appropriate for CLIP data. In general, STAR was shown to perform best of all non-commercial tools in a benchmark of RNAseq aligners [Baruzzo *et al.*, 2017]. The commercially available RNA aligners Novoalign (Novocraft Technologies Sdn Bhd.) and CLC (QIAGEN) perform alignment similarly well [Baruzzo *et al.*, 2017], but need also similar amounts of RAM [Donato *et al.*, 2021]. Moreover, STAR has a lower run-time than tools with less RAM requirements. This can be exemplified by TopHat2, which has a 10-fold higher run-time and in addition, was also shown to perform alignment less well than the STAR [Baruzzo *et al.*, 2017, Donato *et al.*, 2021]. All in all, it would, therefore, at present not be recommendable to exchange the usage of STAR in the racoon_clip workflow with another aligner.

Thirdly, a possibility to remove the dependency on compute clusters would be to include the racoon_clip workflow into the Galaxy platform. Galaxy is a non-commercial browser-accessible workbench for bioinformatic analyses [Afgan *et al.*, 2022]. It allows users to run tools like STAR via a publicly available server. The tools are pre-installed and are provided with a clickable browser interface. Galaxy also provides a collection of publicly available workflows. These can be built with the Python library BioBlend which provides an interface for interacting with Galaxy [Sloggett *et al.*, 2013]. As not all tools used by racoon_clip are available on Galaxy it would be necessary to either create a Galaxy plug-in of these tools or to search for alternatives. The advantages of providing racoon_clip on Galaxy would be the large user base, its free compute resources and the easy usability. The main disadvantage of the publicly available Galaxy servers is that the memory per user is limited to 250 GB. This would roughly be enough for a CLIP data set with up to

10 samples, which is sufficient for most basic use cases. A Galaxy workflow will probably not be used by researchers performing CLIP analyses regularly with sufficient computing resources and expertise. These researchers will gain much more from the scalability of the command line tool racoon_clip, which makes it relatively easy to process several CLIP data sets for comparisons. Still, providing a Galaxy version of racoon_clip in addition to the command line tool in the future might make the workflow accessible to a different type of user, who just wants to analyse one specific CLIP data set with minimal computational difficulties.

In summary, a bottleneck of racoon_clip is its relatively high RAM requirement, which it obtains by using the RAM-intensive STAR for read alignment. There are no equally well-performing aligners with lower RAM usage that are suitable for CLIP data, and this step should, therefore, remain unchanged. Instead, providing racoon_clip additionally on Galaxy for users with limited compute resources might be a good solution.

3.7.4 Peak calling as the next step downstream of racoon_clip

After running racoon_clip, at the moment the user still has to perform peak calling and subsequent binding site definition separately (**Chapter 2.4.3**). In consequence, these steps could also be included in racoon_clip in the future, opening the question of which peak caller to use.

There are around ten peak callers available for iCLIP and/or eCLIP data [Hafner *et al.*, 2021] including CLIPer [Van Nostrand *et al.*, 2020a], Skipper [Boyle *et al.*, 2023], PureCLIP [Krakau *et al.*, 2017], CLIP Tool kit [Shah *et al.*, 2016], iCount [Wang *et al.*, 2010], omniCLIP [Drewe-Boss *et al.*, 2018], Piranha [Uren *et al.*, 2012] and PIPE-CLIP [Chen *et al.*, 2014]. In general, a comprehensive benchmark of the performance of these CLIP peak calling tools is missing, although some publications introducing a new tool provide some comparison to other tools [Krakau *et al.*, 2017, Boyle *et al.*, 2023]. Therefore, at the moment a decision for a certain peak calling algorithm can not be based on the overall performance.

The biggest differences between peak callers are whether they provide single-nucleotide peaks as output and whether they need a control experiment for their peak statistic. According to my survey, all peak callers except PureCLIP and iCount depend on control data to infer the gene expression. However, it has been controversial in the CLIP community whether input controls improve or rather bias the analysis [Chakrabarti *et al.*, 2018, Hafner *et al.*, 2021]. It has been argued that CLIP experiments in contrast to ChIP-seq and RNA immunoprecipitation experiments are so optimised for the removal of noise during the experiment that it is unnecessary to include background into the analysis [Chakrabarti *et al.*, 2018]. Following this argument, iCLIP protocols are performed without a control [Buchbender *et al.*, 2020]. In contrast, eCLIP protocols include so-called SMInput samples as control, which contain all crosslinked RNAs without selecting those bound by the RBP of interest by immunoprecipitation [Van Nostrand *et al.*, 2016].

In my example, racoon_clip processing of U2AF2 analysis with the iCLIP data set did not have SMInput samples while the eCLIP data did. I did not include the SMInput of the eCLIP data, but they can be included in racoon_clip processing if this is chosen by the user. For this, SMInput samples would be processed by racoon_clip as separate samples and then they can be provided as background to the peak caller. In a different project, I tested racoon_clip processing of an AGO2 (Argonaute RISC Catalytic Component 2) eCLIP data set, followed by peak calling and binding site definition with and without SMInput samples, to further analyse how the usage of an input control influences binding site definition. In this case, the usage of the SMInput showed very little influence on the resulting binding sites (data not shown), strengthening the approach without an SMInput control.

A second criterion that should be considered is the resolution of the peak caller. To obtain the high resolution of narrow binding sites, racoon_clip selects the nucleotide before the reads 5'-end as crosslink position. Therefore, in the following step, a peak caller providing single-nucleotide peaks should be chosen. Single-nucleotide resolution of peaks is provided for example by PureCLIP, omniCLIP and iCount. Many other commonly used tools like PIPE-CLIP, Piranha, CLIPer, which are used by the ENCODE database, and its follow-up tool Skipper do not provide single-nucleotide resolution of called peaks. Interestingly, the publication introducing Skipper even mentions that 'PureCLIP exhibited the most focal signal' in a comparison to Skipper and four other peak callers including omniCLIP [Boyle *et al.*, 2023]. Therefore, again, the best peak calling options considering the peak resolution would be PureCLIP or iCount.

PureCLIP peak calling uses a hidden Markov model, that is built on the complete genome as described in **Chapter 2.4.3**. I used PureCLIP in the binding site definition of PURA binding sites which is described in detail in the next chapter. However, I decided to not include PureCLIP into racoon_clip because it needs run-times of one to three days at a RAM usage of usually above 90 GB when provided with 8 or 10 cores. This is likely because PureCLIP first trains its Hidden Markov Model nucleotide-wise on all chromosomes and then tests the state of every nucleotide. A solution without increasing the run-time of the preprocessing could be to include PureCLIP as a second module to racoon_clip that can be executed after the processing to single-nucleotide resolution if chosen by the user. This second module could then also contain the subsequent binding site definition step. Overall, the user could then run two commands called for example 'racoon_clip process' and 'racoon_clip bindingSites' where the first part would maintain the relatively fast run-time that racoon_clip has now and the second part would inherit the long run-time of PureCLIP, but would only be executed if specifically needed.

Another solution might be to use iCount instead of PureCLIP for peak calling. In contrast, iCount performs peak calling by permutation analysis [Wang *et al.*, 2010]. For this, the counts per gene are randomly distributed within gene regions to generate a background. Then, the observed distribution is compared to the random distribution to calculate a false discovery rate. This approach might be faster because it considers only annotated and expressed genes. At the same time, this is also the drawback of this method as it depends on a complete annotation for the studied organism.

In summary, racoon_clip could be expanded by a peak calling step and the subsequent binding site definition step. A suitable peak caller should function without provided background samples to make it usable with all iCLIP data and should provide the called peaks as single nucleotides to maintain the narrow signal of the single-nucleotide crosslinks. From the CLIP peak callers available at the moment, PureCLIP and iCount meet these expectations and could be tested against each other to decide on the best option.

3.7.5 Conclusion

Taken together, with racoon_clip I can provide a processing workflow for iCLIP and eCLIP data, enhancing the comparability of the two data types. As the read data from both experiments has the read start position next to the UV crosslink, both data types can be processed similarly when accounting for the different read architectures as done with racoon_clip. Indeed, in the example of U2AF2 iCLIP and eCLIP data, crosslink patterns from both experiment types are very similar. This highlights the interchangeability of the data from both experiment types when using the same downstream analysis. Therefore, racoon_clip processing will allow the combination of iCLIP and eCLIP data for analyses like training machine learning approaches or modelling RBP networks [Horlacher *et al.*, 2023].

Furthermore, with racoon_clip single-nucleotide crosslinks can be obtained for both iCLIP and eCLIP data. This increases the resolution of the binding patterns in comparison to the usage of full reads. To keep the high resolution of the crosslink profiles in the following peak calling and binding site definition, a peak caller that can provide single-nucleotide peaks should be used.

I automatised racoon_clip by implementing it as a Snakemake workflow. This provides racoon_clip with all the advantages of workflow management including organisation and parallelisation of steps, separate software environments for the steps and portability to other computational systems. To allow the user to keep track of the automatically performed steps, I included a detailed HTML report that contains statistics and visualisations of all steps.

To simplify the usage of racoon_clip, the Snakemake workflow is hidden behind a command line interface. All parameters can be passed to racoon_clip via command line options or in a configuration file. Furthermore, as all needed software is provided in racoon_clip‘s mamba environments, racoon_clip‘s only dependencies are Python and mamba, facilitating its installation. I, therefore, believe that racoon_clip will be very useful both for a broad community of people performing CLIP experiments in need of robust processing of their data and for comparisons of multiple new and publicly available CLIP data sets.

Chapter 4

The cellular role of the RNA-binding protein PURA and its connection to cytoplasmic granules

A major part of the analyses shown in this chapter have been published in

Depletion of the RNA-binding protein PURA triggers changes in posttranscriptional gene regulation and loss of P-bodies.
Molitor L*, Klostermann M*, Bacher S, Merl-Pham J, Spranger N, Burczyk S, Ketteler C, Rusha E, Tews D, Pertek A, Proske M, Busch A, Reschke S, Feederle R, Hauck SM, Blum H, Drukker M, Fischer-Posovszky P, König J, Zarnack K[$], Niessing D[$] (2023); Nucleic Acids Research, 51(3):12971316. DOI http://dx.doi.org/10.1093/nar/gkac1237. (* *both authors contributed equally;* [$] *shared correspondence*)

The figures listed in **Table 4.1** were done by me for this publication and are taken from there.

This thesis	[**Molitor *et al.*, 2023**]
Figure 4.1 B	Figure 2C (modified)
Figure 4.1 C	Figure 3A
Figure 4.2	Figure S2 B
Figure 4.3 A	Figure 3B (modified)
Figure 4.3 B	Figure S2A

Figure 4.3 B	Figure S6H (modified)
Figure 4.5 B	Figure 3D
Figure 4.8 B	Figure S4 B (modified)
Figure 4.8 B	Figure S4 C (modified)
Figure 4.11 A	Figure S6 J
Figure 4.11 B	Figure 3E (modified)
Figure 4.12	Figure 3F
Figure 4.13 B	Figure S6 I
Figure 4.14 B	Figure 4A
Figure 4.14 C	Figure 4B
Figure 4.15 A	Figure 4C
Figure 4.15 B	Figure S8 A
Figure 4.15 C	Figure 4D
Figure 4.16 A	Figure S9 B
Figure 4.16 B	Figure S9 E
Figure 4.17 A	Figure S9 A
Figure 4.17 B	Figure S9 D
Figure 4.17 C	Figure S9 C
Figure 4.17 D	Figure S9 F
Figure 4.18	Figure 4E
Figure 4.19 A, B	Figure S6 E, F
Figure 4.21 B	Figure 2C (modified)
Figure 4.21 C, D	Figure S5 F (modified)
Figure 4.21 E	Figure 2F
Figure 4.21 F	Figure 2E
Figure 4.22 A, B, C	Figure S5 A, B, C
Figure 4.23 A, B	Figure A, B
Figure 4.24 A , B	Figure S10 E, F (modified)
Figure 4.24 C, D	Figure S10 B,C (modified)
Figure 4.24 E	Figure S10 D
Figure 4.24 F	Figure S10 A
Figure 4.24 G	Figure 5A
Figure 4.25 A, B	Figure 6 C, D (done by Sabrina Bacher)
Figure 4.25 C	Figure 5E (modified)

Table 4.1: Figures reused from our previous publication.

4.1 Research objectives

At the start of this project, PURA was known primarily as a DNA-binding protein, with limited evidence for its interaction with RNAs in the cellular context. The primary aims of this research were threefold:

1. **Global characterisation of PURA's RNA binding behaviour:** The first objective was to comprehensively assess PURA's ability to bind RNA molecules in living cells. This involved not only determining the extent of PURA's RNA-binding capacity but also delineating key characteristics of its binding behaviour, including its preferred transcript regions and binding motifs.

2. **Revealing PURA's cellular function:** Although PURA has been implicated in several diseases, its precise role in cellular processes has remained elusive. To address this, my research aimed to elucidate the cellular functions of PURA RNA binding. This was done by analysing the effects of *PURA* depletion on RNA and protein expression in cells to uncover the functional implications.

3. **Characterising PURA's associations with neuronal cells and cytoplasmic granules:** My investigation connected PURA to various cellular functions. From there, I was especially interested in PURA's RNA-binding in neuronal cells, as well as its association with stress granules and P-bodies. Therefore, I aimed to gain a more complete understanding of the role of PURA in these specific cellular contexts using publicly available data.

4.2 Data and Methodology

Please note that the methods used here and, therefore, the description of the method greatly overlap with the methods from [Molitor *et al.*, 2023]. I paraphrased the methods from this manuscript as much as possible. For the methods used for the experiments please refer to [Molitor *et al.*, 2023] as they were performed by our collaborators.

4.2.1 Binding site definition, characterisation and comparison

In this project, I use four different PURA iCLIP data sets (**Table 4.2**). Three of them were performed in HeLa cells and a fourth in neuronal precursor cells (NPC PURA iCLIP). From the PURA iCLIPs in HeLa cells, one was performed with endogenous PURA expression (endogenous PURA iCLIP) and two with an ectopic overexpression of a FLAG-PURA construct. In the first FLAG-PURA iCLIP (FLAG-PURA IH-AB iCLIP), the immunoprecipitation was performed with the same in-house anti-PURA antibody (anti-PURA^{12D11}) that was also used in the endogenous PURA iCLIP and the NPC PURA iCLIP. In the second FLAG-PURA iCLIP, the immunopecipitation was done with an anti-FLAG antibody (FLAG-PURA FL-AB iCLIP). The experimental details and the preprocessing steps from raw reads to single-nucleotide crosslinks can be found in the methods section of [Molitor *et al.*, 2023] (see chapters Generation of a PURA-specific antibody, Generation of inducible PURA overexpression HeLa cell line, Neural progenitor cell derivation from human induced pluripotent stem cells and Processing of iCLIP reads).

Binding site definition

I defined 5-nt wide binding sites on the iCLIP data of all three PURA iCLIP experiments from HeLa cells with the method described by [Busch *et al.*, 2020]. The NPC PURA iCLIP had a very limited signal, which was not sufficient for binding site definition. For the theory behind peak calling and binding site definition see **Chapter 2.4.3**.

	Target	**Cells**	**Antibody**
Endogenous PURA	endogenous PURA	HeLa cells	anti-PURA
FLAG-PURA IH-AB	ectopic FLAG-PURA (overexpression)	HeLa cells	anti-PURA
FLAG-PURA FL-AB	ectopic FLAG-PURA (overexpression)	HeLa cells	anti-FLAG
NPC PURA	endogenous PURA	neuronal precursor cells	anti-PURA

Table 4.2: Used iCLIP data sets.

The processed crosslink events of the four biological replicates from the endogenous PURA iCLIP and the FLAG-PURA IH-AB iCLIP were merged into two pseudo-replicates and subjected to peak calling by PureCLIP (version 1.3.1) [Krakau *et al.*, 2017] with default parameters. The two biological replicates of the FLAG-PURA iCLIP were directly used for peak calling with PureCLIP.

Then, to define binding sites, PureCLIP sites in a distance of less than 5 nt were merged into regions, and isolated PureCLIP sites without an adjacent site within 4 nt were discarded. Binding site centres were defined iteratively as the position with the highest number of crosslink events and enlarged by 2 nt on both sides to obtain 5-nt binding sites. Binding site centres were required to harbour the maximum PureCLIP score within the binding site. Furthermore, binding sites were filtered for reproducibility by requiring a binding site to be supported by a sufficient number of crosslink events in at least two out of four replicates for the endogenous PURA iCLIP. For the FLAG-PURA IH-AB iCLIP, I deemed binding sites reproducible if they were supported by three of the four replicates and for the FLAG-PURA FL-AB iCLIP, both replicates had to support the binding site. The threshold for sufficient crosslink event coverage, using the 0.05 percentile, was determined as described in [Busch *et al.*, 2020]. In addition, as the FLAG-PURA IH-AB iCLIP data set had a very high signal depth, global binding sites enclosing less than 6 crosslinks events were filtered out and for each gene only binding sites with a PureCLIP score in the upper 25% quantile of the gene were kept.

The shown genome browser screenshots (**Figures 4.1 B; 4.4; 4.7 C; 4.21 B, C, D; 4.23**) are made with the GVIZ package (version 1.42.1) [Hahne and Ivanek, 2016].

Transcript location of binding sites and crosslink events

To know the transcript locations of the binding sites from all three PURA iCLIP data sets in HeLa cells (**Figure 4.1 & 4.3 & 4.10**), binding sites were mapped onto the gene and transcript annotation from GENCODE (release 31, genome version GRCh38.p12) [Frankish *et al.*, 2018]. As in most cases, multiple transcripts overlapped, and it is not apparent from the CLIP data which transcript was bound, I defined a set of rules to resolve these overlaps by considering the following.

In the annotation from GENCODE both genes and transcripts are annotated with a gene level and transcript support level. This level describes the certainty that this gene or transcript was annotated correctly. For genes, a level 3 gene is a gene that was only automatically annotated. A level 2 gene is a gene whose automatic annotation was manually curated and a level 1 gene is a gene that was experimentally verified. Transcript support levels range from 1 to 5, where one is a transcript where all splice junctions were supported and five is a model structure that has never had experimental support. Transcript support levels can also be annotated as 'NA' meaning that the transcript support was not analysed yet [Frankish *et al.*, 2018]. To resolve overlapping annotation of multiple genes or transcripts, I include genes of gene level 3 only when no overlapping genes of higher level were present. Similarly, transcripts of transcript support level 'NA' were included only when no transcripts of levels 1-3 were annotated for the gene. When a binding site overlaps with two or more genes, one is chosen at random.

When assigning binding sites to transcript regions, I decided to prioritise transcripts in which the binding site was located to the 3'UTR over transcripts in which the same binding site would be located in another region. To resolve multiple overlapping transcripts with different transcript regions in the same position, I, therefore, selected the bound transcript region by a hierarchical rule with 3'UTR > 5'UTR > CDS > intron, which I established from visual evaluation of the crosslink event and binding site distribution. To normalise the number of binding sites per region to the lengths of the respective regions, I calculated a mean region length for all genes with at least one PURA crosslink in any region over all transcripts annotated for a gene. Then, I divided the number of binding sites per region by the sum of the mean region lengths of all selected genes (**Figure 4.3 C**).

I split the PURA-bound mRNAs into groups according to the location of the strongest PURA binding site on the transcript (3'UTR, n = 2,462; 5'UTR n = 98; CDS, n = 1,646; intron, n = 1; noncoding RNA = 185) (**Figure 4.5**). The metaprofiles (**Figure 4.5 B**) of the distribution of crosslink events along the transcript regions for mRNAs with strongest binding sites in the 3'UTR or CDS were calculated and visualised in R using the Bioconductor package cliProfiler (http://bioconductor.org/packages/release/bioc/html/cliProfiler.html).

Comparison of crosslinks

To compare crosslink patterns, I used heatmaps (**Figure 4.8 B**) of the first 300 nucleotides of all 3'UTRs or CDS regions with crosslinks in this region. Furthermore, I built metaprofiles (**Figure 4.8 C**) that were centred on the endogenous PURA binding sites. For both, I used window-wise min-max normalisation to remove biases from strongly expressed transcripts. Min-max normalisation is calculated by

$$x_{norm} = \frac{x_i - min(x)}{max(x) - min(x)} \tag{4.1}$$

where x is the vector of all crosslinks per window, x_i is the number of crosslink at a given position and $x_{i,norm}$ is the normalised value.

In addition, the normalised values were subjected to cubic spline-smoothing and dimension inflation for the heatmap visualisation. This approach was inspired by a similar approach used by a previous study [Heyl and Backofen, 2021]. Cubic spline smoothing is a method to regress sequential data. In short, it performs interval-wise regression using cubic functions where the intercept of an interval is always the endpoint (knot) of the previous interval. The regression of each interval is called a spline and the total of all splines can be described by

$$f(x) = \begin{cases} a_1x^3 + b_1x^2 + c_1x + d_1 & \text{if } x \in [x_1, x_2] \\ a_2x^3 + b_2x^2 + c_2x + d_2 & \text{if } x \in (x_2, x_3] \\ \dots \\ a_nx^3 + b_nx^2 + c_nx + d_n & \text{if } x \in (x_n, x_{n+1}] \end{cases} \tag{4.2}$$

where $x_1..x_n$ are the knots.

To determine the best fit of the splines, the difference between predicted values and data points should be minimal. This can be calculated with the residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{4.3}$$

where y_i are all data points and $f(x_i)$ is the predicted value.

A higher number of knots (interval ends) results in a better regression fit but also increases the roughness of the curve. Spline smoothing, therefore, uses a high number of knots but penalises the roughness of the curve by the second derivative of the splines.

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt \tag{4.4}$$

Larger λ values result in smoother curves, while smaller λ values lead to more rough curves.

Inflation of dimensions means to predict more data points from the regression than were used for the prediction. For example, for my heatmaps, I looked at a 300 nt window but show 500 predicted points for visualisation. This method can improve visualisation and refine local features.

To visualise the crosslink patterns across the three PURA iCLIP experiments in HeLa cells, I generated heatmaps (**Figure 4.8 B**) of the first 300 nt of 1,550 3'UTRs or 573 CDS with intermediate PURA crosslink coverage (10^2-10^6 crosslink events per window). The crosslink events in each window were scaled to the minimum (set to 0) and maximum (set to 1) therein by min-max normalisation followed by spline-smoothing using the smooth.spline function (R stats package version 4.1.1, spar = 0.5) and inflated dimensions (dim = 500) as described above. To compare the differences in the crosslink distributions between the three data sets, I made metaprofiles (**Figure 4.8 C**) of the min-max normalised signal in a 65-nt window around the PURA binding sites from the endogenous PURA iCLIP data set.

Comparison of binding sites and differential binding analysis

To compare binding sites between the three PURA iCLIPs in HeLa cells, I performed a differential binding analysis. To accommodate all three data sets in the analysis, I first defined a merged set of binding sites from all three data sets (**Figure 4.8 A**). Then, I used pairwise comparisons of FLAG-PURA IH-AB against endogenous PURA and FLAG-

PURA FL-AB against endogenous PURA for differential analysis (**Figure 4.9 A, B**). As a negative control, I compared samples 1 and 2 against samples 3 and 4 of the FLAG-PURA IH-AB. Finally, the empirical cumulative distribution function (ecdf) was used to compare the different pairwise comparisons (**Figure 4.9 D, E**).

The theoretical background of differential analysis and the approach to differential binding analysis with BindingSiteFinder are introduced in **Chapter 2.4.2**. Here, I additionally want to introduce the ecdf, which I refer to as cumulative distribution in the results. The ecdf displays the fraction of the sum of all data points that are smaller than a given x. For example, if x is 1, it sums up the number of all data points smaller than 1 and divides this by the total number of data points. This behaviour is given by

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x) \tag{4.5}$$

where n is the number of data points and $(X_1, .. \ X_i)$ are the individual data points. A possible result might be that a fraction of 0.5 of all data points (50% of all data points) are smaller than 1.

The following steps were performed with the BindingSiteFinder package (Bioconductor developmental version 3.18, BindingSiteFinder version 1.7.11, https://doi.org/doi:10.18129/B9.bioc.BindingSiteFinder). I merged the binding sites of the three iCLIP data sets to obtain a set of shared binding sites with the combineBSF() function, resulting in 144,370 binding sites. Then, the gene background was calculated with the calculateBsBackground() function. The background was filtered with filterBsBackground() removing genes with less than 1,000 crosslinks (minCounts.cutoff parameter) and removing extreme cases of background to binding sites ratios (balanceBackground.cutoff.bs = 0.2, balanceBackground.cutoff.bg = 0.8). Changes in PURA binding were then calculated with the calculateBsFoldChange() function. Binding sites changing with a P value <

0.01 were deemed significantly changing. *P* values were adjusted for multiple testing by BindingSiteFinder using the Independent Hypothesis Weighting method [Ignatiadis *et al.*, 2016].

Differential binding analysis was performed in a Docker container (Docker Desktop version 4.24.2) with the September 2023 development version of Bioconductor (bioconductor/bioconductor_docker:devel) [Lawrence *et al.*, 2013], as differential binding was not included in the main version of BindingSiteFinder at the time of analysis.

Motif analysis

I used several methods to analyse the sequence preferences of PURA RNA binding: Firstly, I counted the numbers of purines in PURA binding sites and generate a general sequence logo of the binding sites and the immediate surroundings. Secondly, I used the STREME algorithm to *de novo* predict sequence motifs and thirdly, I performed k-mer enrichment analyses. As PURA predominantly acts on mature RNAs, all motif analyses were performed on the sequences of spliced transcripts. Binding sites were transferred from genomic to transcriptomic coordinates with the R/Bioconductor package GenomicFeatures (version 1.45.2) [Lawrence *et al.*, 2013] using GENCODE transcript annotation (release 31, genome version GRCh38.p12) [Frankish *et al.*, 2018]. In this approach, 49,602 binding sites could be unambiguously assigned and were used for further analyses. Sequences in and around binding sites were extracted with the getSeq() function (R/Bionconductor package Biostrings, version 2.61.2) (https://bioconductor.org/packages/release/bioc/html/Biostrings.html).

In **Figure 4.11 A**, I counted the number of purines in PURA binding sites and compared it to randomly selected 5 nt long sequences in the 3'UTR or CDS of expressed transcripts (3'UTR - n = 23,436, CDS - n = 24,627). Furthermore, I calculated the distribution of nucleotides in a 15-nt window centred at the binding site as a position weight matrix

(PWM). In a PWM, every row corresponds to one nucleotide and every column to a certain position. The values in the matrix represent the likelihood or strength of each nucleotide occurring at a specific position within the motif. I visualised the PWM on the 15 nt around PURA binding sites as sequence logo (**Figure 4.11 B**) using the R package ggseqlogo (version 0.1).

For *de novo* motif discovery, STREME builds a Markov model of order k-1 on the background sequences, where k is the chosen motif length. It builds suffix trees for both the background and the control sequences to count the exact occurrence of each short sequence of k length (word). The counts are used to calculate the enrichment of the words in the primary sequence in contrast to the background and afterwards, the number of approximate matches for the top occurring words is counted. The words are turned into PWMs and the PWMs are iteratively refined. Finally, the significance of the PWMs are calculated and the PWM with the highest significance is kept. This PWM is removed from the initial set of sequences and the process is repeated until no significant motifs are found in a certain number of subsequent repetitions.

Here, I used the STREME algorithm as implemented in the R/Bioconductor package memes (version 1.6.1) [Bailey *et al.*, 2009, Nystrom, 2023] for *de novo* motif prediction searching for a motif size from 4 to 8 nt in 45 nt windows around endogenous PURA binding sites in 100 iterations (parameters: minw = 4, maxw = 8, niter = 100, patience = 10, **Figure 4.11 C, D**). This analysis was performed once for all binding sites, using a background of randomly selected 45 nt windows from expressed transcripts (background - n = 48,063) and once specifically for 3'UTR binding sites and CDS binding sites using backgrounds from the same regions (3'UTRs: n = 24,553, background - n = 23,436, CDS: n = 25,049, background - n = 24,627), respectively (**Figure 4.11 C**). Discovered motifs with a P value < 0.05 were deemed significant.

I performed k-mer analysis (**Figure 4.12**). In a k-mer analysis, a sliding window approach is used to split up the sequence in all possible sequences of length k. Then, the occurrences of each possible k-mer are counted and compared to a background sequence. In this dissertation, I searched for enriched motifs of 5-nt length (5-mers) within and around the PURA binding sites. For this, I calculated the frequencies of all overlapping 5-mers with the oligonucleotideFrequency() (R/Bioconductor package Biostrings, version 2.61.2) (https://bioconductor.org/packages/release/bioc/html/Biostrings.html) in three windows: (i) [-7 nt; 7 nt], i.e., including the binding sites plus 5 nt flanking on either side and (ii) [-27 nt; -8 nt] and [8 nt; 27 nt], i.e., representing the 20 nt on either side of the binding site-containing windows. For comparison, I randomly selected 49,602 5-nt windows from expressed transcripts (i.e., harbouring at least one PURA crosslink event).

Accessibility prediction

I used RNAplfold (ViennaRNA package version 2.4.17) [Bernhart *et al.*, 2006] to predict the accessibility of transcripts in and around PURA binding sites. RNAplfold predicts the unpaired probability or accessibility of each nucleotide in an RNA sequence based on the computation of RNA secondary structures from thermodynamic parameters and condenses multiple possible secondary structures into an unpaired probability for each position. To account for longer transcript sequences, RNAplfold uses a window of size w sliding along the RNA sequence and calculates in each window the base pair probabilities with a maximum span width of l for each nucleotide position. The probability is then averaged over all windows containing the nucleotide position.

To compute the accessibility of PURA binding sites, I used 501-nt windows either around PURA binding sites (n = 48,525) or randomly selected positions (n = 10,017) and predicted the probability for each single nucleotide to be single-stranded. Please note that binding sites that were closer than 250 nt to the transcript start or end positions were excluded from this analysis. I calculated the unpaired probability with RNAplfold with a sliding window w = 100 nt and a maximum span l = 30 nt. As the obtained unpaired probabilities follow a bimodal distribution, I transferred them to log-odds to obtain a bell-shaped distribution with

$$log(oddsratio) = log(\frac{prob}{1 - prob}) \tag{4.6}$$

where $prob$ is the unpaired probability. The log-odds were further converted into nucleotide-wise z-scores using the mean log-odds ratios of the binding sites versus the mean and standard deviation of the background regions (drawing 1,000 regions 1,000 times).

$$z - score = \frac{mean(prob_{PURA\_bound}) - mean(prob_{random\_1000x1000})}{sd(prob_{random\_1000x1000})} \tag{4.7}$$

P values were calculated from the z-scores using

$$P = 2 * pnorm(-abs(z - score)) \tag{4.8}$$

and then adjusted for multiple testing with Benjamini–Hochberg correction.

The results are displayed only for the 201 nt in the centre of the 501-nt window as RNA folding is strongly biased toward single-strandedness at the edges of a given RNA sequence. In addition, unpaired probabilities were calculated separately for binding sites in 3'UTR (n = 22,572) and CDS (n = 23,678) binding sites using the same mean and standard deviation of the background regions as for all binding sites. Again, only binding sites with the 501-nt window positioned on the respective transcript were used.

4.2.2 RNAseq and proteomics differential analyses

RNAseq and shotgun proteomics were performed in HeLa cells using *PURA* siRNA knock-down and control cells in four replicates each. For a detailed experimental description, see the methods chapters RNA sequencing, Shotgun proteomics of [Molitor *et al.*, 2023]. The theory behind differential expression analyses is described in **Chapter 2.4.2**.

The RNAseq analysis proceeded as follows: First, the reads were aligned to the human genome (GRCh38.p12 from GENCODE) using STAR (version 2.7.6a) [Dobin *et al.*, 2013], allowing up to 4% mismatches and prohibiting multimapping. Subsequently, the count of reads per gene was determined using htseq-count (version 0.11.3) [Anders *et al.*, 2014] with default settings and GENCODE gene annotation (release 31) [Frankish *et al.*, 2018]. The comparison between *PURA* knockdown and control samples for differential expression analysis was carried out via DESeq2 (version 1.33.4) [Anders and Huber, 2010], considering a significance cut-off at an adjusted P value of less than 0.01 (Benjamini-Hochberg correction). It is important to note that in this context, the term 'transcripts' is used interchangeably with 'genes'. To generate the heatmap, the library size-corrected RNAseq read counts per sample underwent a log transformation (DESeq2, version 1.33.4) [Anders and Huber, 2010]. For visualisation purposes, these counts were converted into row-wise z-scores.

The analysis of differential protein expression relied on MaxQuant label-free quantification values (LFQ) using DEqMS (R/Bioconductor package, version 1.12.1) [Zhu *et al.*, 2020]. Handling missing values followed specific rules: If protein abundance was not measured in any samples within one condition, it is set to zero, resulting in infinite values for calculated ratios. Peptides with two or more missing values in one condition are considered non-quantifiable, and no ratios are computed. In cases where only one value was missing per condition, ratios were calculated based on the other three samples.

To account for multiple testing, P values underwent Benjamini-Hochberg correction. DEqMS has a distinct advantage as it scales (adjusted) P values according to the number of detected unique peptides. Proteins were considered significant if their scaled-adjusted P value was less than 0.05. In the heatmap the displayed protein abundance is, like the

RNAseq gene counts, depicted after z-score normalisation.

4.2.3 Comparison to published transcriptomes

I used already processed RNAseq data from studies comparing RNAs in stress granules (human osteosarcoma U-2 OS cells) or P-bodies (human embryonic kidney HEK293 cells) against total RNA [Khong *et al.*, 2017, Hubstenberger *et al.*, 2017]. RNAs with a false discovery rate (FDR) < 0.01 and a $\log_2$-transformed fold change > 0, as determined by the cited authors, are designated to the respective P-body or stress granule transcriptome.

For identifying the overlap between PURA-bound RNAs and dendritically localised RNAs, I used previously published data from [Middleton *et al.*, 2019]. This study combines insights from eight studies characterising the dendritic transcriptome in mouse neurons through RNA sequencing enrichment in dendrites compared to complete neurons. Only RNAs matching their human orthologs as ‘one-to-one orthologs‘ in the Ensembl database (version GRCh38.p13) were included (n = 5,070). Among these, I considered only RNAs expressed in HeLa cells, i.e. that had a baseMean > 0 in our RNAseq data (n = 4,522). From the dendritically localised RNAs, I selected RNAs enriched in at least three of the eight studies from [Middleton *et al.*, 2019] (n = 337).

To assess significance in overlaps, Fisher’s exact test was used, while differences in fold change distributions between groups underwent evaluation using a two-sided Wilcoxon Rank-sum test.

Functional enrichment analyses

Gene set enrichment analyses were performed on the genes of PURA-bound RNAs and the genes of RNAs whose expression levels changed upon *PURA* depletion. For this, the functional annotations from REACTOME (version 7.5.1) and Gene Ontology (GO) (version 7.5.1) databases [Fabregat *et al.*, 2017, Carbon *et al.*, 2018] are used. The over-representation or enrichment of target genes in these sets is computed assuming a hypergeometric distribution for non-enriched gene sets by using the 'hypergeometric' mode of the R/Bioconductor package hypeR (version 1.9.1). HypeR adjusts the gene sets connected to every term specifically to genes present in a given background set. Here, I used all genes with at least one PURA crosslinked nucleotide as the background set for the enrichment of genes bound by PURA and all genes with a TPM > 0 in the RNAseq experiment as the background set for enrichment of RNA expression changes in *PURA* depletion.

For visualisation, I calculated the gene ratios by dividing the number of target genes belonging to a given term by the number of genes from the background set of the term. *P* values obtained from HypeR are adjusted using the Benjamini Hochberg method. **Figure 4.19 & 4.20** display the 25 or 20 terms with the lowest *P* value for the genes of PURA-bound RNAs and the genes of RNAs changing significantly upon *PURA* depletion. A full list with the statistics of all terms can be found for PURA-bound RNAs in Supplementary Table S3 of [Molitor *et al.*, 2023]. A similar list for RNAs changing upon *PURA* depletion is provided in the GitHub repository of this thesis at https://github.com/MelinaKlostermann/phD_thesis_additional_code/tree/main/PURA/10_GeneOntology.

Other

Boxplots: In all shown boxplots, the box depicts the range from the 25% quantile to the 75% quantile, the line annotates the median and the whiskers range to the 5% and 95% quantiles of the respective data. Data points outside of the whiskers are depicted as dots.

Venn diagramms: Venn diagrams are computed with the R package eulerr (version 7.0.0). The areas of the circles and overlapping areas are fitted to the sizes of respective sets.

Machines used for computation

All command line tools were executed on the Ebersberger group's compute cluster at the Goethe University Frankfurt, named ebersberger-46-150.biologie.uni-frankfurt.de (Kernelx86_64 Linux 5.4.0-165-generic, CPU: Intel Xeon E5-2609 v4 @ 8x 1.7GHz [40.0°C]). The operating system was 'Ubuntu 20.04 focal' and the shell used was bash (version 5.0.17). Access was via SSH and job management was via the SLURM workload manager. Command line tools were installed using anaconda (conda version 4.8.4, except ViennaRNA [Lorenz *et al.*, 2011], which was downloaded as a pre-compiled executable from the ViennaRNA website. All other computations were performed in R (version 4.2.2) with RStudio (2022.12.0.353, memory limit 8 GB) on a MacBook Pro with a 2.4GHz quad-core Intel Core i5 processor.

4.2.4 Data and code availability

All sequencing data was stored in the Gene Expression Omnibus (GEO) database under the accession GSE193905. The proteomics data is stored at the ProteomeXchange Consortium with the PRIDE partner repository [Perez-Riverol *et al.*, 2019] under the identifier PXD030266.

In addition, the following previously published data sets were used: dendritic RNAs [Middleton *et al.*, 2019], P-body transcriptome [Hubstenberger *et al.*, 2017], stress granule transcriptome [Khong *et al.*, 2017].

The code used in [Molitor *et al.*, 2023] was submitted to Zenodo (https://doi.org/10.5281/zenodo.7273088), and can also be found on GitHub (https://github.com/ZarnackGroup/Publications/tree/main/Molitor_et_al_2022). The code used for the additional analyses in this thesis is stored in a separate GitHub repository (https://github.com/MelinaKlostermann/phD_thesis_additional_code).

The supplementary tables listed in **Table 4.3** can be found either in the supplementary tables of [Molitor *et al.*, 2023] or in the GitHub repository of this thesis (https://github.com/MelinaKlostermann/phD_thesis_additional_code/tree/main/PURA/11_supplementary_tables).

Content of table	**Storage of table**
Endogenous PURA binding sites	[Molitor *et al.*, 2023] Supplementary Table 2
FLAG-PURA IH-AB binding sites	GitHub repository Table 1
FLAG-PURA FL-AB binding sites	GitHub repository Table 2
Differential binding	GitHub repository Table 3-5
STREME logo prediction for endogenous binding sites	GitHub repository Table 6-8
Enriched REACTOME and GO cellular component terms for PURA-bound transcripts	[Molitor *et al.*, 2023] Supplementary Table 3
Differential RNA expression upon *PURA* knockdown	[Molitor *et al.*, 2023] Supplementary Table 4
Differential protein abundance upon *PURA* knockdown	[Molitor *et al.*, 2023] Supplementary Table 5
Enriched REACTOME and GO cellular component terms for differentially expressed RNAs in *PURA* knockdown	GitHub repository Table 9-10

Table 4.3: List of supplementary tables.

4.3 Results

4.3.1 PURA as global RNA-binding protein

At the beginning of this project, several findings pointed towards the role of PURA in binding RNAs. To investigate this aspect, I examined three PURA iCLIP experiments in HeLa cells (**Figure 4.7 A**) to determine binding patterns and characteristics of PURA RNA-binding in living cells.

In the subsequent sections, I describe in detail the characteristics of the PURA binding sites defined on an endogenous PURA iCLIP data set in HeLa cells, the first of the three PURA iCLIPs. Then, I globally compare the crosslink patterns and binding sites of the three iCLIP experiments. Afterwards, I analyse the endogenous PURA binding sites to determine whether PURA has preferences for specific nucleotide compositions or secondary structures.

Binding sites of endogenously expressed PURA

The endogenous PURA iCLIP data set shows a considerable RNA-binding of PURA. This can already be seen from the PURA crosslink patterns here exemplarily shown in the 3'UTR of the *STARD7* transcript (**Figure 4.1 A, B**). To capture the binding behaviour, I defined five nucleotide-wide PURA binding sites on the PURA crosslinking patterns. This resulted in 55,031 PURA binding sites that were found on the transcripts of 4,391 genes. The majority of the PURA-bound transcripts (4,206, 95.5%) come from protein-coding genes, but PURA binding sites are also found on several classes of long and small non-coding RNAs (**Figure 4.1 C**).

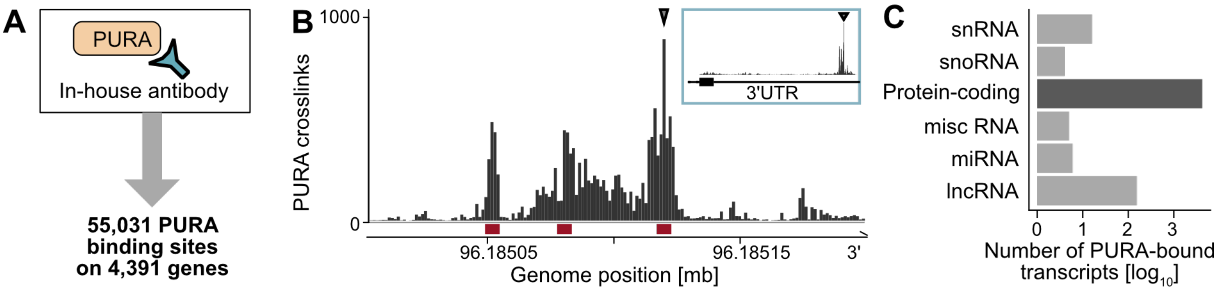


Figure 4.1: Binding site definition on endogenous PURA iCLIP results in over 50,000 binding sites on 4,391 genes.
(A) Schematic on endogenous PURA iCLIP pulled down with anti-PURA in-house antibody.
(B) Exemplary binding sites (red boxes) corresponding to the crosslink profile in the 3'UTR of the *STARD7* RNA. The blue box on the top depicts the complete 3'UTR.
(C) Transcript types bound by PURA. PURA has a preference for protein-coding RNAs (dark-grey).

The support of binding sites by the individual replicates can be demonstrated through pairwise comparisons of the number of crosslinks per binding site between two replicates (**Figure 4.2**). These comparisons reveal that most binding sites have a similar number of crosslinks between the replicates. The number of crosslinks per binding site peaks at around 16 crosslinks (4 on a $\log_2$ scale) for all replicates (**Figure 4.2** density plots on the diagonal). However, replicate 4 and, less pronounced, replicate 2 exhibit a greater proportion of binding sites with nearly zero crosslinks. This trend can be attributed to the lower library depth of replicates 2 and 4, compared to replicates 1 and 3 (**Figure 4.7 B**). Consistent with this observation, the strongest correlation is observed between replicates 1 and 3 (R = 0.71, P value < 0.001, Pearson correlation), while the weakest correlation occurs between replicates 2 and 4 (R = 0.24, P value < 0.001, Pearson correlation). In conclusion, the crosslink distributions of PURA in its binding sites remain comparable across all replicates regardless of the varying library sizes, thus emphasising the reproducibility of the defined PURA binding sites.

In summary, endogenous PURA reproducibly binds 4,391 RNAs in over 50,000 binding sites with a preference for protein-coding RNAs. This reveals a new role of PURA as an RNA-binding protein *in vivo*.

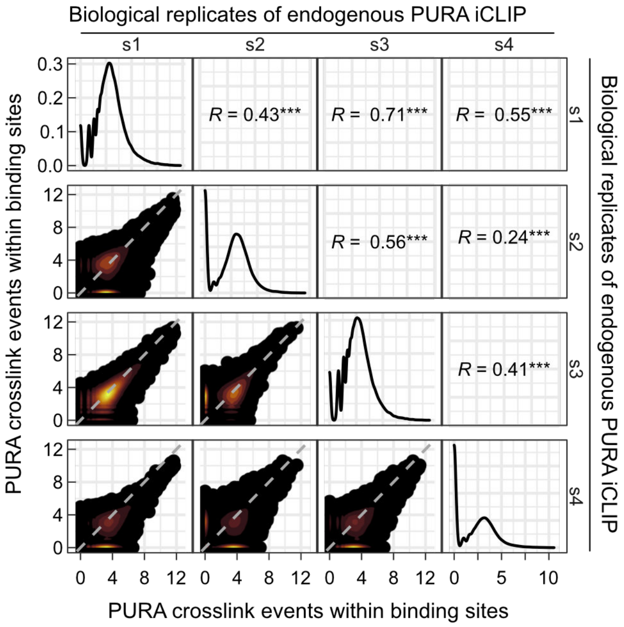


Figure 4.2: Reproducibility of PURA crosslinking signal per binding site. Shown is a matrix of the four replicates from the endogenous PURA iCLIP. In the left triangle of the matrix, binding sites are shown as points and the number of crosslinks in one of the replicates is shown on the x-axis and of a second replicate on the y-axis ($\log_2$ scale). The density of the points is depicted in a colour scale from yellow over red to black and the diagonal is shown as a grey line. In the diagonal plots of the matrix, the density of crosslinks per binding site is shown for each replicate individually. In the right triangle of the matrix, the R values of the pairwise Pearson correlation are given together with their significance (*** - P value < 0.001).

Regional preferences of PURA binding

The functional relevance of an RBP is often connected to the RNA region it binds. The distribution of PURA binding sites across various transcript regions reveals that PURA binds to both the 3' untranslated region (3'UTR) (46.3%) and the coding sequence (CDS) (47.3%) to a similar extent (**Figure 4.3 A**), whereas it rarely binds to the 5' untranslated region (5'UTR) and introns. This pattern is already evident in the crosslink distribution of PURA crosslinks before defining the binding sites (**Figure 4.3 B**).

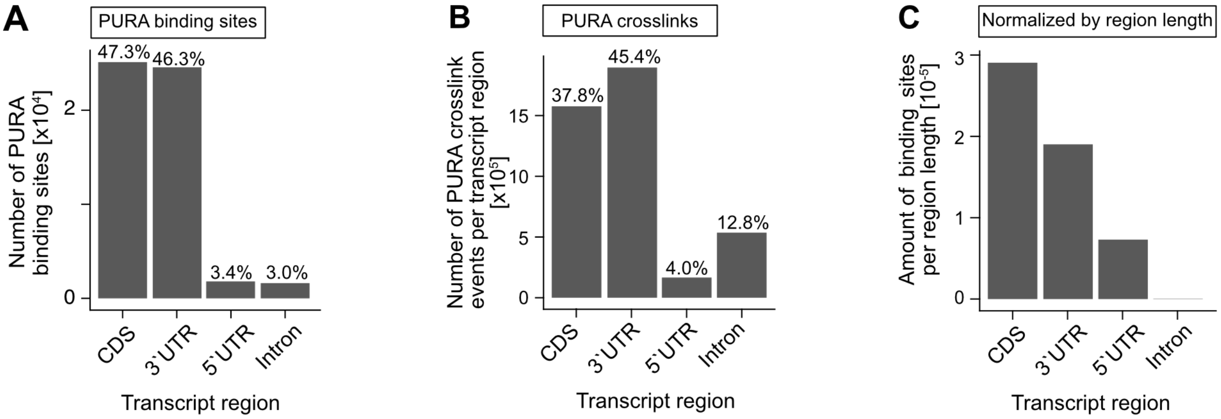


Figure 4.3: Distribution of PURA binding sites and crosslinks in transcript regions.
(A) Distribution of PURA binding sites across transcript regions.
(B) Distribution of PURA crosslinks across transcript regions.
(C) Relative amount of PURA binding sites per region, after normalisation for region length.

When comparing the distribution of binding sites in different regions, the different lengths of the regions can be a confounding factor. For instance, the mean length of 3'UTRs is greater than that of CDS. With a random distribution of binding sites, one would obtain more binding sites in the 3'UTR relative to the CDS. To overcome this, the number of binding sites in each region can be divided by their respective lengths. For the PURA binding sites, there is a rise in the proportion of CDS binding sites compared to 3'UTR binding sites (**Figure 4.3 C**). Moreover, the number of normalised 5'UTR binding sites increases since 5'UTRs are typically short, whereas intronic binding sites decrease owing to the comparatively greater length of introns. In all three different quantifications of PURA RNA-binding regions, there is always little intron binding of PURA. The absence of intron binding implies that PURA binds to mature mRNAs instead of pre-mRNAs. This suggests that PURA predominantly localises in the cytoplasm, as pre-mRNAs are exclusive to the nucleus while mature mRNAs are exported to the cytoplasm. Our collaborators provided evidence for this notion through both immunostainings and nuclear-cytoplasmic fractionation experiments, indicating that PURA is primarily located in the cytoplasm

[Molitor *et al.*, 2023].

In conclusion, 3'UTR and CDS are the most prevalent binding regions of PURA. I was, therefore, interested in comparing PURA binding in those regions and will further characterise the differences between CDS and 3'UTR binding in the next section.

Differences in 3'UTR and CDS binding

Interestingly, a manual inspection of RNAs bound in 3'UTR and CDS regions shows different PURA crosslink patterns. Mainly 3'UTR-bound transcripts such as *STARD7* (*StAR-Related Lipid Transfer Domain Protein 7*), *CTNNA1* (*Catenin Alpha 1*) and *DCP1A* (*mRNA-Decapping Enzyme 1A*) display a few narrow and high peaks resembling a bell shape (**Figure 4.4 A, B, C**); whereas mainly CDS-bound transcripts like *CNOT1* (*CCR4-NOT Transcription Complex Subunit 1*), *YBX1* (*Y-Box Binding Protein 1*) and the later shown *LSM14A* (*LSM14A mRNA processing body assembly factor*) are bound continuously throughout the entire region (**Figure 4.4 D, E & Figure 4.23 A**). Again other transcripts like *SQSTM1* (*Sequestosome-1*) and the later shown *DDX6* (*DEAD-Box Helicase 6*) display both strong binding in CDS and 3'UTR regions (**Figure 4.4 F & Figure 4.23 B**). Therefore, I aimed to investigate the dissimilarities between these two types of PURA binding.

While more than two-thirds of PURA-bound transcripts (n = 2,991) are only bound either in the CDS or the 3'UTR, 1,178 transcripts are bound by PURA in both regions (**Figure 4.5 A**). The strongest binding site per transcript is more often located in the 3'UTR (n = 2,462) compared to the CDS (n = 1,646) (**Figure 4.5 B**). In a global metaprofile over all transcripts, PURA crosslinks are distributed evenly throughout the whole CDS and 3'UTR for both transcripts with the strongest binding in the CDS and the 3'UTR (**Figure 4.5 C**). This shows that although 3'UTR crosslink patterns in individual transcripts form distinct narrow peaks, these peaks occur at different parts of the 3'UTRs in

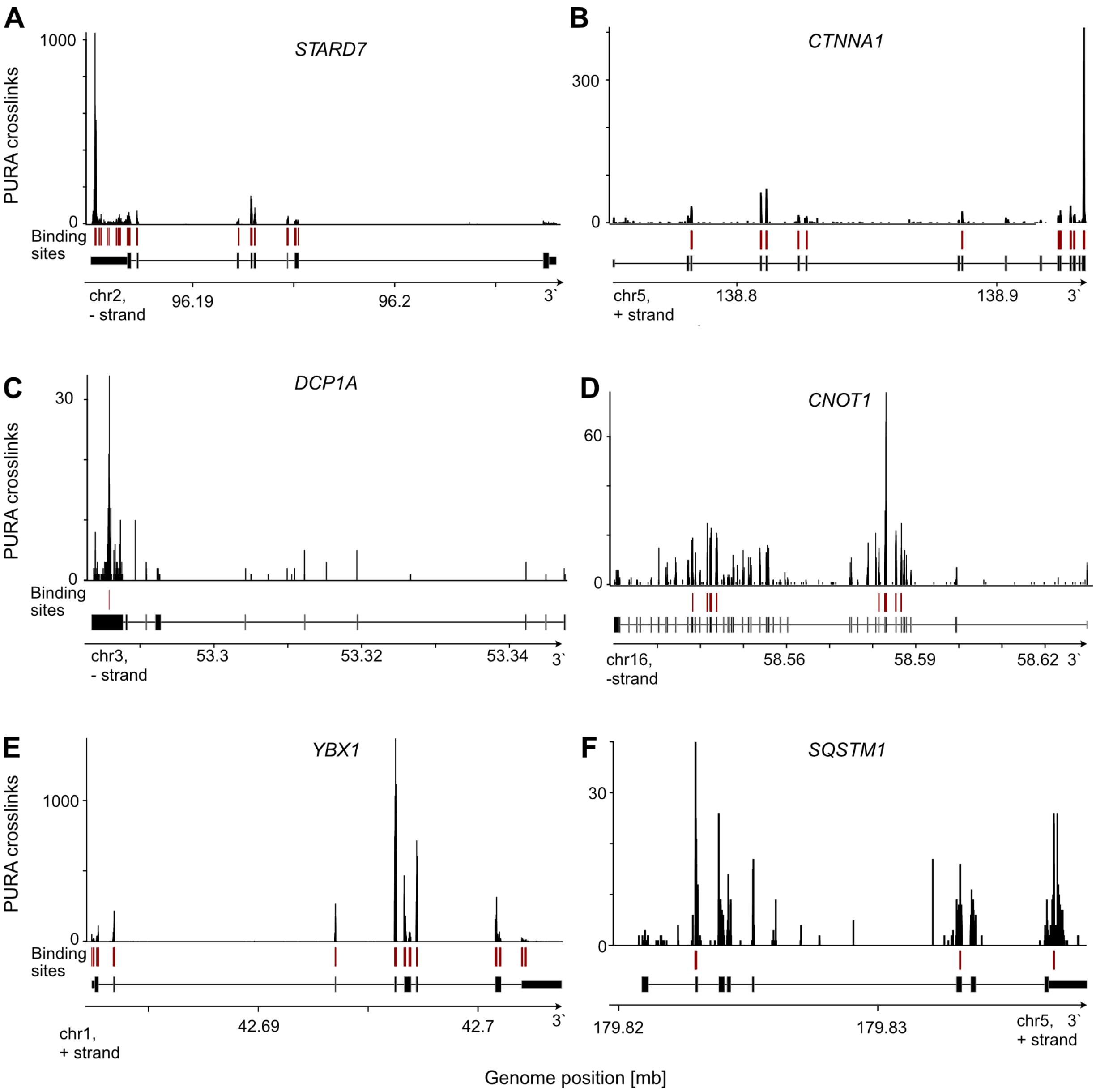


Figure 4.4: PURA binding patterns vary between 3'UTR-bound and CDS-bound transcripts.
Shown are exemplary PURA crosslink profiles on the transcripts of *STARD7* (**A**), *CTNNA1* (**B**), *DCP1A* (**C**), *CNOT1* (**D**), *YBX1* (**E**) and *SQSTM1* (**F**). The number of PURA crosslinks per nucleotide is shown in the bargraphs on the top, red boxes mark the positions of defined PURA binding sites, the transcript architecture (tall box - coding region, stout box - UTR, line - intron) and the genomic position are denoted at the bottom.

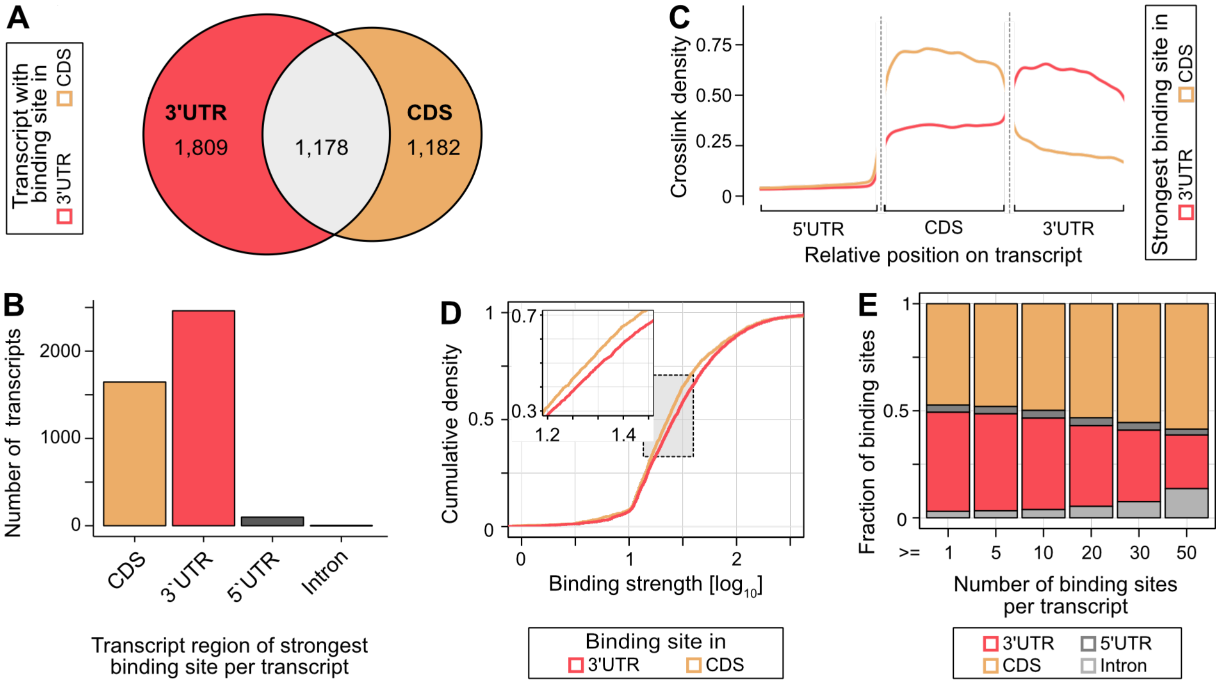


Figure 4.5: Comparison of PURA-binding to 3'UTR and CDS.
(A) Venn diagram showing the overlap of transcripts with a PURA binding site in the 3'UTR (salmon) and CDS (yellow). 1,809 and 1,182 transcripts are only bound either in the 3'UTR or the CDS, respectively, while 1,178 transcripts are bound in both regions.
(B) The position of the strongest PURA binding site is displayed. The number of transcripts is shown against the region encompassing its strongest PURA binding site.
(C) Density of PURA crosslinks across transcript regions. PURA crosslinks are split up by the position of the strongest binding site on the respective transcript. Salmon - 3'UTR; yellow - CDS.
(D) Cumulative density of the binding site strength of all 3'UTR or CDS binding sites ($\log_{10}$ scale). Salmon - 3'UTR; yellow - CDS.
(E) Shown is how the distribution of bound regions changes with the number of binding sites per transcript. The fraction of binding sites located in the different transcript regions (y-axis) is displayed for transcripts with at least a given number of binding sites (x-axis). Salmon - 3'UTR, yellow - CDS, dark grey - 5'UTR, light grey - intron.

different transcripts. There is, therefore, no preference of PURA to bind preferentially in the beginning, middle or end of the 3'UTR but the position of crosslink peaks is randomly distributed over the different transcripts. Similarly, CDS crosslinks have no preference for a certain position in the CDS but are distributed throughout as well. The overall strength of the binding sites is slightly higher for 3'UTR binding sites compared to CDS binding sites (**Figure 4.5 D**). On the other hand, transcripts with many binding sites tend to accumulate CDS binding sites (**Figure 4.5 E**). These observations result in a picture of more narrow RNA binding in 3'UTR binding sites and broader binding spread over the CDS of the same or other transcripts.

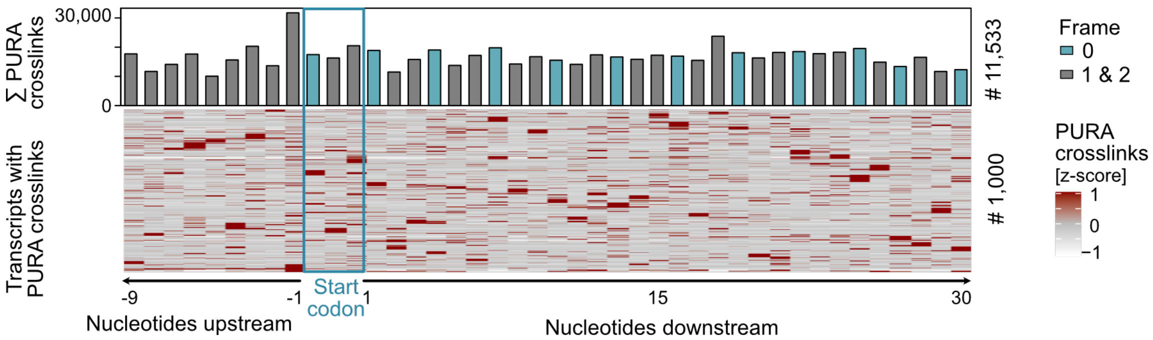


Figure 4.6: PURA crosslinks in the region of the start codon do not display a ribosome-like pattern.
The heatmap (bottom) depicts 1,000 transcripts with PURA crosslinks in the region from 9 nt before until 30 nt after the start codon. The number of PURA crosslinks is normalised in a row-wise z-score that is depicted as a colour scale. The position of the start codon is marked by the blue box. The metaprofile (top) displays the sum of PURA crosslinks across all 11,533 transcripts with any crosslinks in this window at the given position. Beginning from the start codon, the first position of each codon is coloured in blue.

One possible explanation for the widespread CDS binding could be a connection between PURA and ribosome activity, given that the PURA crosslink profiles in the CDS are spread over the complete length of the CDS. If PURA bound RNAs in a complex with ribosomes while the ribosomes move along the CDS, crosslink signals would appear wherever the ribosomes are located during UV radiation. However, close examination of the crosslink distribution around the start codon does not show similarities with a typical ribosomal pattern that recurrently crosslinks every three nucleotides (**Figure 4.6**) [Ingolia, 2016]. Therefore, it is unlikely that the observed binding of PURA in the CDS is linked to ribosomal activity. (Please note that this feature will be revisited in **Chapter 4.3.2**.)

Comparison of different PURA iCLIPs

The PURA iCLIP experiment described until now was performed by our collaborators in four replicates from HeLa cells expressing PURA endogenously (endogenous PURA iCLIP). Furthermore, they performed two iCLIP experiments in HeLa cells with an ectopic overexpression of a FLAG-PURA fusion protein (**Figure 4.7 A**). In one of the two, PURA was pulled down with the in-house anti-PURA antibody (FLAG-PURA IH-AB iCLIP, comprising four replicates), in the other with a commercial anti-FLAG antibody (FLAG-PURA FLAG-AB iCLIP, comprising two replicates). The two additional experiments aid in verifying the proper function of the PURA in-house antibody. In addition, they can be used to look for differences in PURA RNA-binding when overexpressed. Please note that the numbers of endogenous PURA binding sites in this section differ slightly from the endogenous PURA binding sites reported in all other sections as this analysis was performed at a later time point and the gene annotation was filtered more stringently.

The library depth varied between the experiments, resulting in two to four times more crosslinks in the FLAG-PURA IH-AB iCLIP compared to the endogenous PURA iCLIP (**Figure 4.7 B**). Still, the crosslinking patterns of all three experiments look very similar in a manual comparison (**Figure 4.7 C**). However, in a union of 144,370 binding sites that are found in any of the three experiments a considerable amount of binding sites is only found in one of the three experiments (endogenous PURA - 25,541 of 54,867, FLAG-PURA IH-AB - 64,286 of 103,260, FLAG-PURA FL-AB - 13,993 of 33,118) (**Figure 4.8 A**). To better understand this effect, I compared crosslink patterns in all 3'UTRs and CDS with intermediate crosslink coverage (10^2-10^6 crosslink events per window) across

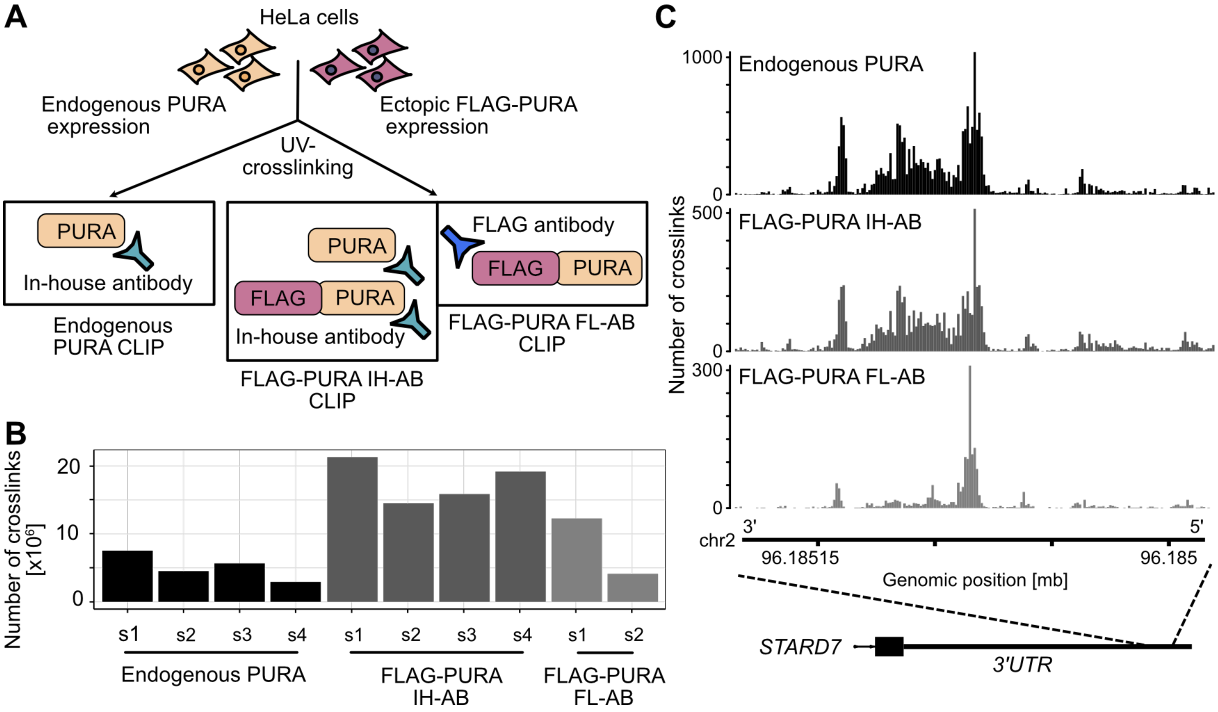


Figure 4.7: PURA iCLIP experiments elucidate its binding capacity.
(A) Three iCLIP experiments were performed in HeLa cells with two different antibodies. In the first experiment, endogenous PURA was pulled down with a specifically made anti-PURA in-house antibody (endogenous PURA CLIP). In the second experiment, the in-house antibody was used in a HeLa cell line that ectopically expressed a FLAG-PURA fusion construct (FLAG-PURA IH-AB CLIP) and in the third an anti-FLAG antibody was used in the same cell line (FLAG-PURA FL-AB CLIP).
(B) Number of PURA crosslinks for the replicates from the three experiments, black – PURA CLIP, dark grey - FLAG-PURA IH-AB CLIP, light grey – FLAG-PURA FL-AB CLIP.
(C) An exemplary crosslink profile in the 3'UTR of the *STARD7* RNA of all three experiments, top, black – endogenous PURA CLIP; middle, dark grey - FLAG-PURA IH-AB CLIP; bottom, light grey – FLAG-PURA FL-AB.

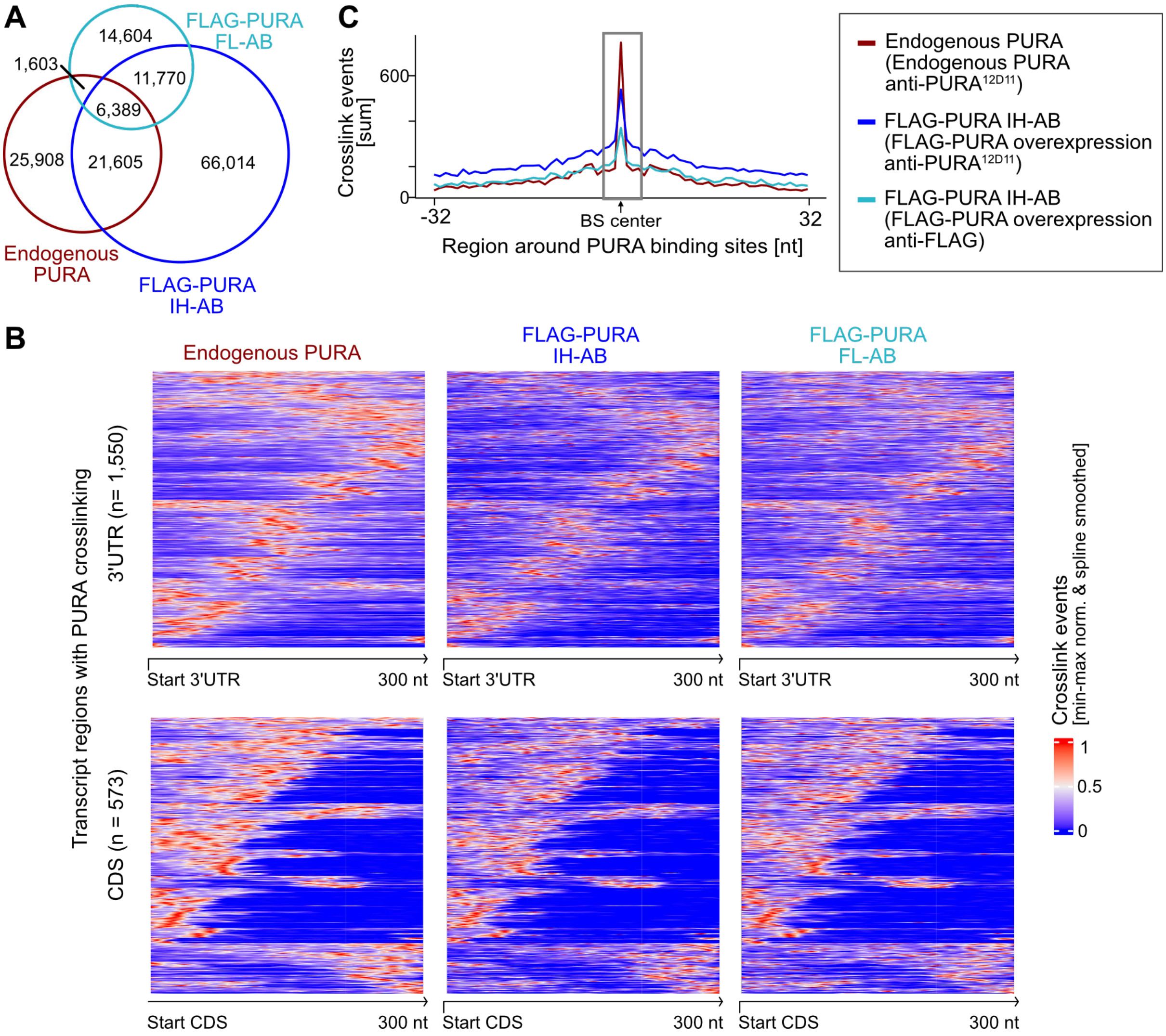


Figure 4.8: FLAG-PURA overexpression leads to a higher background signal in PURA binding compared to endogenous PURA.
(A) Overlaps of the binding sites defined on the three iCLIP sets: endogenous PURA (n = 55,505, red), FLAG-PURA IH-AB (n = 105,778, dark blue), FLAG-PURA FL-AB (n = 34,366, light blue).
(B) Heatmaps of crosslink patterns in the first 300 nt of all 3'UTRs (top, n = 1,550) or CDS (bottom, n = 573) that had intermediated crosslink coverage (10^2-10^6 crosslink events per window). Shown are the crosslink patterns from endogeous PURA (left), FLAG-PURA IH-AB (middle) and FLAG-PURA FL-AB (right) after min-max normalisation and spline smoothing.
(C) Metaprofile of crosslinks around endogenous PURA binding sites. Shown is a window from 32 nt before until 32 nt after the binding site. Red - endogenous PURA, dark blue - FLAG-PURA IH-AB, light blue - FLAG-PURA FL-AB.

all three experiments (**Figure 4.8 B**). The overall patterns are similar, but it is also apparent that there are more highly or poorly crosslinked positions for endogenous PURA in comparison to the two FLAG-PURA experiments. A metaprofile of crosslinks surrounding endogenous PURA binding sites displays reduced crosslinks in the binding sites‘ centre and increased crosslinks in their surrounding area for the FLAG-PURA experiments (**Figure 4.8 C**). Taken together, although the binding sites appear to be quite different in the three experiments, the crosslinking patterns are very similar but contain a higher background signal in the FLAG-PURA experiments.

The observation that the crosslinking distributions of all three experiments peak in the middle of the endogenous PURA binding sites shows that the in-house anti-PURA antibody and the anti-FLAG antibody detect the same crosslinking patterns. This indicates that the in-house anti-PURA antibody is suitable for iCLIP experiments. The major advantage of the in-house antibody is that it does not cross-react with the PURA paralog PURB ([Molitor *et al.*, 2023]) and, therefore, the defined RNA binding sites can be confidently attributed to PURA alone.

To delve deeper into the differences of endogenous PURA binding and binding of overexpressed FLAG-PURA, I analysed pairwise differential binding of FLAG-PURA IH-AB against endogenous PURA (**Figure 4.9 A**) and FLAG-PURA FL-AB against endogenous PURA (**Figure 4.9 B**). I found that a large number of binding sites vary significantly across experiments (P value < 0.01, test performed by DESeq2 method, P value is adjusted by IHW method; FLAG-PURA IH-AB against endogenous PURA - n = 65,396; FLAG-PURA FH-AB against endogenous PURA - n = 46,020). In both comparisons, the binding strength increases significantly in the PURA overexpression condition in the majority of binding sites (**Figure 4.9 A, B**), with 83.2% (54,377 / 65,396) being bound with increased strength in the IH-AB data set and 82.8% (38,083 / 46,020) in the FL-AB data set.

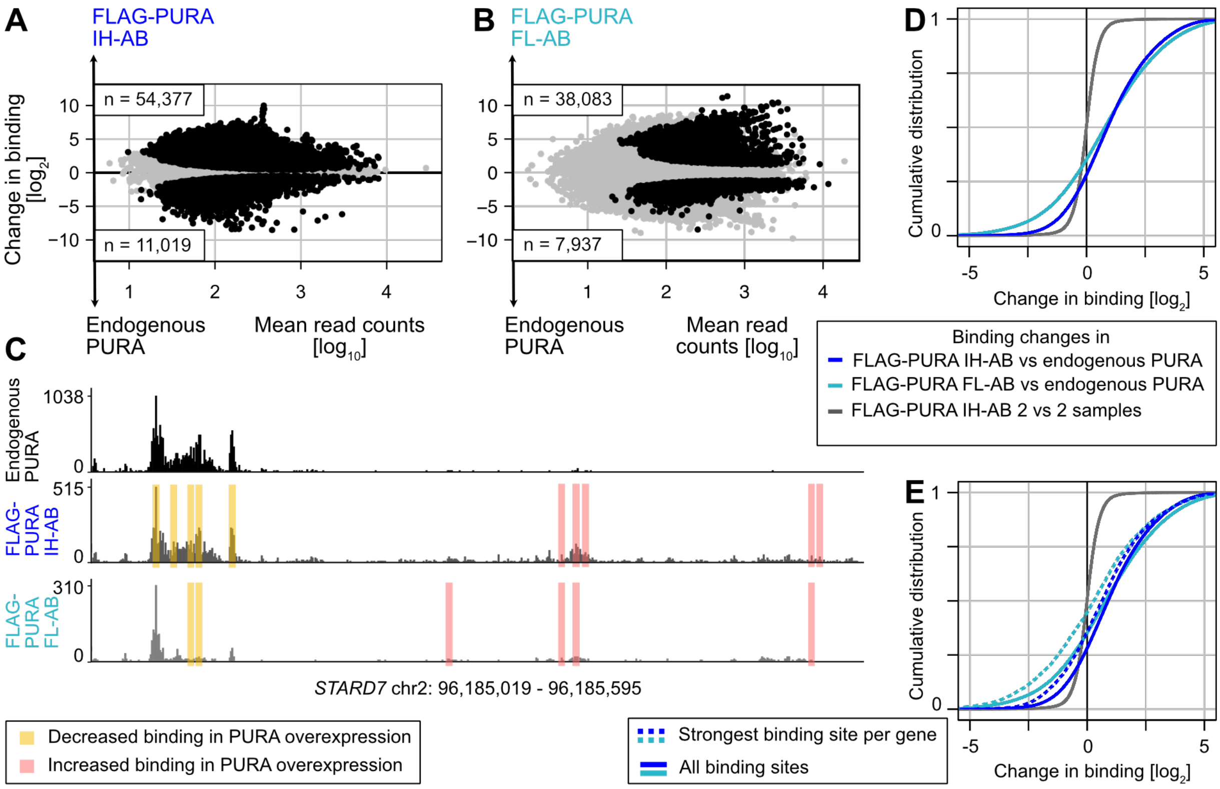


Figure 4.9: PURA binding changes considerably between the PURA iCLIPs. **(A, B)** Differential binding of FLAG-PURA IH-AB against endogenous PURA (A) and FLAG-PURA IH-AB against FLAG-PURA FL-AB (B) in MA-plots. Shown are the mean read counts per binding site on the x-axis ($\log_{10}$ scale) and the change in binding signal on the y-axis ($\log_2$ scale). The number of binding sites with increased binding (top) and decreased binding (bottom) are given. Black - significantly changing binding sites (P value < 0.01, test performed by DESeq2 method, P value is adjusted by IHW method), grey - non-significant.
(C) Genome browser view from the *STARD7* RNA. Depicted are the crosslink profiles from FLAG-PURA IH-AB (top) and endogenous PURA (bottom) and binding sites changing in the comparison of FLAG-PURA IH-AB against endogenous PURA. Yellow - decreased in FLAG-PURA conditions; pink - increased in FLAG-PURA conditions. Please note, that the y-axis are scaled differently.
(D, E) Cumulative distribution of binding changes. Dark blue - FLAG-PURA IH-AB against endogenous PURA; light blue - FLAG-PURA FL-AB against endogenous PURA; grey - FLAG-PURA IH-AB replicates 1 and 2 against replicates 3 and 4; dotted lines - strongest binding site per gene.

An example of this change in binding can be observed in the binding sites in the *STARD7* 3'UTR (**Figure 4.9 C**). Multiple weaker new binding sites emerge in FLAG-PURA IH-AB and FLAG-PURA FL-AB. These sites are almost absent in endogenous PURA and are thus strongly upregulated in the FLAG-PURA data sets. In contrast, the strongest binding sites are bound significantly less in FLAG-PURA IH-AB. This also holds true on a global level when comparing the cumulative distributions of the changes in PURA binding (**Figure 4.9 D**). As a negative control, I performed a differential binding analysis on replicates 1 and 3 against replicates 3 and 4 from the FLAG-PURA IH-AB experiment resulting in only 4 significantly changing binding sites (P value < 0.01, data not shown) and a cumulative distribution of the binding changes close to zero. In comparison, the binding changes for both FLAG-PURA IH-AB and FLAG-PURA FL-AB against endogenous PURA are strongly shifted towards more binding in the overexpression conditions. In addition, the changes of the strongest binding site per gene are shifted towards weaker binding in the PURA overexpression conditions in comparison to all changes (**Figure 4.9 E**).

Despite these differences in PURA binding, binding sites defined in the three experiments always show a preference for PURA to bind the 3'UTRs and CDS of protein-coding transcripts (**Figure 4.10 A, B**). The same holds true when selecting only binding sites found in all three experiments. Also, the majority (n = 3,774) of bound genes are found both in the endogenous and the IH-AB condition, while 512 genes are only bound by endogenous PURA and 1,576 are bound only in one of the two FLAG-PURA conditions (**Figure 4.10 A, B**). The IH-AB and FL-AB conditions overlap greatly, although fewer genes are bound in the FL-AB condition, which is likely due to the lower signal depth of the experiment. Overall, these comparisons strengthen the notion that PURA binds 3'UTRs and CDS of protein-coding genes.

In summary, FLAG-PURA overexpression leads to similar PURA binding patterns, but binding sites become less distinct due to more crosslinking in the surroundings of the binding sites. The differential binding analysis shows that this increased background signal even leads to the detection of additional binding sites in the surroundings of strong

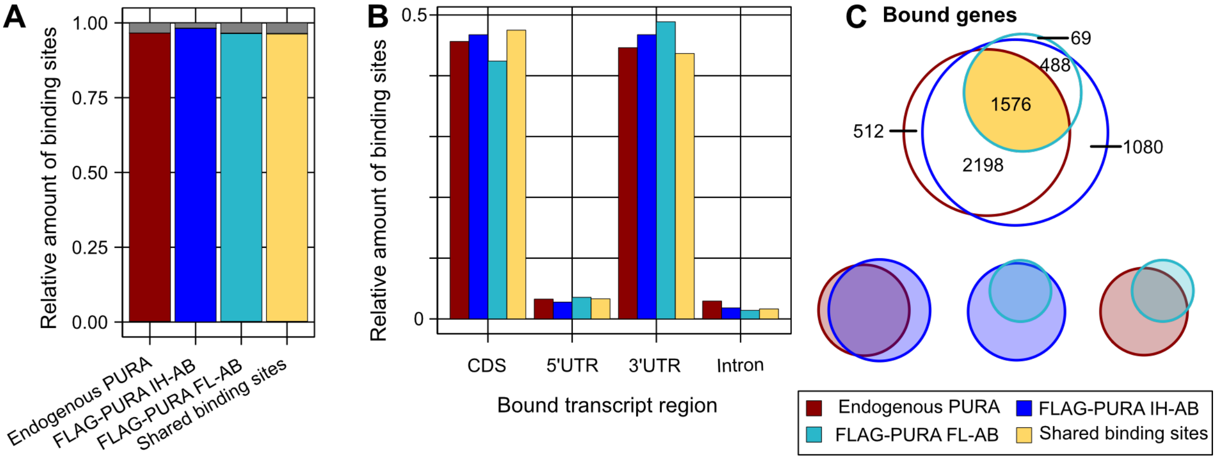


Figure 4.10: PURA binding characteristics stay the same across all data sets.
(A) Given is the relative amount of bound gene types among the bound genes from the three iCLIP experiments and the overlapping binding sites from all three data sets (coloured - protein-coding, grey - other; endogenous PURA - n = 4,355, FLAG-PURA IH-AB - n = 5,531, FLAG-PURA FL-AB - n = 2,242).
(B) Given is the relative amount of binding sites across transcript regions. Red - endogenous PURA; dark blue - FLAG-PURA IH-AB; light blue - FLAG-PURA FL-AB; yellow - Shared binding sites.
(C) Top - Overlap of bound genes between the three experiments. Bottom - Pairwise comparisons of bound genes.

binding sites, that have not been present in the endogenous PURA data. Furthermore, the binding strength in the highest binding site per gene globally decreases in FLAG-PURA overexpression.

It can, therefore, be concluded that FLAG-PURA overexpression leads to a reduced binding specificity of PURA to its target transcripts, resulting in a decrease of specific binding sites and an increase of binding in the surrounding area. Although the emergence of new binding sites in the FLAG-PURA IH-AB experiment might be partially a technical effect due to the higher signal depth of the data set, this is not the case for the FLAG-PURA FL-AB experiment. Therefore, the FLAG-PURA FL-AB experiment underlines that FLAG-PURA in an overexpression indeed binds differently to endogenous PURA. The difference in binding between endogenous PURA and overexpressed FLAG-PURA, could be due to the FLAG fusion to the PURA protein or due to the overexpression of

PURA levels in the cell. On the one hand, the unphysiologically high PURA in the cell could lead to an increase of unspecific PURA binding throughout the transcripts. On the other hand, the FLAG part of fusion to the protein could disturb PURA binding to some degree, making binding less specific. Furthermore, differences in binding could also be due to differential gene expression in the two conditions. A transcript might, therefore, not be bound in one or the other because it's not expressed in the cells in this condition. This is accounted for in the calculation of differential binding (**Figure 4.9**) - which includes the background crosslinks per gene as estimate of expression into the model - but for example not in the overlaps of binding sites (**Figure 4.8 A**) or bound genes (**Figure 4.10 C**).

PURA's sequence preference

PURA derives its name, purine-rich binding element alpha, from its preference for binding purine-rich DNA and RNA sequences. However, a concrete motif for PURA binding has not yet been described. Consequently, I aimed to determine the sequence preferences of PURA and to address whether PURA binding to RNAs could be attributed to a defined PURA binding motif.

PURA binding sites demonstrate enrichment in purines (**Figure 4.11 A**) with 61% of PURA binding sites containing at least three purines. A sequence logo over the positions of all PURA binding sites (**Figure 4.11 B**) displays a sequence pattern that is not particularly specific. There is a general increase in the frequency of guanines and adenines, as well as uridines. Comparable outcomes are achieved in independent analyses for only binding sites in 3'UTR or CDS (data not shown).

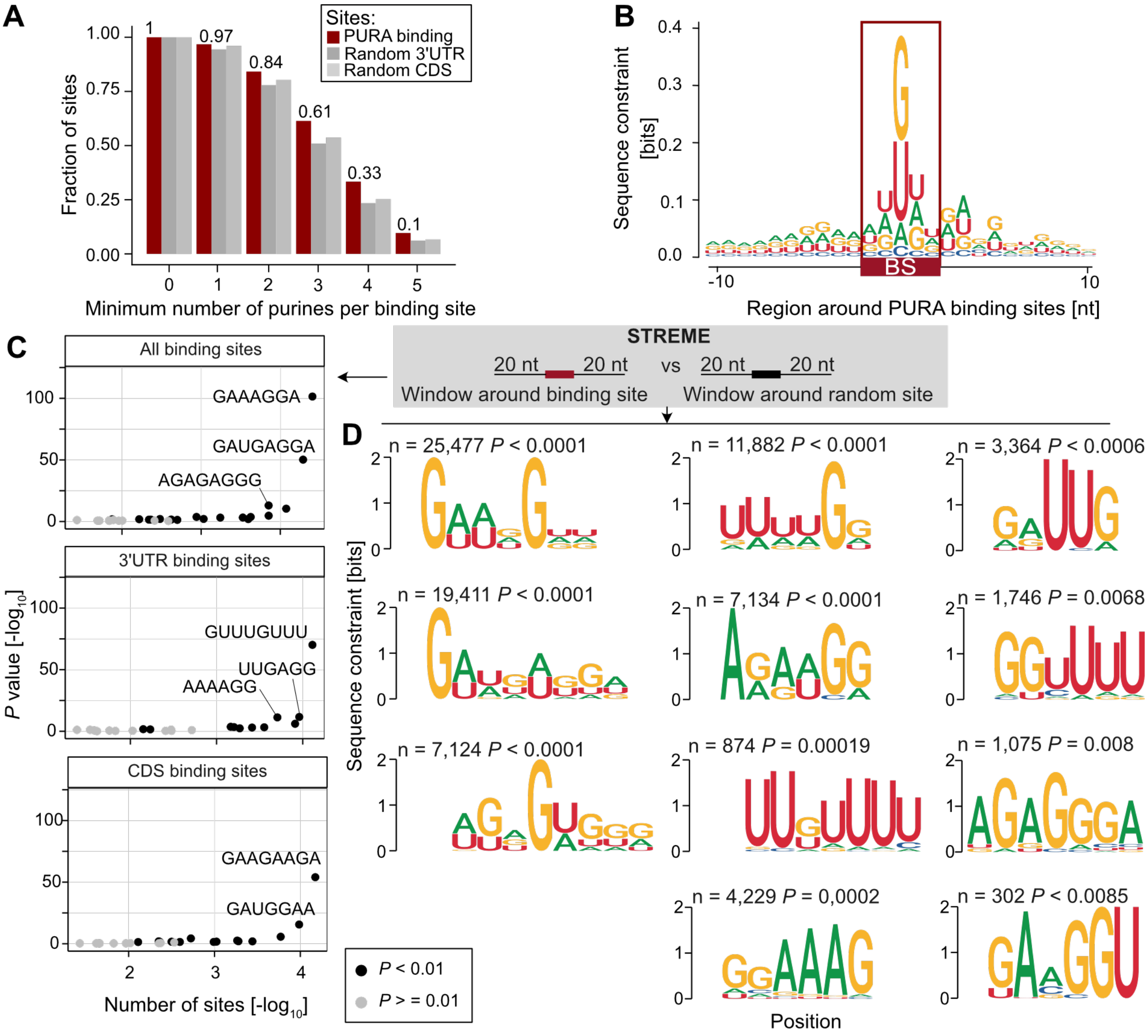


Figure 4.11: PURA binds a degenerate purine-rich motif.
(A) Shown is the fraction of PURA binding sites (red), random 3'UTR (dark grey) and random CDS (light grey) sites containing a minimum number of purines (x-axis). The fraction of PURA binding sites is given on top of the bars. Random sites (3'UTR - n = 23,436, CDS - n = 24,627) are 5 nt long like the PURA binding sites and are picked from the 3'UTRs and CDSs of all expressed transcripts.
(B) Sequence logo on PURA binding sites (red box) together with the 10 nt upstream and downstream of the binding sites.
(C) *De novo* motifs detected with STREME on 45 nt centred on the PURA binding sites. The number of sites per motif ($\log_2$ scale) is shown against the P value ($\log_{10}$ scale, Fisher's Exact Test computed in STREME). The top three motifs are labels with their seed sequence. Top - all PURA binding sites, middle - PURA binding sites in 3'UTRs, bottom - PURA binding sites in CDS, black dots - motif with a P value < 0.01, grey dot - motif with a P value ≥ 0.01.
(D) Sequence logos of motifs detected by STREME for all PURA binding sites. All significant logos with a P value < 0.01 are shown.

Next, I used *de novo* motif prediction on a 45-nt window centred on the PURA binding sites resulting in several degenerate motifs consisting of guanines, adenines and uridines (**Figure 4.11 C, D**). The three most enriched sequence logos have consensus sequences consisting of several guanines interspaced with adenines and uridines. 3'UTR binding sites have 'GUUGUU' as their strongest enriched consensus sequence, while CDS binding sites have 'GAAGAAGA'. However, since the logos for these motifs are degenerated (**Figure 4.11 D**), it is difficult to determine whether the observed difference is a result of the bound regions or simply a random fluctuation of the same degenerate motif.

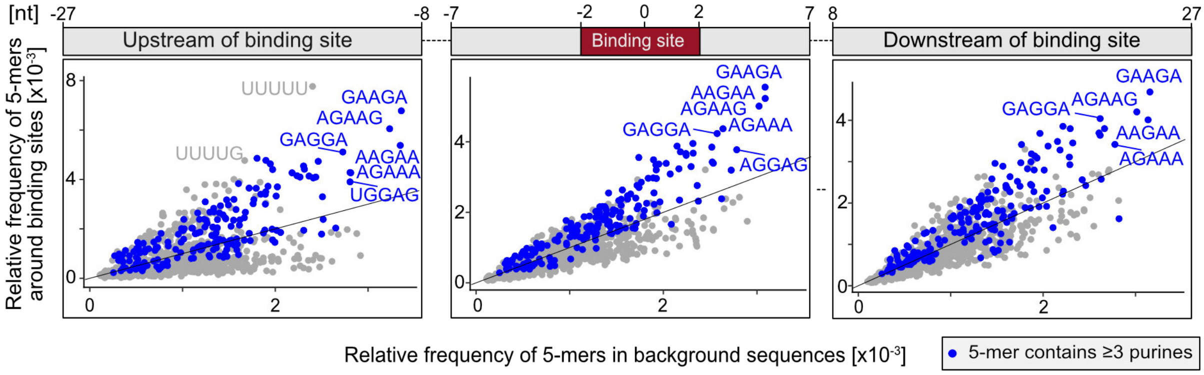


Figure 4.12: 5-mer sequences containing five purines are most enriched in and around PURA binding sites. Shown is the enrichment of 5-mers in three different regions around PURA binding sites: within 7 nt before and after the PURA binding site centre (middle), followed by a window of 20 nt before the central window (upstream window, left) and a window of 20 nt after the central window (downstream window, right). The enrichment of the 5-mers around random positions on expressed transcripts (x-axis) is displayed against the enrichment around PURA binding sites (y-axis). The position of the three windows relative to the binding site (red box) is annotated on the top. 5-mers containing ≥3 purines are coloured blue.

In a further in-depth approach (**Figure 4.12**), I calculated the enrichment of 5 nt long sequence combinations (5-mers) in three different regions around PURA binding sites: within 7 nt before and after the PURA binding site centre, a window of 20 nt before the central window (upstream window) and a window of 20 nt after the central window (downstream window). In general, 5-mers with more than three purines were most enriched in all three windows. The strongest enrichment can be found for 5-mers consisting of different combinations of five purines also in all three windows. Only the upstream window shows an additional enrichment for five uridines or four uridines with one guanine in the beginning.

The uridine enrichment that can be observed in and around the PURA binding sites and also in the *de novo* motif discovery is likely due to a previously described bias of iCLIP experiments [Sugimoto *et al.*, 2012, Krakau *et al.*, 2017]: UV crosslinking works best on uridines, therefore, uridines often appear more enriched in CLIP experiments than in other methods. This effect was for example also described for CPSF6 in [Ghanbari and Ohler, 2020]. To summarise, I verified PURA's preference to bind to purine-rich sequences via a degenerate motif [Graebsch *et al.*, 2009, Weber *et al.*, 2016].

Structural accessibility of PURA binding sites

As the binding motif of PURA is rather degenerate, I wondered whether PURA recognises its targets additionally by a specific secondary structure. For this, I used RNAfold prediction on the 500 nucleotides around PURA binding sites and compared it to a permutation background using 1,000 sets of 1,000 sequences from 10,000 randomly picked transcript sequences.

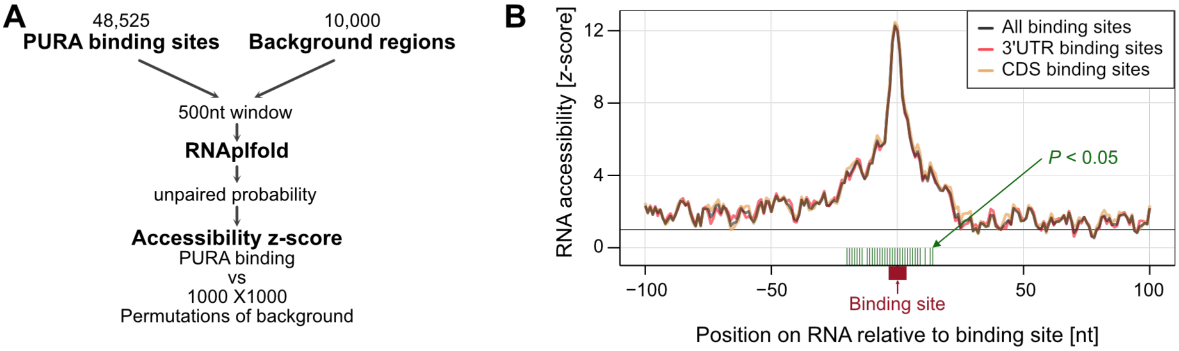


Figure 4.13: PURA preferentially binds single-stranded RNAs. RNA accessibility of bound RNAs is predicted with RNAplfold and given as a z-score using random sequences as background. Shown is the region from 100 nucleotides before until 100 nucleotides after the PURA binding site (red box). Positions with a P value < 0.05 (calculated from z-score) are marked with green bars. Grey - all binding sites, pink - binding sites in 3'UTRs, yellow - binding sites in CDS

This approach shows a significantly increased probability of the PURA binding sites being single-stranded. The same pattern can also be observed for the subgroups of 3'UTR binding sites and CDS binding sites. There is no region in the surroundings of the PURA binding sites that has an increased probability of being double-stranded, which would hint to a specific secondary structure. In conclusion, PURA binding sites are preferentially single-stranded.

Summary

Taken together, I could establish that PURA is a global RNA-binding protein that binds 4,391 transcripts in over 50,000 binding sites that are located prevalently in the 3'UTR and CDS of protein-coding transcripts. PURA RNA binding has a preference for single-stranded purine-rich sequences. Furthermore, overexpression of PURA leads to a decreased specificity of PURA binding. The widespread binding of PURA to cellular RNAs leads to the question of what the cellular function of PURA RNA binding might be, which will be investigated further in the next sections.

4.3.2 The impact of reduced PURA levels on RNA and protein expression

To examine the cellular function of PURA, I analysed the impact of reduced PURA levels on the global RNA and protein expression in HeLa cells. PURA protein levels were reduced by *PURA* siRNA knockdown and the reduction of PURA protein levels to around 50% was confirmed in Western blots [Molitor *et al.*, 2023]. In the following, I describe how the RNA and proteomics expression changes were analysed and then integrated with the previously defined PURA binding sites.

RNA and protein expression changes in PURA-depleted cells

I analysed RNA expression in an RNAseq experiment from both control cells and *PURA* siRNA knockdown cells using four replicates each (**Figure 4.14 A**). A total of 3,415 RNAs are significantly deregulated (P value < 0.01, Benjamini–Hochberg adjusted, **Figure 4.14 B**). Deregulated RNAs are evenly distributed between up-regulated (1,663, 48%) and down-regulated (1,752, 52%). Of note, the transcript of the PURA paralog PURB does not change in its expression and the transcript of the second PURA paralog PURG is not expressed in HeLa cells.

I also examined alternative splicing between these conditions and discovered that *PURA* depletion has little influence on splicing, as demonstrated by only 103 significantly regulated local splice variants (P|δPSI>0.05| $>$ 95%, data not shown). Therefore, I conclude that PURA does not function as a splicing regulator and instead focus on its extensive effect on RNA levels.

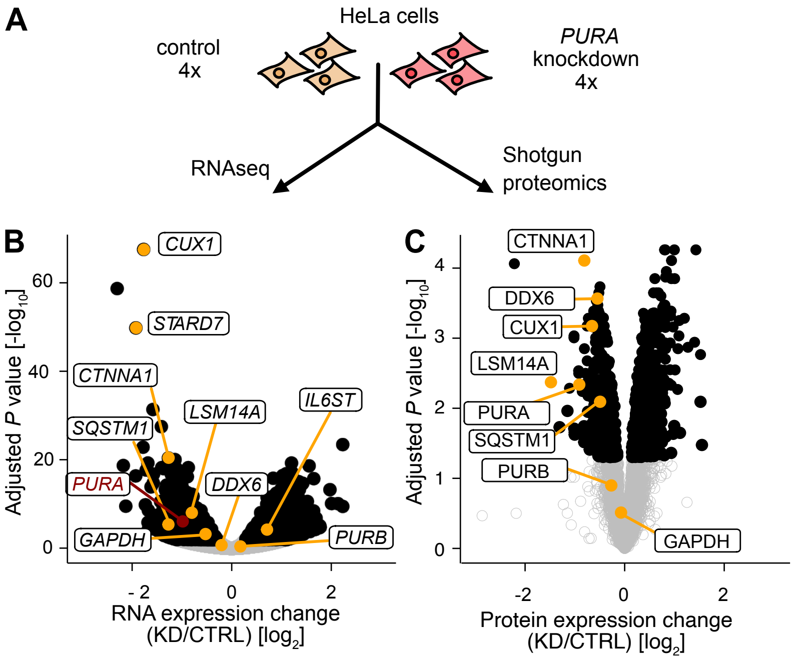


Figure 4.14: *PURA* depletion results in global dysregulation of RNA and protein expression.
(A) Schema of performed experiments. RNAseq and shotgun proteomics were performed on four *PURA* knockdown and control samples.
(B, C) Volcano plot of RNA (B) or protein (C) expression changes upon *PURA* knockdown against control in HeLa cells. Shown is the expression change between *PURA* knockdown and control on the x-axis ($\log_2$ scale) and the P value (-$\log_{10}$ scale, adjusted using Benjamini–Hochberg method) on the y-axis. Black - significant RNAs (P value $<$ 0.01) or proteins (P value $<$ 0.05); orange - selected targets with labels; grey - all other RNAs or proteins.

Similarly, I compared protein expression data from *PURA* knockdown and control cells, each with four replicates (**Figure 4.14 C**). As a result of the analysis, I observed that 995 proteins are dysregulated in the *PURA* knockdown (P value < 0.05, Benjamini–Hochberg adjustment). Furthermore, PURA-regulated proteins demonstrate an even distribution between up- (589, 59%) and down-regulation (406, 41%). Again, PURB expression remains unchanged and PURG is not detected. While it could be hypothesised that PURB and PURG, whose function is little known and who have similar PUR-domains as PURA, could have a similar function to PURA, the unchanged PURB and PURG levels suggest that PURB and PURG can not directly compensate for PURA functionality. Overall, both experiments indicate that a decrease of PURA to half of its physiological amount has a global effect on RNA and protein expression.

The relation of RNA and protein changes in *PURA* depletion

Next, I compared the alterations in RNA and protein expression in *PURA* knockdown. This comparison may reveal possible impacts on RNA translation, in case protein levels vary differently from RNA levels.

Firstly, the RNA and protein expression of targets that are significantly regulated on both RNA and protein levels (**Figure 4.15 A**) indicates that protein levels in general follow RNA levels. Out of the 334 targets that show significant regulation on both RNA and protein levels, 249 are regulated in the same direction (149 up- and 100 down-regulated).

Examining all targets detected in both conditions (**Figure 4.15 B**), it becomes apparent that a considerable number of targets (3,081 RNAs and 662 proteins) change significantly in only one of the data sets. However, it should be noted that the comparison between RNAseq and proteomics data can be limited because high-throughput sequencing and mass spectrometry vary considerably in their detection methodologies which may result in superior or inferior detection of certain targets. To overcome the limitations of statistical significance in both sets, I examined the relationship between all RNA and protein changes (**Figure 4.15 C**). Notably, this exhibits a similar trend of protein levels following RNA levels, which might suggest that PURA does not affect RNA translation.

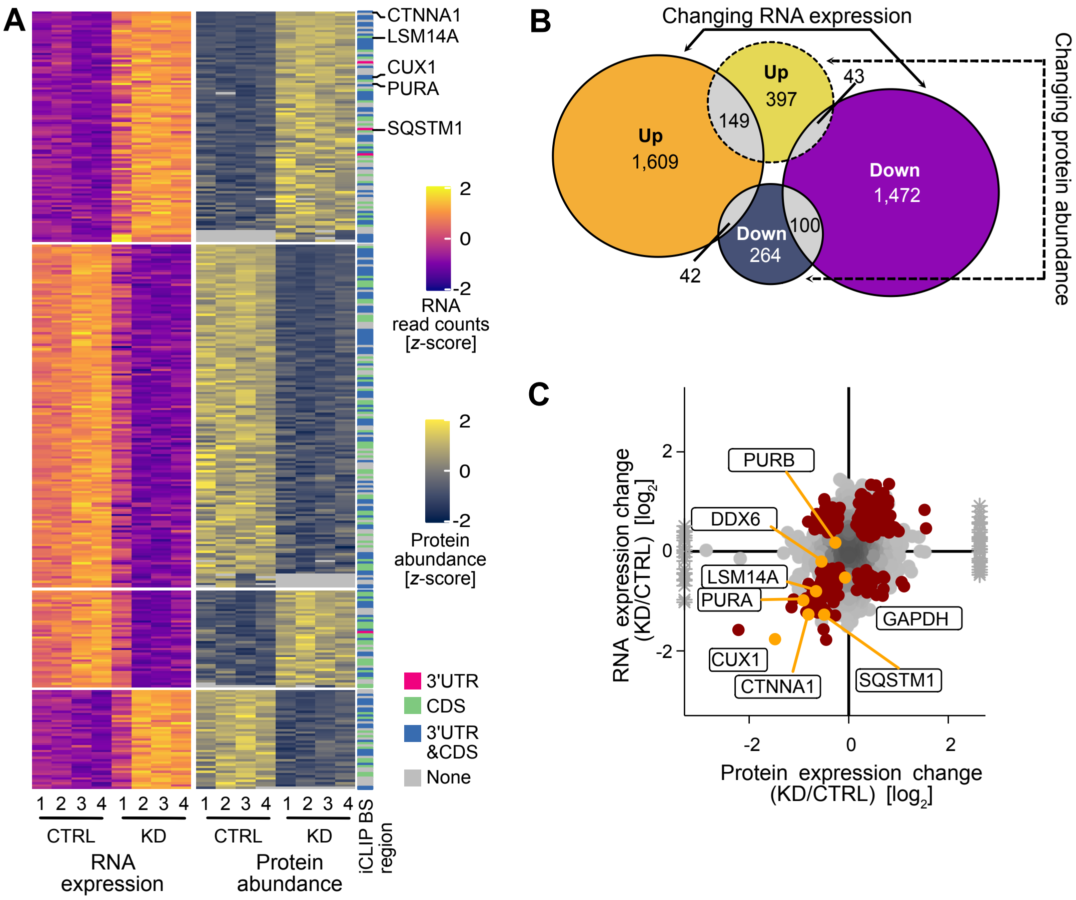


Figure 4.15: In *PURA* depletion, protein changes tend to follow RNA changes.

(A) Heatmap depicting the expression of all targets changing significantly in both RNA and protein expression. Given are the z-scores of RNA and protein expression across all samples using an orange to purple colour scale for RNA expression and a yellow to dark blue colour scale for protein expression. Targets are grouped by the direction of their expression with 100 targets upregulated on RNA and protein level (first block), 149 targets downregulated on RNA and protein expression (second block), 42 targets downregulated on RNA and upregulated on protein level (third block) and 43 targets up-regulated on RNA and downregulated on protein level. On the right, the region of PURA binding in the respective RNA is shown (pink - 3'UTR, green - CDS, blue - both 3'UTR and CDS, grey - neither 3'UTR or CDS). Selected targets are annotated.

(B) Venn diagram displaying the overlap of significantly regulated RNAs (orange - upregulated, n = 1,800; purple - downregulated, n = 1,615) and proteins (yellow - upregulated, n = 589; dark blue - downregulated, n = 406).

(C) Comparison of overall RNA (y-axis) and protein (x-axis) expression changes. Red - significantly changing targets; orange - selected targets with labels; Stars - values with an infinite protein change, which is the result of a protein being detected only in one of the two conditions.

Expression changes of PURA-bound and unbound RNAs and proteins

The observed global changes can broadly be categorised as direct or indirect effects of PURA. A direct effect refers to RNAs or proteins that PURA acts upon directly, while indirect effects are secondary effects that result from the direct effects of PURA on another target. Here, I used PURA binding as an indicator of PURA's direct effect on RNA to distinguish between direct and indirect effects and search for differences between them.

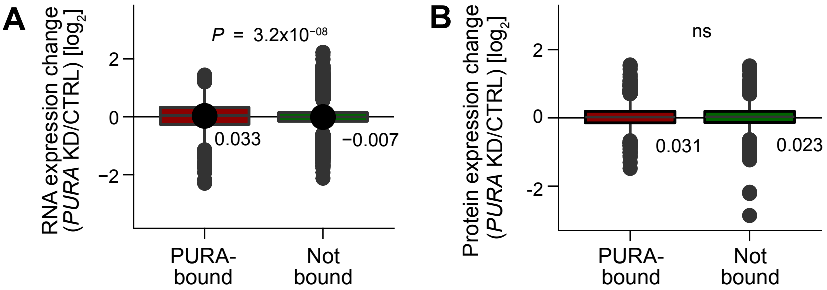


Figure 4.16: PURA-bound and unbound targets have similar overall changes in both RNA and protein levels.
Shown are the RNA **(A)** and protein expression changes **(B)** of *PURA* knockdown against control for PURA-bound (red) and not bound (green) RNAs as boxplots ($\log_2$ scale). The mean expression change is annotated below the respective box. P value is obtained from the Wilcoxon rank-sum test.

There is no noticeable overall difference in changes between PURA-bound (direct) and unbound (indirect) targets for both RNA (**Figure 4.16 A**) and protein (**Figure 4.16 B**). PURA-bound RNAs exhibit a mean expression change of 0.033, whereas non-PURA-bound RNAs show a mean expression change of -0.007. Although this results in an extremely low P value (P value < 0.0001, Wilcoxon rank-sum test), it appears to be a mere technical effect due to the high number of RNAs in both groups. Proteins encoded by PURA bound RNAs display a mean alteration in expression of 0.031, whereas non-associated proteins only show a mean alteration of 0.023 (not significant, P value $>$ 0.05, Wilcoxon rank-sum test). Similarly, the different binding regions have no impact on target regulation (**Figure 4.17**). PURA-bound RNAs and their respective proteins are dysregulated for all PURA-bound regions with an equal distribution of up and down-regulation. This results in all cases in a mean fold change close to zero.

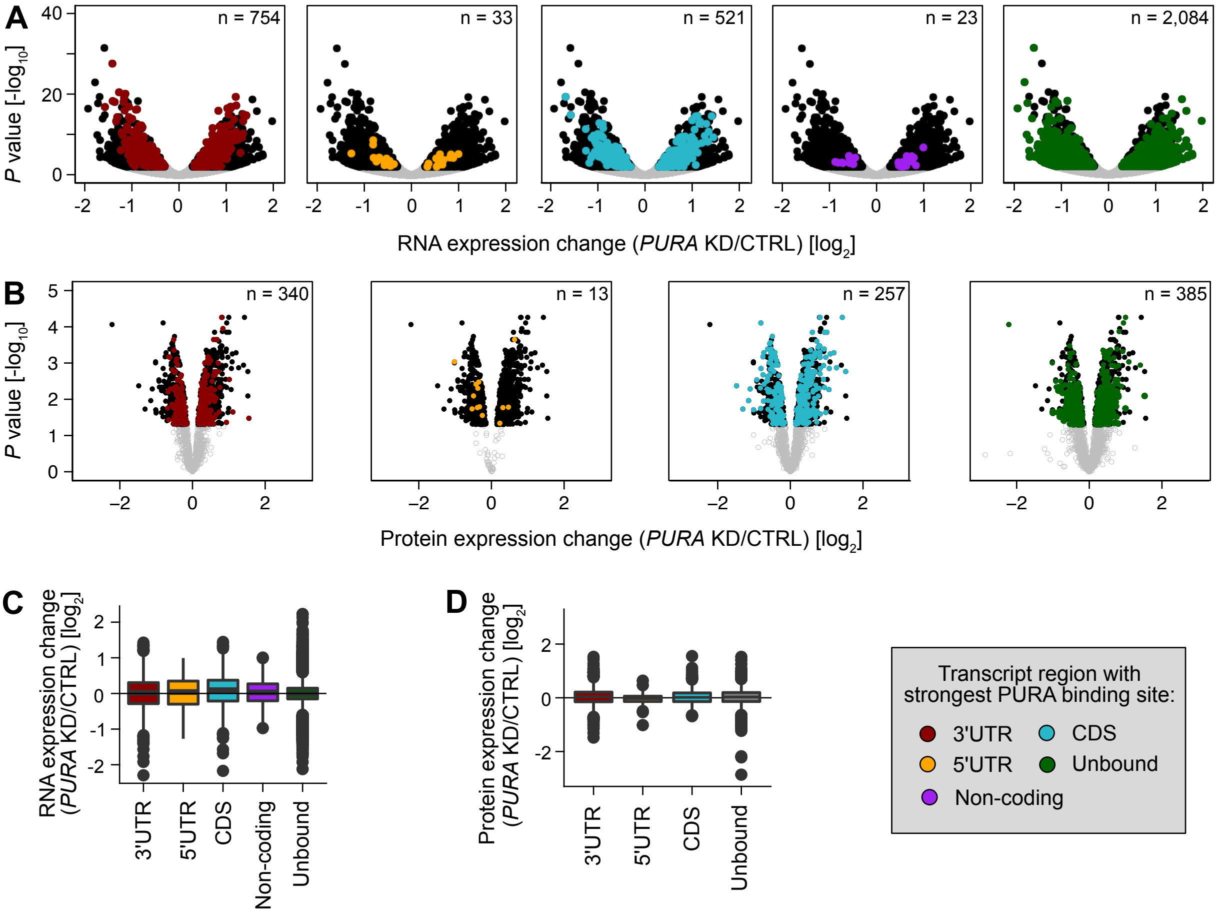


Figure 4.17: Directionality of RNA and protein expression changes upon *PURA* depletion is unconnected to the PURA-bound RNA region.
(A, B) Volcano plot of RNA (A) or protein (B) expression changes upon *PURA* knockdown, coloured by the PURA-bound region. Shown is the expression change between *PURA* knockdown and control on the x-axis ($\log_2$ scale) and the P value ($-\log_{10}$ scale, adjusted using Benjamini-Hochberg method) on the y-axis. Significant targets (RNAs - P value < 0.01 or proteins - P value < 0.05) with the respective binding are coloured. Red - bound in 3'UTR, orange - bound in 5'UTR, blue - bound in CDS, purple - bound non-coding RNAs, black - all other significant targets; grey - non-significant targets.
(C, D) Boxplots depict expression changes of RNAs (C) and proteins (D) for the different PURA-bound regions ($\log_2$ scale). Red - bound in 3'UTR, orange - bound in 5'UTR, blue - bound in CDS, purple - bound non-coding RNAs, grey - not PURA-bound RNAs.

The observation that PURA-bound RNAs are regulated in both directions speaks against the effect of PURA on RNA stability. For example, a knockdown of *UPF1* which is involved in mRNA decay leads mainly to positive RNA fold changes [Kishor *et al.*, 2020]. At the same time, translational effects should be observable in one-directional protein changes for PURA-bound RNAs. Therefore, the effect of *PURA* depletion on PURA-bound RNAs shows no clear evidence of PURA acting on RNA stability or RNA translation.

Direct PURA targets from the integration of PURA binding and regulated targets

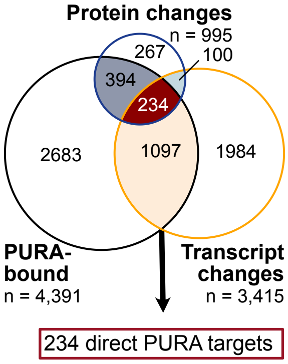


Figure 4.18: The overlap of PURA-bound and regulated targets can be used to define a set of 234 high-confidence direct PURA targets.
The overlap of PURA-bound RNAs (n = 4,391, black circle) with RNAs (n = 3,415, yellow circle) and proteins (n = 995, blue circle) dysregulated by *PURA* depletion is depicted as a Venn diagram. The 234 targets that overlap between all three groups are deemed direct PURA targets. 1,097 targets are bound by PURA and regulated only on the RNA level (yellow area), 394 are bound by PURA and regulated only on the protein level (blue area) and 100 targets change on RNA and protein levels but are not bound by PURA.

To generate a set of high-confidence direct PURA targets, I intersected targets found in all three data sets (**Figure 4.18**): PURA-bound transcripts, RNAs dysregulated in response to *PURA* depletion, and proteins dysregulated by *PURA* depletion. The largest overlap of two features (1,331 targets) consists of RNAs that are bound by PURA and change in their expression in *PURA* depletion. 394 targets are bound by PURA and change at the protein level but not the RNA level, whilst 100 targets undergo changes at both the protein and RNA levels but are not bound by PURA. I used the overlap of all three features to establish high-confidence direct PURA targets, resulting in 234 direct PURA targets which serve as starting points for more comprehensive investigations. Interesting direct targets include the autophagy receptor SQSTM1, the mitochondrial STARD7, the neurodevelopmental transcription factor CUX1 (Cut Like Homeobox 1), and the essential P-body factor LSM14A. A connection of PURA to SQSTM1 was previously described in a zebrafish model of C9ORF72 ALS [Swinnen *et al.*, 2020]. For STARD7, CUX1 and LSM14A, there was no previous description of a connection to PURA, therefore, they provide new links to pathways influenced by PURA. Downregulation of SQSTM1 and LSM14A was confirmed in orthogonal Western blots by our collaborators [Molitor *et al.*, 2023]. Taken together I define 234 direct PURA targets, whose functions could provide valuable insight into the cellular function of PURA.

Summary

RNAseq and proteomics experiments of *PURA* depletion display changed expression of 3,415 RNAs and 995 proteins. These changes are divided equally between up- and down-regulation and, therefore, show no clear evidence for a role of PURA in RNA stability or translation. The integration with PURA binding allows the identification of 234 direct PURA targets that aid further investigation of PURA-connected pathways.

4.3.3 The connection of PURA to neuronal contexts and cytoplasmic granules

To deduce the role of PURA in more physiological contexts from the previous results, I first used information from the public databases GeneOntology (GO) [Carbon *et al.*, 2018] and REACTOME [Fabregat *et al.*, 2017]. I looked for the enrichment of the PURA-bound or PURA-regulated targets in certain cellular compartments or pathways. From the various contexts thereby linked to PURA, I further analysed the connection of PURA to neuronal cells and its association with cellular granules. I explored the role of PURA binding in neurons, using a PURA iCLIP obtained from neuronal precursor cells (NPCs) and previously published dendritic RNAs [Middleton *et al.*, 2019]. For the investigation into PURA's connection to cytoplasmic granules, I integrated my findings for PURA with previously published transcriptomes of stress granules [Khong *et al.*, 2017] and P-bodies [Hubstenberger *et al.*, 2017].

Cellular pathways and components enriched in PURA targets

To get an overview of PURA-related pathways and cellular components, I performed enrichment analyses on all three PURA data sets. The 234 direct PURA targets were a too small set to use for enrichment analyses and did not enrich any pathway or component significantly (data not shown). Instead, I looked for terms enriched for PURA-bound targets or for targets with changing RNA levels in *PURA* depletion.

Enrichment analyses of PURA-bound RNAs in REACTOME pathways and GO cellular components revealed 477 and 214 terms significantly enriched, respectively (P value $\leq$ 0.001, test against hypergeometric distribution as implemented in HypeR). The 25 most enriched are shown here (**Figure 4.19**). For a full list of terms please refer to Supplementary Table 3 of [Molitor *et al.*, 2021]. The top 25 terms include among others terms related to the cell cycle, the immune system and viral infections, the mitochondria, the lifecycle

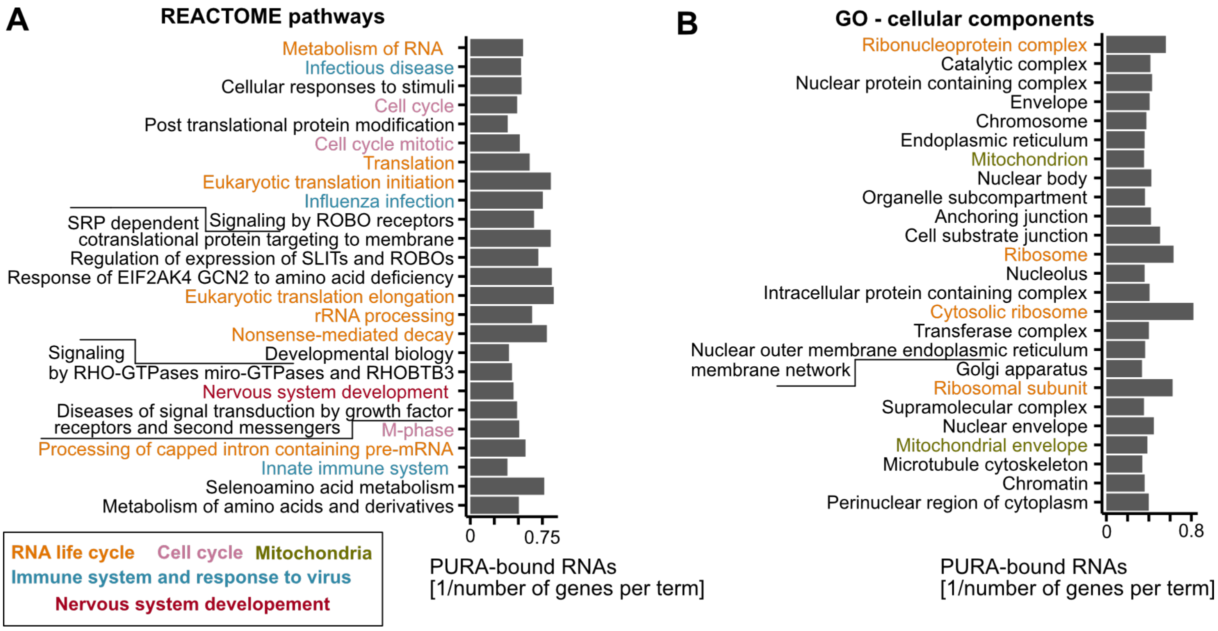


Figure 4.19: Top 25 enriched REACTOME pathways and cellular components from GeneOntology for PURA-bound RNAs.
Shown are the most enriched (lowest P value) REACTOME pathways **(A)** and cellular component terms from GeneOntology **(B)**. Bars depict the ratio of PURA-bound RNAs belonging to each term, divided by all expressed RNAs belonging to the term. P values of all shown terms are < 0.001.

of RNAs and the nervous system development. Similar terms can be found among the 20 most enriched terms for transcripts dysregulated in *PURA* depletion (**Figure 4.20**).

PURA has been implicated in many of these functions before: For instance, it was shown that PURA injection led to cell cycle arrest in the G2 phase [Stacey *et al.*, 1999, White *et al.*, 2009]. Also, PURA was found to be located in viral particles in cells infected with Human Immunodeficiency Virus 1 (HIV-1) [Garcia-Moreno *et al.*, 2023]. Additionally, the neuronal context of PURA is emphasised. In the next section, I further elucidate PURA functions in neuronal contexts.

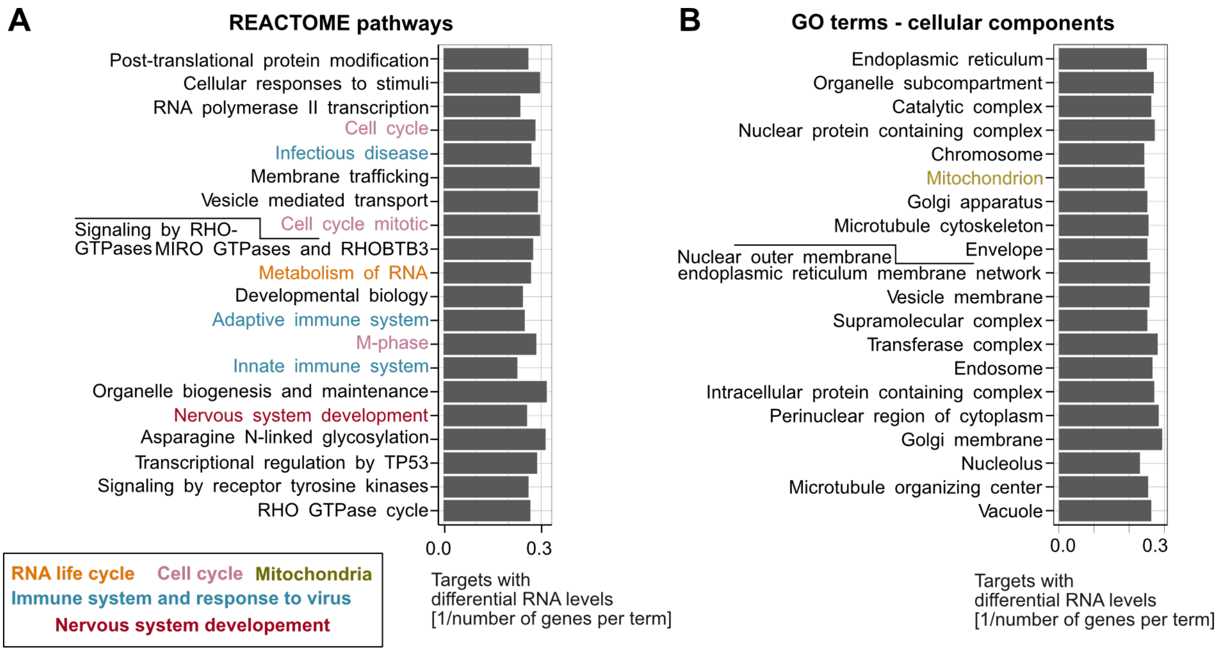


Figure 4.20: Top 20 enriched REACTOME pathways and cellular components from GeneOntology for PURA-bound RNAs.
Shown are the most enriched (lowest P value) REACTOME pathways **(A)** and cellular component terms from GeneOntology **(B)**. Bars depict the ratio of RNAs with significantly ($P < 0.01$) changing expression in *PURA* knockdown belonging to each term, divided by all expressed RNAs belonging to the term. P values of all shown terms are $<$ 0.001.

PURA RNA binding in the context of neuronal precursor cells and dendritic RNAs

As PURA plays a role in the neurodevelopmental PURA Syndrome, we wanted to investigate PURA RNA binding in a cellular context that is closer to a neurodevelopmental phenotype. For this, our collaborators differentiated induced pluripotent stem cells into neuronal precursor cells (NPCs) and then performed PURA iCLIP in the NPCs (NPC iCLIP). The NPC iCLIP data set exhibits limited signal depth and hence cannot be used for binding site definition (**Figure 4.21 A**). A reason for this might be that PURA is very lowly expressed in cell types of early neuronal development and is expressed stronger only in later stages as can be seen in the LIBD stem cell browser (http://stemcell.libd.org/scb/PURA/ENSG00000185129.5, accessed 12.12.2023) [Burke *et al.*, 2020]. Nevertheless, manual comparisons of regions with elevated PURA crosslinks in the NPC iCLIP exhibit comparable patterns of PURA crosslinks as in HeLa cells as can be seen in the examples of the *STARD7*, *SEC61A1* (*SEC61 Translocon Subunit Alpha 1*) and *YWHAE* (*Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Epsilon*) transcripts (**Figure 4.21 B, C, D**). In a metaprofile over all endogenous PURA binding sites from HeLa cells, the crosslinks in the NPC iCLIP peak in the centre of PURA HeLa binding sites (**Figure 4.21 E**). This hints at a similar RNA-binding behaviour of PURA in NPCs, although lowly expressed. Even though HeLa cells are physiologically very different from neuronal cells, RNAs bound by PURA in HeLa cells are enriched in 10 cellular compartments linked to neurons (**Figure 4.21 F**). Taken together, these observations suggest a similar behaviour of PURA in neuronal cells as in HeLa cells. Additional experiments should be performed in the future to confirm this finding.

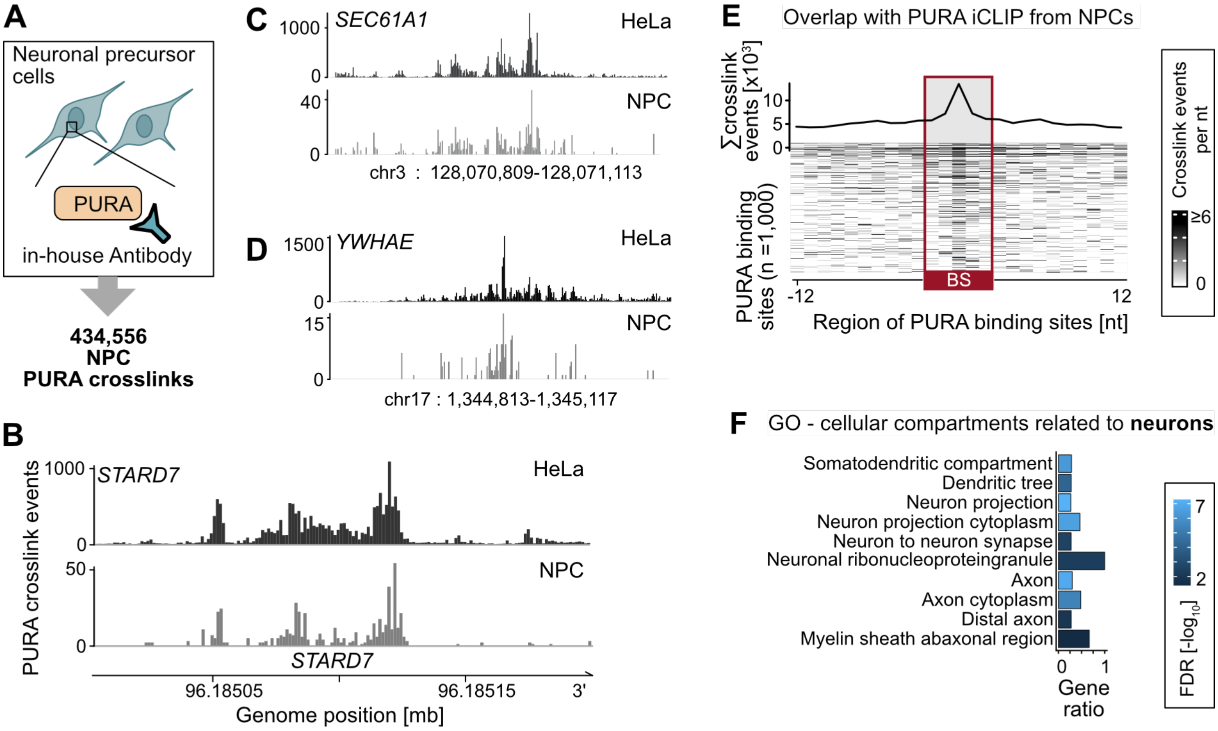


Figure 4.21: PURA RNA binding in neuronal precursor cells shows a comparable signal to PURA RNA binding in HeLa cells despite a very limited signal depth.
(A) Schema of NPC PURA iCLIP. Endogenous PURA was immunoprecipitated with the in-house anti-PURA antibody resulting in 434,556 PURA crosslinks.
(B, C, D) Exemplary crosslink profiles from PURA iCLIP in HeLa (top, black) and NPC cells (bottom, grey) in the 3'UTR of *STARD7* (B), *SEC61A1* (C) and *YWHAE* (D).
(E) PURA crosslinks from the PURA NPC iCLIP accumulate in the positions of PURA binding sites in HeLa cells. Shown are the relative positions from 12 nt upstream to 12 nt downstream of PURA binding sites. Top - metaprofile depicting the sum of PURA NPC crosslinks for each position; bottom - heatmap of PURA NPC crosslinks from the 1,000 regions with the highest amount of PURA crosslinks. The colour scale depicts the number of crosslinks from zero (white) to six or more (black) crosslinks. The binding site from PURA CLIP in Hela cells is highlighted as a red box.
(F) Enriched GeneOntology cellular compartments related to neurons for PURA-bound RNAs from the endogenous PURA iCLIP in HeLa cells. Bars depict the ratio of PURA-bound RNAs belonging to each term, divided by all expressed RNAs belonging to the term. *P* value is adjusted by Benjamini-Hochberg method and given by the colour scale.

PURA has been shown to play a role in dendritic transport mechanisms and specifically to transport the *Bc1* (Brain Cytoplasmic RNA 1), *CamkII* (calcium/calmodulin dependent protein kinase II) and *Acr* (acrosin) RNA [Ohashi *et al.*, 2002, Kanai *et al.*, 2004] in mouse neurons. To follow this up on a global level, I tested the overlap of PURA binding to RNAs that are located in the ends of dendrites in neurons. For this, I used a list of dendritic RNAs in mouse neurons that were previously published [Middleton *et al.*, 2019]. This publication performed a combinatorial analysis combining their own dendritic RNAseq with 5 other published dendritic RNAseq experiments from mouse neurons and listed the number of supporting studies for each dendritic RNA. I overlapped PURA-bound RNAs with these reported dendritic RNAs by transferring them to their mouse 1-2-1 orthologs and comparing only RNAs expressed in both dendrites and in HeLa cells (**Figure 4.22 A**). PURA binds 50.4% of all dendritic RNAs that were detected in at least four studies (**Figure 4.22 B**). Moreover, the fraction of bound dendritic RNAs increases with the number of supporting studies from 37.6% for one supporting study to 66.7% for five supporting studies (**Figure 4.22 C**). In summary, PURA does bind to a considerable amount of dendritic RNAs. It is, therefore, tempting to speculate that PURA transports some of its here-described bound RNAs along dendrites in a neuronal context.

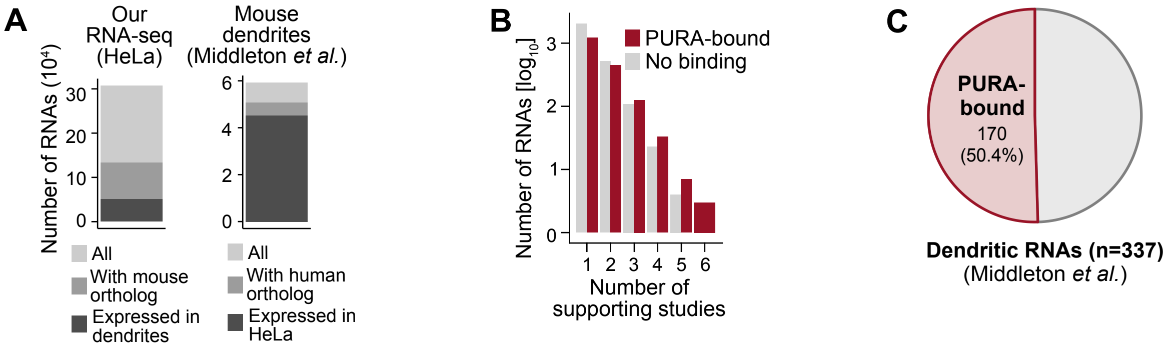


Figure 4.22: PURA binds to dendritic RNAs.
(A) To compare PURA-bound RNAs from HeLa cells and mouse dendritic RNAs from [Middleton *et al.*, 2019], RNAs with 1-2-1 orthologs between human and mouse that were expressed in both cell types were selected. Shown are the fractions of RNAs with orthologs (light grey) and expressed in both cell types (dark grey) for RNAs from our RNAseq (left) and the dendritic RNAs (right).
(B) The fraction of PURA-bound RNAs increases with supporting studies reporting an RNA to be dendritic. The number of PURA-bound (red) and not-bound (grey) dendritic RNAs is shown against the number of studies supporting the dendritic localisation of the RNA [Middleton *et al.*, 2019].
(C) PURA binds 50.4% of all dendritic RNAs when considering RNAs where the dendritic localisation was supported by at least 4 studies.

A connection of PURA to P-bodies

Given that according to our data, PURA seems to be neither an effector on transcript stability nor on translation and plays a role in RNA transport in neurons, I looked for a connection of PURA to RNA localisation in HeLa cells. In this context, I realised that the RNA of LSM14A, an essential P-body factor, is bound (**Figure 4.23 A**) and regulated by PURA (**Figure 4.14**). In addition, the RNA of DDX6, an interactor of LSM14A and similarly an essential P-body factor, is also bound by PURA (**Figure 4.23 B**), but is regulated only on protein not on RNA level (**Figure 4.14**). This led me to investigate PURA's connection to P-bodies and stress granules.

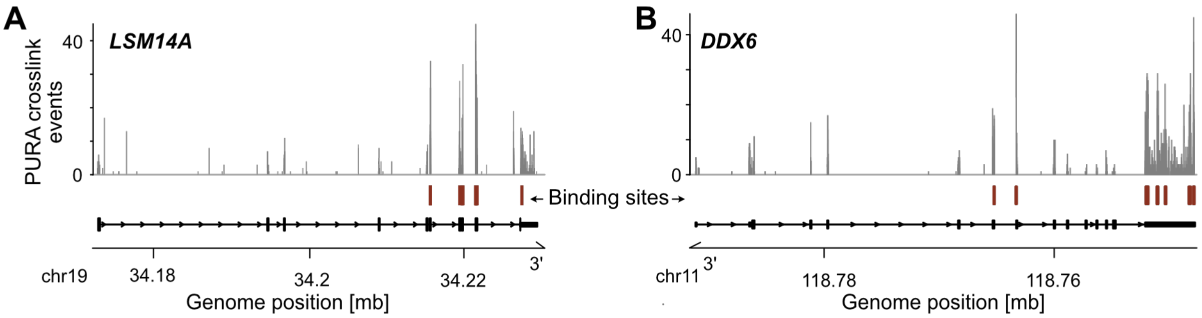


Figure 4.23: PURA binds two essential P-body factors.
Genome browser shots of the *LSM14A* **(A)** and *DDX6* **(B)** transcripts show PURA crosslinks (grey) and PURA binding sites (red boxes) together with the longest transcript annotation per gene (black schema).

To elucidate PURA's connection to RNAs stored in P-bodies and stress granules, I used a previously published P-body transcriptome [Hubstenberger *et al.*, 2017] and stress granule transcriptome [Khong *et al.*, 2017]. Both publications performed RNAseq of the RNAs in the respective granule compared to whole-cell RNA, allowing them to define RNA enrichment in the granule. I compared the overall enrichment of RNAs in either granule type between PURA-bound RNAs and an expression-matched control of unbound RNAs (**Figure 4.24, A, B, C, D**). PURA-bound RNAs have significantly higher enrichment in both granule types (P value < 0.0001, unpaired two-sided Wilcoxon rank-sum test).

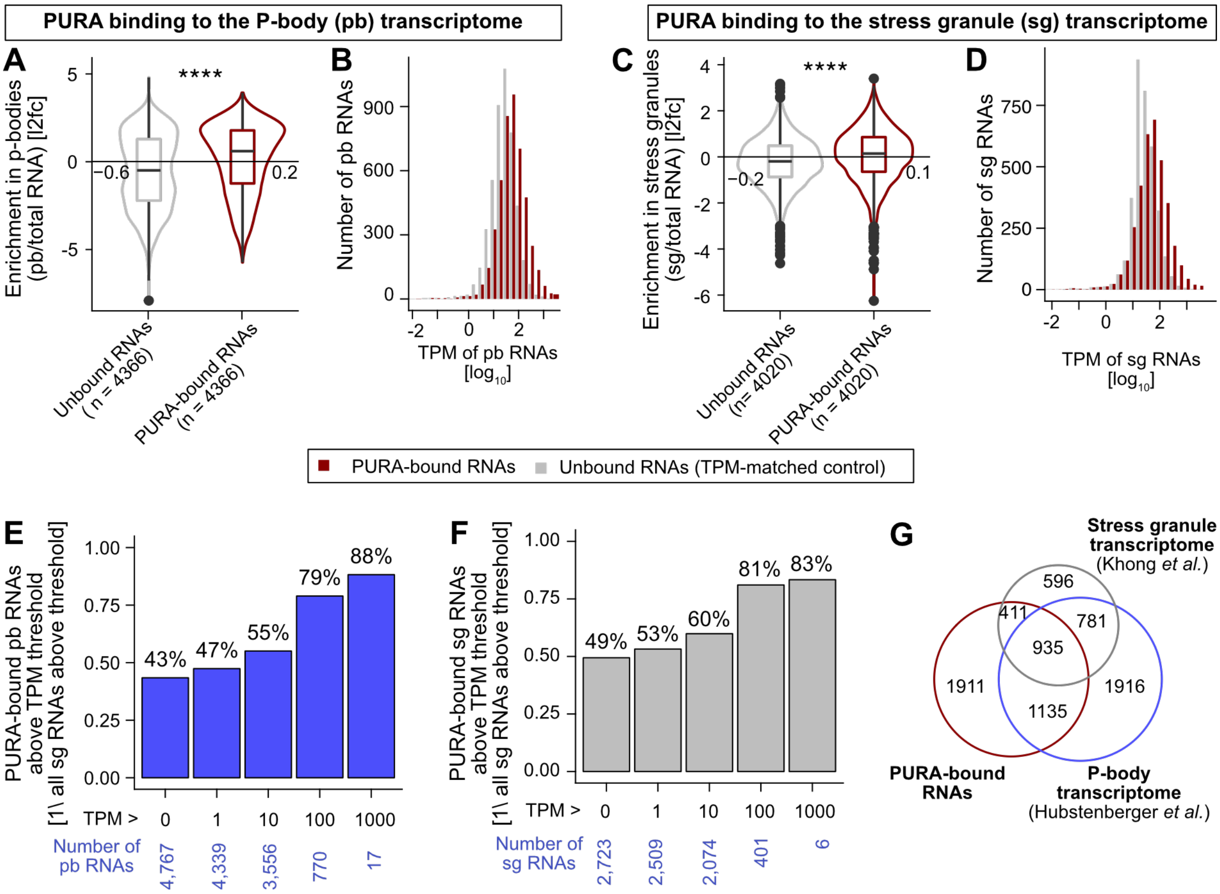


Figure 4.24: PURA-bound RNAs are enriched in the P-body and stress granule transcriptome.

(A, C) Violin and boxplot show enrichment of RNAs in P-bodies (A) and stress granules (B) for PURA-bound and a TPM-matched control of unbound RNAs. Only RNAs expressed in both HeLa cells as well as the HEK293 cells (n = 4,767, P-body transcriptome, [Hubstenberger *et al.*, 2017]) or the U2OS cells (stress granule transcriptome, n = 2,723, [Khong *et al.*, 2017]) are considered. **** - P value < 0.0001, unpaired two-sided Wilcoxon rank-sum test.

(B, D) Expression distribution of RNAs from the P-body transcriptome (C) or stress granule transcriptome (D) (red) and the respective expression-matched control (grey) is shown along TPM expression bins.

(E, F) Relative amount of PURA-bound RNAs in the P-body (E) and stress granule (F) transcriptome across RNA expression levels. Minimum RNA expression is given as TPM on the x-axis. The number of RNAs over the respective expression cutoff is given below (blue). Percentages of PURA-bound RNAs are displayed above bars.

(G) Overlap of PURA-bound RNAs (red) with the P-body (blue, n = 4,767, odds ratio 5.7, P value < 0.0001, Fisher‘s exact test) and stress granule transcriptome (grey, n = 2,723, odds ratio 12.0, P value < 0.0001, Fisher‘s exact test).

Then, I defined RNAs significantly enriched (P value < 0.01, $\log_2$ expression change > 0) in P-bodies or stress granules as the respective P-body or stress granule transcriptome. A comparison to PURA-bound RNAs shows that indeed 49% of the stress granule transcriptome and 43% of the P-body transcriptome are bound by PURA (P-body - P value < 0.0001, odds ratio 5.7, Fisher‘s exact test; stress granule - P value < 0.0001, odds ratio 12.0, Fisher‘s exact test; **Figure 4.24 A**). This enrichment further increases with the expression level of the granule RNAs (**Figure 4.24 E, F**). Taken together, this hints at a strong connection of PURA to RNAs located in P-bodies and stress granules.

Our collaborators could solidify the connection to P-bodies by showing that PURA localises in P-bodies in immunostainings of HeLa cells and normal human dermal fibroblasts (**Figure 4.25 A**) [Molitor *et al.*, 2023]. Even more, the number of P-bodies significantly decreased in *PURA*-depleted cells (**Figure 4.25 A, B**). In addition, I compared RNA regulation in *PURA* depletion between P-body and non-granule RNAs and found that *PURA* depletion - which can subsequently be connected to a loss of P-bodies - leads to an overall upregulation of P-body RNAs (**Figure 4.25 G**).

In summary, firstly, PURA binds P-body and stress granule RNAs, secondly, it is located in these granules [Molitor *et al.*, 2021, Tian *et al.*, 2022, Proske *et al.*, 2023] and thirdly, *PURA* depletion leads to a loss of P-bodies. Several possible mechanisms could be behind these observations as discussed later.

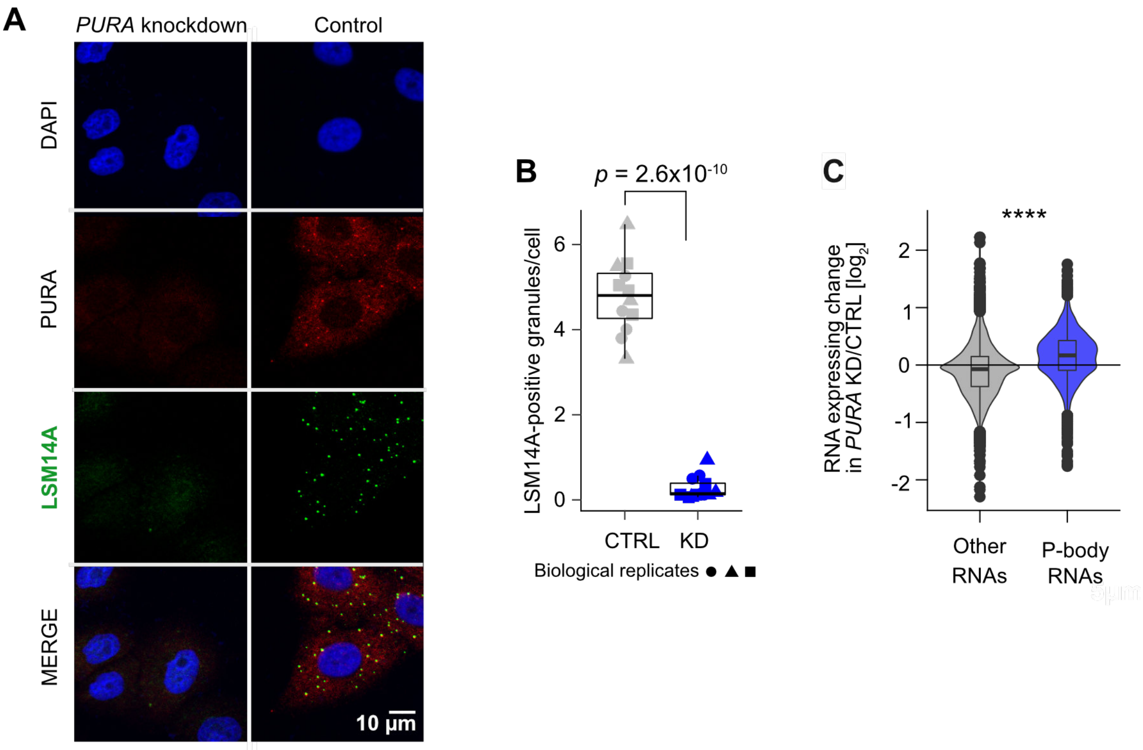


Figure 4.25: PURA localises to P-bodies and *PURA* depletion leads to a loss of P-bodies.
(A) Confocal micrographs of PURA (red), P-body marker LSM14A (green), nuclei (blue) and colocalised signal PURA and LSM14A signal (yellow). Top: control condition, Bottom: *PURA* knockdown condition, Scale bars - 10 µm. Figure done by Sabrina Bacher.
(B) Quantification (ImageJ) of LSM14A-positive granules on approximately 400 cells per replicate (n = 3) comparing control and PURA knockdown cells. Shown is the number of granules positive for LSM14A per cell for all replicates. 3 biological replicates (independent cultures) are indicated by different icons, with each containing 4 technical replicates (multiple images). *P* value from unpaired two-sided Student's t-test on all replicates. ImageJ analysis by Sabrina Bacher, plot and statistics by me.
(C) Violin and boxplots show RNA expression changes in *PURA* depletion against control for RNAs belonging to the P-body transcriptome (grey) and RNAs that neither overlapped with the P-body nor the stress granule transcriptome (blue).

Summary

PURA-bound RNAs and RNAs regulated by *PURA* depletion can link PURA to multiple cellular functions (**Figure 4.19, 4.20**). Some of them have been previously linked to PURA like its impact on the cell cycle [Stacey *et al.*, 1999, White *et al.*, 2009] and its interaction with viruses [Krachmarov *et al.*, 1996, Garcia-Moreno *et al.*, 2023] and can now more specifically be attributed to PURA's RNA-binding capacity. PURA RNA binding in neuronal precursor cells is comparable to PURA binding in HeLa cells (**Figure 4.21**) making it tempting to speculate that PURA's essential role in neurodevelopment can be attributed to PURA's RNA binding and PURA's neuronal transport function. Furthermore, PURA binds a considerable amount of dendritic RNAs (**Figure 4.22**), which might hint to PURA globally transporting RNAs along dendrites, as has been exemplarily shown for the RNAs *Bc1*, *CamkII* and *Acr* in mouse neurons [Ohashi *et al.*, 2002, Kanai *et al.*, 2004]. Of note, a new area of PURA function is cytoplasmic granules and especially P-bodies. PURA-bound RNAs are enriched in the P-body and stress granule transcriptome (**Figure 4.24**). Moreover, PURA localises to P-bodies and *PURA* depletion leads to a loss of P-bodies (**Figure 4.24**).

4.4 Discussion

4.4.1 PURA as an RNA-binding protein

Here, I show that endogenous PURA reproducibly binds 4,391 RNAs in over 50,000 binding sites with a preference for protein-coding RNAs (**Chapter 4.3.1**). This reveals a new role of PURA as an RNA-binding protein, which is consistently observed among three PURA iCLIP experiments in HeLa cells. Moreover, two concurrent studies identified PURA RNA-binding in KYSE 510 [Tian *et al.*, 2022] and Jurkat cells [Garcia-Moreno *et al.*, 2023], solidifying PURA's role in RNA binding. This opens a new perspective on the protein PURA, which was previously mostly described as a transcription factor.

PURA binding in 3'UTR and CDS

In HeLa cells, I found that PURA preferentially binds both 3'UTRs and CDS with binding sites evenly distributed between CDS and 3'UTR regions. Interestingly, in KYSE 510 cells 63% of PURA binding sites are located in the 3'UTR and only 17% in the CDS cells [Tian *et al.*, 2022], while in Jurkat cells over 50% of PURA binding sites are located in the CDS and roughly one third in the 3'UTR (exact numbers are not given in the preprint manuscript) [Garcia-Moreno *et al.*, 2023]. The variance in the relative amounts of PURA binding sites in one or the other region is likely due to the different cell types, the signal depth of the experiments and the different quantification approaches. However, 3'UTR and CDS always remain the two most bound regions.

The binding of PURA in 3'UTR and CDS regions displays notable differences that may suggest distinct modes of PURA binding and potentially may be connected to different cellular functions or a single function that combines binding in both regions. In two back-to-back studies [Van Nostrand *et al.*, 2020b, Van Nostrand *et al.*, 2020a] eCLIP experiments were conducted for 150 RBPs. Of note, the study connected 3'UTR-binding proteins to functions such as translational regulation, 3'end processing, mitochondrial regulation, as well as P-bodies and stress granules. RBPs primarily binding in the CDS were also associated with P-bodies and stress granules, ribosomes, basic translation, and splicing. Furthermore, they identified a group of RBPs that bind both CDS and 3'UTR, including translation repressors such as TIA1 (T-Cell-Restricted Intracellular Antigen 1), TIAL (T-Cell-Restricted Intracellular Antigen like), PUM2 (Pumilio homolog 2), and the RNA decay factor UPF1 (Up-Frameshift Suppressor 1 Homolog), as well as DDX6, which plays a role in both translation and mRNA decay and is furthermore connected to cytoplasmic granules. This demonstrates that RBP binding in multiple transcript regions is not an exclusive trait of PURA and that some RPB functions, such as RNA stability regulation and translation, may also involve binding in both regions. Moreover, it might hint at PURA functioning in similar contexts as the other RBPs from this group.

Specificity of PURA binding

I found that PURA binds a degenerate purine-rich motif (**Chapter 4.3.1**). Similar sequence motifs were obtained in *in vitro* PURA binding studies [Graebsch *et al.*, 2009, Weber *et al.*, 2016] and in the other two PURA iCLIP studies [Tian *et al.*, 2022, Garcia-Moreno *et al.*, 2023]. In contrast to PURA's described *in vitro* binding, the three CLIP studies all additionally show an enrichment of uridines in PURA binding sites. This is likely introduced by the UV crosslinking bias to uridines [Sugimoto *et al.*, 2012, Krakau *et al.*, 2017]. A possibility to verify this would be the use of orthogonal methods to query PURA binding to RNAs like RNA Bind-n-Seq [Lambert *et al.*, 2014] or SELEX for RNA-binding [Bouvet, 2001, Manley, 2013]. In RNA Bind-n-Seq, a pool of RNAs are *in vitro* subjected to the RBP of interest, then the RBP is pulled down and the RNAs are sequenced. The sequences of the bound RNAs can then be queried for the enrichment of specific k-mers, which were preferentially bound by the RBP. Systematic Evolution of Ligands by Exponential Enrichment (SELEX) selects RBP bound RNAs from a randomised RNA pool and this is repeated multiple times to increase the enrichment of RNAs specific to the RBP of interest. As both techniques do not depend on UV crosslinking, they will not display the same uridine bias.

The degeneration of the motif might question the specificity of PURA target recognition. Indeed, PURA RNA binding can be found on a high number of RNAs and is also broadly spread across RNAs, especially in CDS regions. It should be noted that RBPs recognise relatively short sequences (3-7 nt) in comparison to DNA-binding proteins [Lunde *et al.*, 2007, Jankowsky and Harris, 2015, Feng *et al.*, 2019]. Several RBPs, including TIA1 and HNRNPC (Heterogeneous Nuclear Ribonucleoprotein C), have been reported to bind to degenerated motifs such as U-rich or AU-rich elements [König *et al.*, 2010, Wang *et al.*, 2010]. It has been proposed that the bound sequence motif contains limited information content for many RBPs and that other mechanisms uphold RBP target recognition in addition [Lunde *et al.*, 2007, Jankowsky and Harris, 2015, Feng *et al.*, 2019]. The degenerated motif of PURA should therefore not be interpreted as unspecific binding, but rather as a common RBP characteristic.

In some cases, RBPs may detect their targets through the recognition of a particular RNA secondary structure, such as Staufen [Ramos *et al.*, 2000]. I could not detect a strong structure specificity for PURA binding sites apart from a general preference of PURA to bind single-stranded regions (**Chapter 4.3.1**). It, therefore, seems unlikely that PURA recognises its targets based on secondary structures.

In other cases, such as hnRNPs and SR proteins, it has been suggested that RBPs achieve their specificity by being included in RNP complexes, where cofactors may determine binding specificity [Singh and Valcárcel, 2005]. It is currently unclear whether PURA participates in an RNP complex with other RPBs. A useful resource for RNA-dependent protein-protein interactions is the R-DeeP database that performed density gradient ultracentrifugation with or without RNase treatment followed by mass spectrometry [Caudron-Herger *et al.*, 2019, Caudron-Herger *et al.*, 2020]. The results of this study allow linking proteins to other proteins in the same fraction. However, PURA was found to precipitate in this experiment and an identification of PURA co-factors from this study is not possible. Therefore, the interaction of PURA with possible co-factors requires further investigation. Furthermore, a specific location in the cell can determine which RNAs are bound by an RBP. In this sense, PURA location in P-bodies and stress granules can reduce the number of potential PURA RNA targets. Taken together, while some binding specificity of PURA arises from a degenerated purine-rich motif in a single-stranded RNA region, further factors might influence PURA target recognition such as co-factors and the location of PURA.

PURA as a transcription factor

Historically, PURA was classified as a transcription factor [Bergemann *et al.*, 1992], but when carefully going through literature most evidence for PURA's role in DNA binding comes either from *in vitro* binding assays or from RNA expression changes that could also be caused by PURA binding to the RNAs instead of the genes. PURA binds both DNA and RNA *in vitro* [Graebsch *et al.*, 2009, Weber *et al.*, 2016], but this does not necessarily mean that it will bind both *in vivo*. One could therefore hypothesise, that classification of PURA as transcription factor was a misinterpretation of its RBP binding function. What speaks for this hypothesis is that PURA locates mostly in the cytoplasm as shown in [Molitor *et al.*, 2023], which would not allow PURA to come into contact and bind DNA. In contrast to this hypothesis stands a recent study that queried PURA DNA binding *in vivo* with a ChIP-Seq experiment and described PURA DNA binding to 1,389 genes in the context of Alzheimer‘s disease [Shi *et al.*, 2021]. Furthermore, [Song *et al.*, 2022] showed PURA binding to the promoters of five Alzheimer-related genes using ChIP-qPCR. It might therefore be possible, that PURA has both a cellular role as an RNA binding protein and as transcription factor. Indeed, a recent study showed that a broad range of transcription factors like SOX2 (SRY-box transcription factor 2), MYC

(V-Myc Avian Myelocytomatosis Viral Oncogene Homolog) and CTCF (CCCTC-Binding Factor) also bind RNA *in vivo* [Oksuz *et al.*, 2023]. It might therefore be interesting to consider disentangling the effects of PURA RNA and DNA binding in future studies.

Disentangling PURA and PURB functions

In this discussion, it also has to be taken into account that most PURA antibodies are not specific for PURA but also detect the PURA paralog PURB. Therefore, our iCLIP analysis used a specific in-house anti-PURA antibody that has no cross-reactivity to PURB. Other publications used tagged PURA proteins [Garcia-Moreno *et al.*, 2023], and, therefore, their results can be surely connected to PURA. But as I show in the comparison of the iCLIP data sets, an ectopic expression of a tagged PURA might also decrease the PURA binding specificity in these studies. Other studies might be similarly affected by this and might not properly discriminate the effects of PURA and PURB [Shi *et al.*, 2021, Song *et al.*, 2022, Tian *et al.*, 2022]. The question, which findings belong to PURA alone and which might also be attributed to PURB is especially interesting as little is known about the function of PURB. Some publications describe both PURA and PURB in the same complexes [Ji *et al.*, 2007, Pandey *et al.*, 2020, Garcia-Moreno *et al.*, 2023]. However, it is hard to tell whether PURB has a similar functionality to PURA [Johnson *et al.*, 2013]. I found that PURB expression is not changing upon *PURA* depletion, which suggests that there are no direct feedback mechanisms between the two proteins. Moreover, my study of endogenous PURA with a specific anti-PURA antibody further aids in disentangling unique and shared functions of the two proteins.

4.4.2 PURA in cytoplasmic granules

I could establish a connection of PURA to P-bodies as PURA binds the RNAs of the essential P-body factors DDX6 and LSM14A and in PURA depletion *LSM14A* RNA and both LSM14A and DDX6 proteins are downregulated. Indeed, our collaborators could show that PURA localises to P-bodies and that P-body numbers are strongly reduced upon *PURA* depletion. PURA likely binds P-body and stress granule RNAs because of its localisation in these granules. Also, the previously described enrichment in GO terms of PURA-bound RNAs and RNA regulated by *PURA* depletion connected to RNA degradation and translation might be due to the effect on P-bodies, which were shown to contain mRNAs of post-transcriptional mRNA regulation factors [Hubstenberger *et al.*, 2017]. While the P-body factors LSM14A and DDX6 are downregulated in PURA depletion, I observe a global upward shift in the expression of the P-body transcriptome, clearly indicating that P-body function is impaired.

What role PURA plays in P-bodies can only be speculated at this point. One possibility would be that it directly influences the architecture of the P-bodies and that LSM14A and DDX6 are lost as a secondary effect due to the loss of P-bodies. Inversely, it could also be that P-bodies are lost due to *PURA* depletion downregulating the RNAs and consequently also the proteins of the essential P-body factors DDX6 and LSM14A. To further evaluate these two options, one could try to rescue P-body assembly in PURA-depleted cells by overexpression of DDX6 and LSM14A. Furthermore, while P-bodies are lost due to *PURA* depletion in HeLa cells and normal human dermal fibroblasts [Molitor *et al.*, 2021], this is not reproducible in some other cell lines (Sabrina Bacher, personal conversation). If the mechanism of P-body loss due to *PURA* depletion would be DDX6/LSM14A-dependent, an explanation for this might be that *PURA* depletion possibly does not affect DDX6 and LSM14A expression in other cell lines. This hypothesis can be tested by quantifying LSM14A and/or DDX6 abundances in these cell lines or also by testing whether PURA binds *LSM14A* and *DDX6* RNAs. Taken together, our study clearly establishes a link between PURA and P-bodies in HeLa cells and normal human dermal fibroblasts. This is evident from my analysis on the connection of PURA and the P-body transcriptome, and confirmed by targeted experiments by our collaborators. It will take further studies to understand the mechanism of how PURA influences P-body assembly.

At the same time as our study, two other studies described PURA's localisation in stress granules, a different type of cytoplasmic granules [Tian *et al.*, 2022, Proske *et al.*, 2023]. Interestingly, stress granules were not observed to be lost in *PURA* depletion, but PURA protein variants with mutations that are found in patients with PURA Syndrome did localise significantly less to stress granules than wildtype PURA [Proske *et al.*, 2023]. This speaks against a direct effect of PURA on granule architecture. Concerning PURA's role in stress granules, it was proposed that PURA interacts with PABPC1 (Poly(A) Binding Protein Cytoplasmic 1) in binding RNAs in their 3'UTR and then recruits translation initiation factors sequestering them away from the translation start site and thereby weakens mRNA translation [Tian *et al.*, 2022]. The formation of an RNP consisting of PURA, PABPC1 and their joint RNA targets would explain the localisation of PURA to stress granules. However, in our study, we could not observe global effects of *PURA* depletion on translation (**Chapter 4.3.2**). This should be further investigated possibly with a ribosome footprinting experiment [Ingolia *et al.*, 2009].

It might be tempting to speculate that PURA's location in P-bodies is dependent on the formation of an RNP with a co-factor, similar to the RNP formation of PURA and PABPC1 in stress granules. This cofactor might be LSM14A, which is found together with PURA and PURB in viral particles of HIV-1 [Garcia-Moreno *et al.*, 2023]. These speculations underline that the search for PURA co-factors is an intriguing next step to understanding PURA's cellular function and its role in cellular granules.

4.4.3 PURA's role in neurodevelopment and protection from neurodegeneration

Understanding PURA's cellular function builds a basis to better understand PURA's role in the disease contexts such as PURA Syndrome. While PURA has also been implicated in cancer progression [Tian *et al.*, 2022] and HIV-1 infection [Garcia-Moreno *et al.*, 2023], here, I am focusing on PURA's connection to neuronal developmental and neurodegenerative diseases. PURA has been shown to be present in RNA transport complexes that move along microtubules in neurons [Ohashi *et al.*, 2002] and to be essential for postnatal brain development in mice [Khalili *et al.*, 2003]. This might also explain the strong neurodevelopmental phenotype of mice lacking *PURA* and of PURA Syndrome patients, as a loss of PURA might lead to global dislocation of RNAs in neurons, which might - as previously shown [Kang and Schuman, 1996] - influence neuroplasticity.

However, while PURA is strongly connected to neuronal development and functionality, it is expressed in all cell types and loss of PURA in mice and PURA Syndrome patients influences the organism globally including also for example muscular dysfunctionalities, feeding difficulties and hypotonia [White *et al.*, 2009, Reijnders *et al.*, 2018, Molitor *et al.*, 2021]. It is, therefore, likely that PURA must act also in functions other than neuronal transport. Indeed, we found new cellular functions of PURA connected to P-bodies and mitochondria as described in the next section.

Insights into PURA-related neurodevelopmental disorders from PURA's cellular function

The 50% reduction of PURA in the RNAseq and proteomics experiments was chosen to mimic the situation in PURA Syndrome patients, where the heterozygous mutation is proposed to lead to a loss of half of the functional PURA protein also called haploinsufficiency [Reijnders *et al.*, 2018]. A limitation of this approach is that it cannot be excluded that some patient mutations might lead to a gain of function in PURA, meaning that PURA function might not be lost but be changed by the mutation. However, two patient mutations of PURA were shown to result in a protein that has decreased RNA affinity [Proske *et al.*, 2023], strengthening our approach to use the *PURA* knockdown. The described functions of PURA can also allow some insights into the 5q31.3 microdeletion Syndrome, where PURA is deleted together with two other genes. For 5q31.3 microdeletion Syndrome the 50% PURA reduction only displays a partial spectrum of the lost functions in the disorder as additional genes are missing in the microdeletion. Still a loss of PURA functionalities will be involved in the disease phenotype and my study of PURA's cellular function can, therefore, help elucidate the 5q31.3 microdeletion Syndrome.

In concordance with the neurodevelopmental delay in the two disorders, both targets bound by PURA and targets dysregulated in *PURA* depletion are enriched for factors from nervous system development and several neuronal locations (**Chapter 4.3.3**) as exemplified by *CUX1*. In addition, my results might provide a link between PURA Syndrome and mitochondrial functions, as exemplified by the downregulation of targets like *STARD7* in *PURA* depletion. Interestingly, several neurodevelopmental diseases including Fragile X Syndrome, Rett spectrum disorders and Angelman Syndrome have been associated with impaired mitochondrial respiration [Ortiz-González, 2021, Lin and Beal, 2006]. Moreover, the PURA Syndrome phenotypes of epilepsy, hypotonia and developmental delay have been connected to mitochondrial dysfunction in other contexts [Ortiz-González, 2021]. Taking this together, one might speculate that some of the PURA Syndrome phenotypes are connected to a dysfunction of mitochondria, that might be caused by the depletion of *PURA*. Indeed, mitochondrial function was described to be impaired by PURA loss in cell lines of oesophagal squamous cell carcinoma [Sun *et al.*, 2020].

In addition to the link to mitochondrial dysfunction described above, the loss of P-bodies in *PURA* depletion might be involved in the disease mechanics of PURA-related neurodevelopmental disorders as well. Of note, P-bodies are lost in DDX6 Syndrome, another rare disease which is caused by missense variants in the *DDX6* gene [Balak *et al.*, 2019]. The four clinically described patients with DDX6 Syndrome show phenotypes similar to PURA-related neurodevelopmental disorders, including intellectual disability, delay of speech, motor deficiency, feeding difficulties and facial dysmorphia. The study could show that DDX6 variants from patients disturb P-body assembly by reducing DDX6 binding to its interaction partners including LSM14A. The similarities of DDX6 Syndrome to PURA-related neurodevelopmental disorders and its clear connection to a loss of P-bodies might be indicative of the role of P-body loss in patients with PURA-related neurodevelopmental disorders.

PURA's neuroprotective function in neurodegenerative diseases

In addition to PURA's effect on neuronal development, PURA was implicated in several neurodegenerative diseases, in which PURA can fulfil a neuroprotective function. In both ALS/FTD and FXTAS, PURA overexpression could resolve the formation of neuronal foci. It might be interesting to reconsider this finding with PURA's role in P-bodies and stress granules in mind. While PURA promotes the formation of the physiological P-bodies, it dissolves unphysiological RNA aggregates. It has been proposed that disease-related neuronal aggregates arise by over-solidification - also called metastability - of membranless cellular compartments. In this process the liquid-like granule nucleates into glass-like protein crystals due to the high concentration of certain proteins like FUS (Fused in Sarcoma RNA Binding Protein) and HNRNPA1 (Heterogeneous Nuclear Ribonucleoprotein A1) [Shin and Brangwynne, 2017, Alberti and Hyman, 2021]. Factors that can increase the viscosity of granules therefore play an important role to counteract metastability and balance granule dynamics and it might be tempting to speculate, that PURA is one of these factors.

So far, both neurodevelopmental and neurodegenerative diseases seem to profit from enhancing PURA concentrations and it might be tempting to develop treatments that enhance PURA levels. However, it might be crucial to evaluate the effect of unphysiologically high PURA expression. As I showed PURA overexpression leads to a reduced specificity of PURA binding which could result in off-target effects. Indeed, in addition to the previously described PURA Syndrome that is likely caused by haploinsufficiency of PURA, there has been a clinical case of a patient with PURA duplication (Dierk Niessing, personal communication) and several patients with a 5q31 duplication [Rosenfeld *et al.*, 2011], resulting in a phenotype similar to PURA Syndrome and 5q31 microdeletion Syndrome. It will therefore be important to carefully study the effects of PURA dosage on both neuronal functionality and P-body and stress granule dynamics.

4.4.4 Summary

This study establishes PURA as a global RNA-binding protein, revealing its diverse roles in cellular processes. The findings indicate that PURA interacts with a multitude of transcripts due to its preference for single-stranded purine-rich sequences. These results are also consistent with several published studies.

PURA depletion leads to global RNA and protein expression changes. While no clear evidence for a direct role in RNA stability or translation can be observed from these changes, the identification of 234 direct PURA targets opens up opportunities for further investigation into PURA-associated pathways.

Furthermore, this study provides evidence linking PURA to various cellular functions. The extensive binding of dendritic RNAs strengthens previous reports of a potential role in RNA transport along dendrites. As PURA RNA binding can be transferred to neuronal precursor cells, this might shed light on the dysfunctional neurodevelopment that accompanies heterozygous loss of PURA. Moreover, PURA's involvement with cytoplasmic granules, particularly P-bodies, and its localisation to these structures, underscores a novel area of PURA function.

These findings collectively shed light on the multifaceted roles of PURA in RNA binding and its impact on various cellular processes. Importantly, these multifaceted roles of PURA hold relevance in various disease contexts like the PURA-related neuronal disorders PURA Syndrome and 5q31.3 microdeletion Syndrome and RNA repeat expansion disorders like ALS/FTD and FXTAS.

Chapter 5

Conclusion and Outlook

Establishing PURA as a global RNA-binding protein and characterising its preferred binding sites on RNAs has the potential to significantly enhance our understanding of its function. PURA's RNA-binding function is likely associated with conditions related to PURA. To gain further insight into the cellular mechanisms underlying PURA Syndrome, we simulated heterozygous PURA loss by a 50% depletion and compiled a list of over 200 direct PURA targets with potential implications for the disease context. Intriguingly, we could show that PURA localises in P-bodies and influences the formation of P-bodies in a likely DDX6/LSM14A-dependent manner.

These finding will need to be confirmed in a more physiological cell line and ideally with the usage of PURA patient mutations in contrast to the employed PURA depletion. An outstanding question still remains regarding the extent of the functional overlap of PURA with PURB, its paralog. Furthermore, it would be of interest to investigate which cofactors collaborate with PURA in an RNP, potentially influencing its binding specificity and function.

Our investigation into PURA RNA binding demonstrates the utility of CLIP experiments and their computational analysis. The software racoon_clip streamlines and enhances this analysis. I am certain that it will aid research in gaining a better comprehension of various RNA-binding proteins. An essential benefit of racoon_clip is its ability to process multiple CLIP data types, including those from iCLIP and eCLIP experiments, increasing the comparability between these two data sets. Multiple data sets can be integrated in large-scale approaches, including data from several RBPs and various experiment setups, using racoon_clip.

A paramount objective in further developing racoon_clip is to integrate user feedback to optimize its usability. To facilitate installation, supplying a racoon_clip docker image and possibly a conda installation should be considered. In addition, as there are over 28 types of CLIP experiments, it might be interesting to include further new experiment types. Lastly, the work in this thesis sets the stage to use racoon_clip processing on RBPs that are possible PURA co-factors to further learn about their possible interactions. Taken together, the work presented in this thesis will help to understand PURA functions and its implications in disease which will ultimately aid in elucidating the disease mechanisms and developing possible treatments.

Appendix

CV/ scientific contributions

- **Name:** Melina Klostermann
- **Date of Birth:** 24-July-1994
- **Place of Birth:** Heidelberg, Germany
- **Education:** M. Sc. and B. Sc. Biomolecular Engineering, Technische Universität Darmstadt
- **Address:** Karlstraße 101, 64285 Darmstadt
- **Email:** melinaklostermann@googlemail.com

Peer-reviewed publications

First authorship

Depletion of the RNA-binding protein PURA triggers changes in posttranscriptional gene regulation and loss of P-bodies.
Molitor L*, Klostermann M*, Bacher S, Merl-Pham J, Spranger N, Burczyk S, Ketteler C, Rusha E, Tews D, Pertek A, Proske M, Busch A, Reschke S, Feederle R, Hauck SM, Blum H, Drukker M, Fischer-Posovszky P, König J, Zarnack K$, Niessing D$ (2023); Nucleic Acids Research, 51(3):12971316. DOI http://dx.doi.org/10.1093/nar/gkac1237. (* *both authors contributed equally;* $ *shared correspondence*)

racoon_clip – a complete pipeline for single-nucleotide analyses of iCLIP and eCLIP data. Klostermann M and Zarnack K; *Manuscript is currently in review at Bioinformatics.*

A high-resolution map of functional miR-181 response elements in the thymus reveals the role of coding sequence targeting and an alternative seed match. Verheyden N*, Klostermann M*, Brüggemann M, Scholz A, Amr S, Lichtenthaeler C, Münch C, Schmid T, Zarnack K$^{\$}$, Krueger A$^{\$}$; (** both authors contributed equally;* $^{\$}$ *shared correspondence.*) *Manuscript is currently under revision at Nucleic Acids Research.* A preprint can be found at BioRxiv DOI https://doi.org/10.1101/2023.09.08.556730.

Co-authorships

Modulation of pre-mRNA structure by hnRNP proteins regulates alternative splicing of *MALT1*. Jones AN, Graß , Meininger I, Geerlof A, Klostermann M, Zarnack K, Krappmann D, Sattler M (2022); Science Advances, p9153. DOI http://doi/10.1126/sciadv.abp9153.

Heterogenous nuclear ribonucleoprotein D-like controls endothelial cell functions. Fischer S, Lichtenthaeler C, Stepanenko A, Heyl F, Maticzka D, Kemmerer K, Klostermann M, Backofen R, Zarnack K, Weigand J (2023); Biological Chemistry, DOI https://doi.org/10.1515/hsz-2023-0254.

Conferences and courses

Talks

How does PURA recognize RNA? PURA Syndrome Conference, 23.-25.06.2023 in Hinxton, UK. *Attendance was kindly funded by the GRADE IQbio graduate school.*

Depletion of the RNA-binding protein PURA triggers changes in posttranscriptional generegulation and loss of P-bodies. RNA Salon, 25.04.2023 in Frankfurt.

Using multi-omics to understand the neurodevelopmental PURA syndrome - challenges and insights. MPC-Cafe, 01.03.2023 at Metabolomics and Proteomics Core Helmholtz Zentrum München.

Depletion of the RNA-binding protein PURA triggers changes in posttranscriptional gene regulation and loss of P-bodies. SCALE Minisymposium, 11.11.2022 in Frankfurt.

A systemics view of PURAs cellular RNA-binding function from omics-data gives insights into PURA related diseases. Computational Approaches to RNA

Structure and Function, 07.08.-20.08.2022 in Benasque, Spain.

A systemic view of PURA's global cellular function from "omics" data. BMLS Seminar, 04.11.2021 in Frankfurt.

Poster presentations

A complete automated pipeline pinpoints miRNA-binding from miR-eCLIP data to narrow sites and provides global insights into miRNA binding propensities *in vivo*. International course on Non-coding Genome, 10.05.-17.05.2023 at Institut Curie in Paris, France. *Attendance was kindly funded by the Institut Curie.*

A systemics view of PURA's cellular function from omics data gives insights into PURA-related diseases. RNA Transport Meeting, 11.09. -14.09.2022 in Düsseldorf. *Attendance was kindly funded by the GRADE IQbio graduate school.*

A systemics view of PURA's cellular function from omics data gives insights into PURA-related diseases. Conference on Intelligent Systems for Molecular Biology (ISMB), 10.07.-14.07.2022, online.

A systemics view of PURA's cellular function from omics data gives insights into PURA-related diseases. Advances in Precision Medicine Conference, 29.11. - 01.12.2021 in Qatar. *Attendance was kindly funded by the Hamad Bin Khalifa University, Qatar.*

Teaching

WiSe 2023 Spezialisierung I & II "Spezialisierunspraktikum: RNA Bioinformatik" Modul 19H & 20E FB 15 Goethe Universität. Teaching a practical course in splicing analysis.

Bioinformatics Summer School - Computational RNA Biology of the National Center of Competence in Research, RNA & Disease and the Swiss Institute, 23.08. – 28.08.2020 in Schwarzenberg, Switzerland. Teaching assistant for iCLIP analyses.

List of Figures

List of Tables

Bibliography

[Afgan *et al.*, 2022] Afgan E, Nekrutenko A, Grüning B A, Blankenberg D, Goecks J, Schatz M C, Ostrovsky A E, Mahmoud A, Lonie A J, Syme A, Fouilloux A, Bretaudeau A, Nekrutenko A, Kumar A, Eschenlauer A C, DeSanto A D, Guerler A, Serrano-Solano B, Batut B, Grüning B A, Langhorst B W, Carr B, Raubenolt B A, Hyde C J, Bromhead C J, *et al.* (2022); The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, 50(W1):345–351, URL http://dx.doi.org/10.1093/nar/gkac247.

[Alberti and Hyman, 2021] Alberti S, Hyman A A (2021); Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. *Nature Reviews Molecular Cell Biology*, 22(3):196–213, URL http://dx.doi.org/10.1038/s41580-020-00326-6.

[Alberts, 1998] Alberts B (1998); The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists. *Cell*, 92(3):291–294, URL http://dx.doi.org/10.1016/s0092-8674(00)80922-8.

[Anders and Huber, 2010] Anders S, Huber W (2010); Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, URL http://dx.doi.org/10.1186/gb-2010-11-10-r106.

[Anders *et al.*, 2014] Anders S, Pyl P T, Huber W (2014); HTSeq - A Python framework to work with high-throughput sequencing data. *bioRxiv*, URL http://dx.doi.org/10.1101/002824.

[Andrews, 2010] Andrews S (2010); FastQC A Quality Control tool for High Throughput Sequence Data. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[Antifeeva *et al.*, 2022] Antifeeva I A, Fonin A V, Fefilova A S, Stepanenko O V, Povarova O I, Silonov S A, Kuznetsova I M, Uversky V N, Turoverov K K (2022); Liquid–liquid phase separation as an organizing principle of intracellular space: overview of the evolution of the cell compartmentalization concept. *Cellular and Molecular Life Sciences*, 79(5):251, URL http://dx.doi.org/10.1007/s00018-022-04276-4.

[Arzumanian *et al.*, 2022] Arzumanian V A, Dolgalev G V, Kurbatov I Y, Kiseleva O I, Poverennaya E V (2022); Epitranscriptome: Review of Top 25 Most-Studied RNA Modifications. *International Journal of Molecular Sciences*, 23(22):13851, URL http://dx.doi.org/10.3390/ijms232213851.

[Ayache *et al.*, 2015] Ayache J, Bénard M, Ernoult-Lange M, Minshall N, Standart N, Kress M, Weil D (2015); P-body assembly requires DDX6 repression complexes rather than decay or Ataxin2/2L complexes. *Molecular Biology of the Cell*, 26(14):2579–2595, URL http://dx.doi.org/10.1091/mbc.e15-03-0136.

[Bailey *et al.*, 2009] Bailey T L, Boden M, Buske F A, Frith M, Grant C E, Clementi L, Ren J, Li W W, Noble W S (2009); MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37:W202–W208, URL http://dx.doi.org/10.1093/nar/gkp335.

[Balak *et al.*, 2019] Balak C, Benard M, Schaefer E, Iqbal S, Ramsey K, Ernoult-Lange M, Mattioli F, Llaci L, Geoffroy V, Courel M, Naymik M, Bachman K K, Pfundt R, Rump P, Beest J t, Wentzensen I M, Monaghan K G, McWalter K, Richholt R, Béchec A L, Jepsen W, Both M D, Belnap N, Boland A, Piras I S, *et al.* (2019); Rare De Novo Missense Variants in RNA Helicase DDX6 Cause Intellectual Disability and Dysmorphic Features and Lead to P-Body Defects and RNA Dysregulation. *The American Journal of Human Genetics*, 105(3):509–525, URL http://dx.doi.org/10.1016/j.ajhg.2019.07.010.

[Baltz *et al.*, 2012] Baltz A, Munschauer M, Schwanhäusser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M, Wyler E, Bonneau R, Selbach M, Dieterich C, Landthaler M (2012); The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Molecular Cell*, 46(5):674–690, URL http://dx.doi.org/10.1016/j.molcel.2012.05.021.

[Barbe *et al.*, 2016] Barbe M F, Krueger J J, Loomis R, Otte J, Gordon J (2016); Memory deficits, gait ataxia and neuronal loss in the hippocampus and cerebellum in mice that are heterozygous for Pur-alpha. *Neuroscience*, 337:177–190, URL http://dx.doi.org/10.1016/j.neuroscience.2016.09.018.

[Baruzzo *et al.*, 2017] Baruzzo G, Hayer K E, Kim E J, Camillo B D, FitzGerald G A, Grant G R (2017); Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods*, 14(2):135–139, URL http://dx.doi.org/10.1038/nmeth.4106.

[Bassell *et al.*, 1998] Bassell G J, Zhang H, Byrd A L, Femino A M, Singer R H, Taneja K L, Lifshitz L M, Herman I M, Kosik K S (1998); Sorting of β-Actin mRNA and Protein to Neurites and Growth Cones in Culture. *The Journal of Neuroscience*, 18(1):251–265, URL http://dx.doi.org/10.1523/jneurosci.18-01-00251.1998.

[Bentley *et al.*, 2008] Bentley D R, Balasubramanian S, Swerdlow H P, Smith G P, Milton J, Brown C G, Hall K P, Evers D J, Barnes C L, Bignell H R, Boutell J M, Bryant J, Carter R J, Cheetham R K, Cox A J, Ellis D J, Flatbush M R, Gormley N A, Humphray S J, Irving L J, Karbelashvili M S, Kirk S M, Li H, Liu X, Maisinger K S, *et al.* (2008); Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, URL http://dx.doi.org/10.1038/nature07517.

[Bergemann *et al.*, 1992] Bergemann A D, Ma Z W, Johnson E M (1992); Sequence of cDNA comprising the human pur gene and sequence-specific single-stranded-DNA-binding properties of the encoded protein. *Molecular and Cellular Biology*, 12(12):5673–5682, URL http://dx.doi.org/10.1128/mcb.12.12.5673.

[Bernhart *et al.*, 2006] Bernhart S H, Hofacker I L, Stadler P F (2006); Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–615, URL http://dx.doi.org/10.1093/bioinformatics/btk014.

[Björk and Wieslander, 2017] Björk P, Wieslander L (2017); Integration of mRNP formation and export. *Cellular and Molecular Life Sciences*, 74(16):2875–2897, URL http://dx.doi.org/10.1007/s00018-017-2503-3.

[Boivin *et al.*, 2018] Boivin M, Willemsen R, Hukema R K, Sellier C (2018); Potential pathogenic mechanisms underlying Fragile X Tremor Ataxia Syndrome: RAN translation and/or RNA gain-of-function? *European Journal of Medical Genetics*, 61(11):674–679, URL http://dx.doi.org/10.1016/j.ejmg.2017.11.001.

[Bouvet, 2001] Bouvet P (2001); DNA-Protein Interactions, Principles and Protocols. *Methods in molecular biology (Clifton, N.J.)*, 148:603–610, URL http://dx.doi.org/10.1385/1-59259-208-2:603.

[Boycott *et al.*, 2017] Boycott K M, Rath A, Chong J X, Hartley T, Alkuraya F S, Baynam G, Brookes A J, Brudno M, Carracedo A, Dunnen J T d, Dyke S O, Estivill X, Goldblatt J, Gonthier C, Groft S C, Gut I, Hamosh A, Hieter P, Höhn S, Hurles M E, Kaufmann P, Knoppers B M, Krischer J P, Macek M, Matthijs G, *et al.* (2017); International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *The American Journal of Human Genetics*, 100(5):695–705, URL http://dx.doi.org/10.1016/j.ajhg.2017.04.003.

[Boyle *et al.*, 2023] Boyle E A, Her H L, Mueller J R, Naritomi J T, Nguyen G G, Yeo G W (2023); Skipper analysis of eCLIP datasets enables sensitive detection of constrained translation factor binding sites. *Cell Genomics*, 3(6):100317, URL http://dx.doi.org/10.1016/j.xgen.2023.100317.

[Brown *et al.*, 2013] Brown N, Burgess T, Forbes R, McGillivray G, Kornberg A, Mandelstam S, Stark Z (2013); 5q31.3 Microdeletion syndrome: Clinical and molecular characterization of two further cases. *American Journal of Medical Genetics Part A*, 161(10):2604–2608, URL http://dx.doi.org/10.1002/ajmg.a.36108.

[Buchbender *et al.*, 2020] Buchbender A, Mutter H, Sutandy F X R, Körtel N, Hänel H, Busch A, Ebersberger S, König J (2020); Improved library preparation with the new iCLIP2 protocol. *Methods*, 178:33–48, URL http://dx.doi.org/10.1016/j.ymeth.2019.10.003.

[Burke *et al.*, 2020] Burke E E, Chenoweth J G, Shin J H, Collado-Torres L, Kim S K, Micali N, Wang Y, Colantuoni C, Straub R E, Hoeppner D J, Chen H Y, Sellers A, Shibbani K, Hamersky G R, Bustamante M D, Phan B N, Ulrich W S, Valencia C, Jaishankar A, Price A J, Rajpurohit A, Semick S A, Bürli R W, Barrow J C, Hiler D J, *et al.* (2020); Dissecting transcriptomic signatures of neuronal differentiation and maturation using iPSCs. *Nature Communications*, 11(1):462, URL http://dx.doi.org/10.1038/s41467-019-14266-z.

[Busch *et al.*, 2020] Busch A, Brüggemann M, Ebersberger S, Zarnack K (2020); iCLIP data analysis: A complete pipeline from sequencing reads to RBP binding sites. *Methods*, 178:49–62, URL http://dx.doi.org/10.1016/j.ymeth.2019.11.008.

[Carbon *et al.*, 2018] Carbon S, Douglass E, Dunn N, Good B, Harris N L, Lewis S E, Mungall C J, Basu S, Chisholm R L, Dodson R J, Hartline E, Fey P, Thomas P D, Albou L P, Ebert D, Kesling M J, Mi H, Muruganujan A, Huang X, Poudel S, Mushayahama T, Hu J C, LaBonte S A, Siegele D A, Antonazzo G, *et al.* (2018); The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, URL http://dx.doi.org/10.1093/nar/gky1055.

[Castello *et al.*, 2012] Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann B, Strein C, Davey N, Humphreys D, Preiss T, Steinmetz L, Krijgsveld J, Hentze M (2012); Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell*, 149(6):1393–1406, URL http://dx.doi.org/10.1016/j.cell.2012.04.031.

[Caudron-Herger *et al.*, 2019] Caudron-Herger M, Rusin S F, Adamo M E, Seiler J, Schmid V K, Barreau E, Kettenbach A N, Diederichs S (2019); R-DeeP: Proteome-wide and Quantitative Identification of RNA-Dependent Proteins by Density Gradient Ultracentrifugation. *Molecular Cell*, 75(1):184–199.e10, URL http://dx.doi.org/10.1016/j.molcel.2019.04.018.

[Caudron-Herger *et al.*, 2020] Caudron-Herger M, Wassmer E, Nasa I, Schultz A S, Seiler J, Kettenbach A N, Diederichs S (2020); Identification, quantification and bioinformatic analysis of RNA-dependent proteins by RNase treatment and density gradient ultracentrifugation using R-DeeP. *Nature Protocols*, 15(4):1338–1370, URL http://dx.doi.org/10.1038/s41596-019-0261-4.

[Chakrabarti *et al.*, 2018] Chakrabarti A M, Haberman N, Praznik A, Luscombe N M, Ule J (2018); Data Science Issues in Studying Protein–RNA Interactions with CLIP Technologies. *Annual Review of*

Biomedical Data Science, 1(1):235–261, URL http://dx.doi.org/10.1146/annurev-biodatasci-080917-013525.

[Chen *et al.*, 2014] Chen B, Yun J, Kim M S, Mendell J T, Xie Y (2014); PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome Biology*, 15(1):R18, URL http://dx.doi.org/10.1186/gb-2014-15-1-r18.

[Chi *et al.*, 2009] Chi S W, Zang J B, Mele A, Darnell R B (2009); Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature*, 460(7254):479–486, URL http://dx.doi.org/10.1038/nature08170.

[Cock *et al.*, 2010] Cock P J A, Fields C J, Goto N, Heuer M L, Rice P M (2010); The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, URL http://dx.doi.org/10.1093/nar/gkp1137.

[Cooper *et al.*, 2009] Cooper T A, Wan L, Dreyfuss G (2009); RNA and Disease. *Cell*, 136(4):777–793, URL http://dx.doi.org/10.1016/j.cell.2009.02.011.

[Costa, 2005] Costa F F (2005); Non-coding RNAs: New players in eukaryotic biology. *Gene*, 357(2):83–94, URL http://dx.doi.org/10.1016/j.gene.2005.06.019.

[Craig and Banker, 1994] Craig A M, Banker G (1994); Neuronal Polarity. *Annual Review of Neuroscience*, 17(1):267–310, URL http://dx.doi.org/10.1146/annurev.ne.17.030194.001411.

[Crick, 1958] Crick F H (1958); On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–63.

[Dai *et al.*, 2023] Dai W, Sun Y, Fan Y, Gao Y, Zhan Y, Wang L, Xiao B, Qiu W, Gu X, Sun K, Yu Y, Xu N (2023); A 25 Mainland Chinese cohort of patients with PURA-related neurodevelopmental disorders: clinical delineation and genotype–phenotype correlations. *European Journal of Human Genetics*, 31(1):112–121, URL http://dx.doi.org/10.1038/s41431-022-01217-4.

[Daigle *et al.*, 2016] Daigle J G, Krishnamurthy K, Ramesh N, Casci I, Monaghan J, McAvoy K, Godfrey E W, Daniel D C, Johnson E M, Monahan Z, Shewmaker F, Pasinelli P, Pandey U B (2016); Pur-alpha regulates cytoplasmic stress granule dynamics and ameliorates FUS toxicity. *Acta Neuropathologica*, 131(4):605–620, URL http://dx.doi.org/10.1007/s00401-015-1530-0.

[Das *et al.*, 2021] Das S, Vera M, Gandin V, Singer R H, Tutucci E (2021); Intracellular mRNA transport and localized translation. *Nature Reviews Molecular Cell Biology*, 22(7):483–504, URL http://dx.doi.org/10.1038/s41580-021-00356-8.

[Decker and Parker, 2012] Decker C J, Parker R (2012); P-Bodies and Stress Granules: Possible Roles in the Control of Translation and mRNA Degradation. *Cold Spring Harbor Perspectives in Biology*, 4(9):a012286, URL http://dx.doi.org/10.1101/cshperspect.a012286.

[DeJesus-Hernandez *et al.*, 2011] DeJesus-Hernandez M, Mackenzie I, Boeve B, Boxer A, Baker M, Rutherford N, Nicholson A, Finch N, Flynn H, Adamson J, Kouri N, Wojtas A, Sengdy P, Hsiung G Y, Karydas A, Seeley W, Josephs K, Coppola G, Geschwind D, Wszolek Z, Feldman H, Knopman D, Petersen R, Miller B, Dickson D, *et al.* (2011); Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron*, 72(2):245–256, URL http://dx.doi.org/10.1016/j.neuron.2011.09.011.

[Delaunay *et al.*, 2023] Delaunay S, Helm M, Frye M (2023); RNA modifications in physiology and disease: towards clinical applications. *Nature Reviews Genetics*, URL http://dx.doi.org/10.1038/s41576-023-00645-2.

[Dobin *et al.*, 2013] Dobin A, Davis C A, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras T R (2013); STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, URL http://dx.doi.org/10.1093/bioinformatics/bts635.

[Dodt *et al.*, 2012] Dodt M, Roehr J T, Ahmed R, Dieterich C (2012); FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology*, 1(3):895–905, URL http://dx.doi.org/10.3390/biology1030895.

[Doma and Parker, 2007] Doma M K, Parker R (2007); RNA Quality Control in Eukaryotes. *Cell*, 131(4):660–668, URL http://dx.doi.org/10.1016/j.cell.2007.10.041.

[Donato *et al.*, 2021] Donato L, Scimone C, Rinaldi C, D'Angelo R, Sidoti A (2021); New evaluation methods of read mapping by 17 aligners on simulated and empirical NGS data: an updated comparison of DNA- and RNA-Seq data from Illumina and Ion Torrent technologies. *Neural Computing and Applications*, 33(22):15669–15692, URL http://dx.doi.org/10.1007/s00521-021-06188-z.

[Donnelly *et al.*, 2013] Donnelly C J, Zhang P W, Pham J T, Haeusler A R, Heusler A R, Mistry N A, Vidensky S, Daley E L, Poth E M, Hoover B, Fines D M, Maragakis N, Tienari P J, Petrucelli L, Traynor B J, Wang J, Rigo F, Bennett C F, Blackshaw S, Sattler R, Rothstein J D (2013); RNA Toxicity from the ALS/FTD C9ORF72 Expansion Is Mitigated by Antisense Intervention. *Neuron*, 80(2):415–428, URL http://dx.doi.org/10.1016/j.neuron.2013.10.015.

[Drewe-Boss *et al.*, 2018] Drewe-Boss P, Wessels H H, Ohler U (2018); omniCLIP: probabilistic identification of protein-RNA interactions from CLIP-seq data. *Genome Biology*, 19(1):183, URL http://dx.doi.org/10.1186/s13059-018-1521-2.

[Dreyfuss *et al.*, 2002] Dreyfuss G, Kim V N, Kataoka N (2002); Messenger-RNA-binding proteins and the messages they carry. *Nature Reviews Molecular Cell Biology*, 3(3):195–205, URL http://dx.doi.org/10.1038/nrm760.

[Eulalio *et al.*, 2007] Eulalio A, Behm-Ansmant I, Schweizer D, Izaurralde E (2007); P-Body Formation Is a Consequence, Not the Cause, of RNA-Mediated Gene Silencing. *Molecular and Cellular Biology*, 27(11):3970–3981, URL http://dx.doi.org/10.1128/mcb.00128-07.

[Ewels *et al.*, 2016] Ewels P, Magnusson M, Lundin S, Käller M (2016); MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, URL http://dx.doi.org/10.1093/bioinformatics/btw354.

[Fabregat *et al.*, 2017] Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V, D'Eustachio P, Stein L, Hermjakob H (2017); Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics*, 18(1):142, URL http://dx.doi.org/10.1186/s12859-017-1559-2.

[Feng *et al.*, 2019] Feng H, Bao S, Rahman M A, Weyn-Vanhentenryck S M, Khan A, Wong J, Shah A, Flynn E D, Krainer A R, Zhang C (2019); Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites. *Molecular Cell*, 74(6):1189–1204.e6, URL http://dx.doi.org/10.1016/j.molcel.2019.02.002.

[Frankish *et al.*, 2018] Frankish A, Diekhans M, Ferreira A M, Johnson R, Jungreis I, Loveland J, Mudge J M, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Sala S C, Chrast J, Cunningham F, Domenico T D, Donaldson S, Fiddes I T, Girón C G, Gonzalez J M, Grego T, Hardy M, Hourlier T, Hunt T, *et al.* (2018); GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, URL http://dx.doi.org/10.1093/nar/gky955.

[Garcia-Moreno *et al.*, 2023] Garcia-Moreno M, Truman R, Chen H, Iselin L, Lenz C E, Lee J, Dicker K, Noerenberg M, Sohier T J M, Palmalux N, Jaervelin A I, Kamel W, Ruscica V, Ricci E P, Davis I, Mohammed S, Castello A (2023); Incorporation of genome-bound cellular proteins into HIV-1 particles regulates viral infection. *bioRxiv*, URL http://dx.doi.org/10.1101/2023.06.14.544764.

[Gentleman *et al.*, 2004] Gentleman R C, Carey V J, Bates D M, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini A J, Sawitzki G, Smith C, Smyth G, Tierney L, Yang J Y, Zhang J (2004); Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, URL http://dx.doi.org/10.1186/gb-2004-5-10-r80.

[Gerstberger *et al.*, 2013] Gerstberger S, Hafner M, Tuschl T (2013); Learning the language of post-transcriptional gene regulation. *Genome Biology*, 14(8):130, URL http://dx.doi.org/10.1186/gb-

2013-14-8-130.

[Gerstberger *et al.*, 2014] Gerstberger S, Hafner M, Tuschl T (2014); A census of human RNA-binding proteins. *Nature Reviews Genetics*, 15(12):829–845, URL http://dx.doi.org/10.1038/nrg3813.

[Ghanbari and Ohler, 2020] Ghanbari M, Ohler U (2020); Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Research*, 30(2):214–226, URL http://dx.doi.org/10.1101/gr.247494.118.

[Goutman *et al.*, 2022] Goutman S A, Hardiman O, Al-Chalabi A, Chió A, Savelieff M G, Kiernan M C, Feldman E L (2022); Emerging insights into the complex genetics and pathophysiology of amyotrophic lateral sclerosis. *The Lancet Neurology*, 21(5):465–479, URL http://dx.doi.org/10.1016/s1474-4422(21)00414-2.

[Graebsch *et al.*, 2009] Graebsch A, Roche S, Niessing D (2009); X-ray structure of Pur-α reveals a Whirly-like fold and an unusual nucleic-acid binding surface. *Proceedings of the National Academy of Sciences*, 106(44):18521–18526, URL http://dx.doi.org/10.1073/pnas.0907990106.

[Graff-Radford and Woodruff, 2007] Graff-Radford N, Woodruff B (2007); Frontotemporal Dementia. *Semin Neurol*, 27(1):048–057, URL http://dx.doi.org/10.1055/s-2006-956755.

[Guzikowski *et al.*, 2019] Guzikowski A R, Chen Y S, Zid B M (2019); Stress-induced mRNP granules: Form and function of processing bodies and stress granules. *Wiley Interdisciplinary Reviews: RNA*, 10(3):e1524, URL http://dx.doi.org/10.1002/wrna.1524.

[Hafner *et al.*, 2010] Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jungkamp A C, Munschauer M, Ulrich A, Wardle G S, Dewell S, Zavolan M, Tuschl T (2010); Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–41, URL http://dx.doi.org/10.1016/j.cell.2010.03.009.

[Hafner *et al.*, 2021] Hafner M, Katsantoni M, Köster T, Marks J, Mukherjee J, Staiger D, Ule J, Zavolan M (2021); CLIP and complementary methods. *Nature Reviews Methods Primers*, 1(1):20, URL http://dx.doi.org/10.1038/s43586-021-00018-1.

[Hagerman and Hagerman, 2016] Hagerman R J, Hagerman P (2016); Fragile X-associated tremor/ataxia syndrome — features, mechanisms and management. *Nature Reviews Neurology*, 12(7):403–412, URL http://dx.doi.org/10.1038/nrneurol.2016.82.

[Hahne and Ivanek, 2016] Hahne F, Ivanek R (2016); Statistical Genomics, Methods and Protocols. *Methods in Molecular Biology*, 1418:335–351, URL http://dx.doi.org/10.1007/978-1-4939-3578-9_16.

[Hentze *et al.*, 2018] Hentze M W, Castello A, Schwarzl T, Preiss T (2018); A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology*, 19(5):327–341, URL http://dx.doi.org/10.1038/nrm.2017.130.

[Heyl and Backofen, 2021] Heyl F, Backofen R (2021); StoatyDive: Evaluation and classification of peak profiles for sequencing data. *GigaScience*, 10(6):giab045, URL http://dx.doi.org/10.1093/gigascience/giab045.

[Hirose *et al.*, 2014] Hirose T, Virnicchi G, Tanigawa A, Naganuma T, Li R, Kimura H, Yokoi T, Nakagawa S, Bénard M, Fox A H, Pierron G (2014); NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies. *Molecular Biology of the Cell*, 25(1):169–183, URL http://dx.doi.org/10.1091/mbc.e13-09-0558.

[Hirose *et al.*, 2023] Hirose T, Ninomiya K, Nakagawa S, Yamazaki T (2023); A guide to membraneless organelles and their various roles in gene regulation. *Nature Reviews Molecular Cell Biology*, 24(4):288–304, URL http://dx.doi.org/10.1038/s41580-022-00558-8.

[Hirtz *et al.*, 2007] Hirtz D, Thurman D J, Gwinn-Hardy K, Mohamed M, Chaudhuri A R, Zalutsky R (2007); How common are the "common" neurologic disorders? *Neurology*, 68(5):326–337, URL http://dx.doi.org/10.1212/01.wnl.0000252807.38124.a3.

[Hokkanen *et al.*, 2011] Hokkanen S, Feldmann H M, Ding H, Jung C K E, Bojarski L, Renner-Muller I, Schuller U, Kretzschmar H, Wolf E, Herms J (2011); Lack of Pur-alpha alters postnatal brain development and causes megalencephaly. *Human Molecular Genetics*, 21(3):473–484, URL http://dx.doi.org/10.1093/hmg/ddr476.

[Holt *et al.*, 2019] Holt C E, Martin K C, Schuman E M (2019); Local translation in neurons: visualization and function. *Nature structural & molecular biology*, 26(7):557–566, URL http://dx.doi.org/10.1038/s41594-019-0263-5.

[Horlacher *et al.*, 2023] Horlacher M, Cantini G, Hesse J, Schinke P, Goedert N, Londhe S, Moyon L, Marsico A (2023); A systematic benchmark of machine learning methods for protein–RNA interaction prediction. *Briefings in Bioinformatics*, 24(5):bbad307, URL http://dx.doi.org/10.1093/bib/bbad307.

[Horvathova *et al.*, 2017] Horvathova I, Voigt F, Kotrys A V, Zhan Y, Artus-Revel C G, Eglinger J, Stadler M B, Giorgetti L, Chao J A (2017); The Dynamics of mRNA Turnover Revealed by Single-Molecule Imaging in Single Cells. *Molecular Cell*, 68(3):615–625.e9, URL http://dx.doi.org/10.1016/j.molcel.2017.09.030.

[Houseley and Tollervey, 2009] Houseley J, Tollervey D (2009); The Many Pathways of RNA Degradation. *Cell*, 136(4):763–776, URL http://dx.doi.org/10.1016/j.cell.2009.01.019.

[Hubstenberger *et al.*, 2017] Hubstenberger A, Courel M, Bénard M, Souquere S, Ernoult-Lange M, Chouaib R, Yi Z, Morlot J B, Munier A, Fradet M, Daunesse M, Bertrand E, Pierron G, Mozziconacci J, Kress M, Weil D (2017); P-Body Purification Reveals the Condensation of Repressed mRNA Regulons. *Molecular Cell*, 68(1):144–157.e5, URL http://dx.doi.org/10.1016/j.molcel.2017.09.003.

[Hunt *et al.*, 2014] Hunt D, Leventer R J, Simons C, Taft R, Swoboda K J, Gawne-Cain M, Magee A C, Turnpenny P D, Baralle D (2014); Whole exome sequencing in family trios reveals de novo mutations in PURA as a cause of severe neurodevelopmental delay and learning disability. *Journal of Medical Genetics*, 51(12):806, URL http://dx.doi.org/10.1136/jmedgenet-2014-102798.

[Ignatiadis *et al.*, 2016] Ignatiadis N, Klaus B, Zaugg J B, Huber W (2016); Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13(7):577–580, URL http://dx.doi.org/10.1038/nmeth.3885.

[Imamura *et al.*, 2014] Imamura K, Imamachi N, Akizuki G, Kumakura M, Kawaguchi A, Nagata K, Kato A, Kawaguchi Y, Sato H, Yoneda M, Kai C, Yada T, Suzuki Y, Yamada T, Ozawa T, Kaneki K, Inoue T, Kobayashi M, Kodama T, Wada Y, Sekimizu K, Akimitsu N (2014); Long Noncoding RNA NEAT1-Dependent SFPQ Relocation from Promoter Region to Paraspeckle Mediates IL8 Expression upon Immune Stimuli. *Molecular Cell*, 53(3):393–406, URL http://dx.doi.org/10.1016/j.molcel.2014.01.009.

[Ingolia, 2016] Ingolia N (2016); Ribosome Footprint Profiling of Translation throughout the Genome. *Cell*, 165(1):22–33, URL http://dx.doi.org/10.1016/j.cell.2016.02.066.

[Ingolia *et al.*, 2009] Ingolia N T, Ghaemmaghami S, Newman J R S, Weissman J S (2009); Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, 324(5924):218–223, URL http://dx.doi.org/10.1126/science.1168978.

[Ishizuka *et al.*, 1995] Ishizuka N, Cowan W M, Amaral D G (1995); A quantitative analysis of the dendritic organization of pyramidal cells in the rat hippocampus. *Journal of Comparative Neurology*, 362(1):17–45, URL http://dx.doi.org/10.1002/cne.903620103.

[Jankowsky and Harris, 2015] Jankowsky E, Harris M E (2015); Specificity and nonspecificity in RNA–protein interactions. *Nature Reviews Molecular Cell Biology*, 16(9):533–544, URL http://dx.doi.org/10.1038/nrm4032.

[Janowski and Niessing, 2020] Janowski R, Niessing D (2020); The large family of PC4-like domains - similar folds and functions throughout all kingdoms of life. *RNA Biology*, 17(9):1–11, URL http://dx.doi.org/10.1080/15476286.2020.1761639.

[Ji *et al.*, 2007] Ji J, Tsika G L, Rindt H, Schreiber K L, McCarthy J J, Kelm R J, Tsika R (2007); Purα and Purβ Collaborate with Sp3 To Negatively Regulate β-Myosin Heavy Chain Gene Expression duringSkeletal Muscle Inactivity. *Molecular and Cellular Biology*, 27(4):1531–1543, URL http://dx.doi.org/10.1128/mcb.00629-06.

[Jin *et al.*, 2007] Jin P, Duan R, Qurashi A, Qin Y, Tian D, Rosser T C, Liu H, Feng Y, Warren S T (2007); Pur α Binds to rCGG Repeats and Modulates Repeat-Mediated Neurodegeneration in a Drosophila Model of Fragile X Tremor/Ataxia Syndrome. *Neuron*, 55(4):556–564, URL http://dx.doi.org/10.1016/j.neuron.2007.07.020.

[Johnson *et al.*, 2006] Johnson E M, Kinoshita Y, Weinreb D B, Wortman M J, Simon R, Khalili K, Winckler B, Gordon J (2006); Role of Purα in targeting mRNA to sites of translation in hippocampal neuronal dendrites. *Journal of Neuroscience Research*, 83(6):929–943, URL http://dx.doi.org/10.1002/jnr.20806.

[Johnson *et al.*, 2013] Johnson E M, Daniel D C, Gordon J (2013); The pur protein family: Genetic and structural features in development and disease. *Journal of Cellular Physiology*, 228(5):930–937, URL http://dx.doi.org/10.1002/jcp.24237.

[Jumper *et al.*, 2021] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl S A A, Ballard A J, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, *et al.* (2021); Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, URL http://dx.doi.org/10.1038/s41586-021-03819-2.

[Kanai *et al.*, 2004] Kanai Y, Dohmae N, Hirokawa N (2004); Kinesin Transports RNA Isolation and Characterization of an RNA-Transporting Granule. *Neuron*, 43(4):513–525, URL http://dx.doi.org/10.1016/j.neuron.2004.07.022.

[Kanakamani *et al.*, 2021] Kanakamani S, Suresh P S, Venkatesh T (2021); Regulation of processing bodies: From viruses to cancer epigenetic machinery. *Cell Biology International*, 45(4):708–719, URL http://dx.doi.org/10.1002/cbin.11527.

[Kang and Schuman, 1996] Kang H, Schuman E M (1996); A Requirement for Local Protein Synthesis in Neurotrophin-Induced Hippocampal Synaptic Plasticity. *Science*, 273(5280):1402–1406, URL http://dx.doi.org/10.1126/science.273.5280.1402.

[Kedersha and Anderson, 2007] Kedersha N, Anderson P (2007); Mammalian Stress Granules and Processing Bodies. *Methods in Enzymology*, 431:61–81, URL http://dx.doi.org/10.1016/s0076-6879(07)31005-7.

[Kedersha *et al.*, 2005] Kedersha N, Stoecklin G, Ayodele M, Yacono P, Lykke-Andersen J, Fritzler M J, Scheuner D, Kaufman R J, Golan D E, Anderson P (2005); Stress granules and processing bodies are dynamically linked sites of mRNP remodeling. *The Journal of Cell Biology*, 169(6):871–884, URL http://dx.doi.org/10.1083/jcb.200502088.

[Kedersha *et al.*, 1999] Kedersha N L, Gupta M, Li W, Miller I, Anderson P (1999); RNA-Binding Proteins Tia-1 and Tiar Link the Phosphorylation of Eif-2α to the Assembly of Mammalian Stress Granules. *The Journal of Cell Biology*, 147(7):1431–1442, URL http://dx.doi.org/10.1083/jcb.147.7.1431.

[Khalili *et al.*, 2003] Khalili K, Valle L D, Muralidharan V, Gault W J, Darbinian N, Otte J, Meier E, Johnson E M, Daniel D C, Kinoshita Y, Amini S, Gordon J (2003); Puralpha is essential for postnatal brain development and developmentally coupled cellular proliferation as revealed by genetic inactivation in the mouse. *Molecular and Cellular Biology*, 23(19):6857–75, URL http://dx.doi.org/10.1128/mcb.23.19.6857-6875.2003.

[Khong *et al.*, 2017] Khong A, Matheny T, Jain S, Mitchell S F, Wheeler J R, Parker R (2017); The Stress Granule Transcriptome Reveals Principles of mRNA Accumulation in Stress Granules. *Molecular Cell*, 68(4):808–820.e5, URL http://dx.doi.org/10.1016/j.molcel.2017.10.015.

[Kim *et al.*, 2013] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg S L (2013); TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, URL http://dx.doi.org/10.1186/gb-2013-14-4-r36.

[Kishor *et al.*, 2020] Kishor A, Fritz S E, Haque N, Ge Z, Tunc I, Yang W, Zhu J, Hogg J R (2020); Activation and inhibition of nonsense-mediated mRNA decay control the abundance of alternative polyadenylation products. *Nucleic Acids Research*, 48(13):7468–7482, URL http://dx.doi.org/10.1093/nar/gkaa491.

[Krachmarov *et al.*, 1996] Krachmarov C P, Chepenik L G, Barr-Vagell S, Khalili K, Johnson E M (1996); Activation of the JC virus Tat-responsive transcriptional control element by association of the Tat protein of human immunodeficiency virus 1 with cellular protein Purα. *Proceedings of the National Academy of Sciences*, 93(24):14112–14117, URL http://dx.doi.org/10.1073/pnas.93.24.14112.

[Krakau *et al.*, 2017] Krakau S, Richard H, Marsico A (2017); PureCLIP: capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome Biology*, 18(1):240, URL http://dx.doi.org/10.1186/s13059-017-1364-2.

[Köhler and Hurt, 2007] Köhler A, Hurt E (2007); Exporting RNA from the nucleus to the cytoplasm. *Nature Reviews Molecular Cell Biology*, 8(10):761–773, URL http://dx.doi.org/10.1038/nrm2255.

[König *et al.*, 2010] König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner D J, Luscombe N M, Ule J (2010); iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology*, 17(7):909–915, URL http://dx.doi.org/10.1038/nsmb.1838.

[Lambert *et al.*, 2014] Lambert N, Robertson A, Jangi M, McGeary S, Sharp P, Burge C (2014); RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Molecular Cell*, 54(5):887–900, URL http://dx.doi.org/10.1016/j.molcel.2014.04.016.

[Langdon and Gladfelter, 2018] Langdon E M, Gladfelter A S (2018); A New Lens for RNA Localization: Liquid-Liquid Phase Separation. *Annual Review of Microbiology*, 72(1):255–271, URL http://dx.doi.org/10.1146/annurev-micro-090817-062814.

[Lawrence *et al.*, 2013] Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan M T, Carey V J (2013); Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, 9(8):e1003118, URL http://dx.doi.org/10.1371/journal.pcbi.1003118.

[Lee and Ule, 2018] Lee F C, Ule J (2018); Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Molecular Cell*, 69(3):354–369, URL http://dx.doi.org/10.1016/j.molcel.2018.01.005.

[Li *et al.*, 2009] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup G P D P (2009); The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, URL http://dx.doi.org/10.1093/bioinformatics/btp352.

[Licatalosi *et al.*, 2008] Licatalosi D D, Mele A, Fak J J, Ule J, Kayikci M, Chi S W, Clark T A, Schweitzer A C, Blume J E, Wang X, Darnell J C, Darnell R B (2008); HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469, URL http://dx.doi.org/10.1038/nature07488.

[Lin and Beal, 2006] Lin M T, Beal M F (2006); Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature*, 443(7113):787–795, URL http://dx.doi.org/10.1038/nature05292.

[Liu-Yesucevitz *et al.*, 2010] Liu-Yesucevitz L, Bilgutay A, Zhang Y J, Vanderweyde T, Vanderwyde T, Citro A, Mehta T, Zaarur N, McKee A, Bowser R, Sherman M, Petrucelli L, Wolozin B (2010); Tar DNA Binding Protein-43 (TDP-43) Associates with Stress Granules: Analysis of Cultured Cells and Pathological Brain Tissue. *PLoS ONE*, 5(10):e13250, URL http://dx.doi.org/10.1371/journal.pone.0013250.

[Lonsdale *et al.*, 2013] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker L, Jewell S, *et al.*

(2013); The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, URL http://dx.doi.org/10.1038/ng.2653.

[Lorenz *et al.*, 2011] Lorenz R, Bernhart S H, Siederdissen C H z, Tafer H, Flamm C, Stadler P F, Hofacker I L (2011); ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, URL http://dx.doi.org/10.1186/1748-7188-6-26.

[Lunde *et al.*, 2007] Lunde B M, Moore C, Varani G (2007); RNA-binding proteins: modular design for efficient function. *Nature Reviews Molecular Cell Biology*, 8(6):479–490, URL http://dx.doi.org/10.1038/nrm2178.

[Luo *et al.*, 2019] Luo Y, Hitz B C, Gabdank I, Hilton J A, Kagda M S, Lam B, Myers Z, Sud P, Jou J, Lin K, Baymuradov U K, Graham K, Litton C, Miyasato S R, Strattan J S, Jolanki O, Lee J W, Tanaka F Y, Adenekan P, O'Neill E, Cherry J M (2019); New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, 48(D1):D882–D889, URL http://dx.doi.org/10.1093/nar/gkz1062.

[Lykke-Andersen and Jensen, 2015] Lykke-Andersen S, Jensen T H (2015); Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nature Reviews Molecular Cell Biology*, 16(11):665–677, URL http://dx.doi.org/10.1038/nrm4063.

[Ma *et al.*, 1994] Ma Z W, Bergemann A D, Johnson E M (1994); Conservation in human and mouse Purα of a motif common to several proteins involved in initiation of DNA replication. *Gene*, 149(2):311–314, URL http://dx.doi.org/10.1016/0378-1119(94)90167-8.

[Manley, 2013] Manley J L (2013); SELEX to Identify Protein-Binding Sites on RNA. *Cold Spring Harbor Protocols*, 2013(2):pdb.prot072934, URL http://dx.doi.org/10.1101/pdb.prot072934.

[Markmiller *et al.*, 2018] Markmiller S, Soltanieh S, Server K L, Mak R, Jin W, Fang M Y, Luo E C, Krach F, Yang D, Sen A, Fulzele A, Wozniak J M, Gonzalez D J, Kankel M W, Gao F B, Bennett E J, Lécuyer E, Yeo G W (2018); Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. *Cell*, 172(3):590–604.e13, URL http://dx.doi.org/10.1016/j.cell.2017.12.032.

[Masliah *et al.*, 2013] Masliah G, Barraud P, Allain F H T (2013); RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cellular and Molecular Life Sciences*, 70(11):1875–1895, URL http://dx.doi.org/10.1007/s00018-012-1119-x.

[Maynard *et al.*, 2023] Maynard S A, Ranft J, Triller A (2023); Quantifying postsynaptic receptor dynamics: insights into synaptic function. *Nature Reviews Neuroscience*, 24(1):4–22, URL http://dx.doi.org/10.1038/s41583-022-00647-9.

[Mead *et al.*, 2023] Mead R J, Shan N, Reiser H J, Marshall F, Shaw P J (2023); Amyotrophic lateral sclerosis: a neurodegenerative disorder poised for successful therapeutic translation. *Nature Reviews Drug Discovery*, 22(3):185–212, URL http://dx.doi.org/10.1038/s41573-022-00612-2.

[Middleton *et al.*, 2019] Middleton S A, Eberwine J, Kim J (2019); Comprehensive catalog of dendritically localized mRNA isoforms from sub-cellular sequencing of single mouse neurons. *BMC Biology*, 17(1):5, URL http://dx.doi.org/10.1186/s12915-019-0630-z.

[Mikkelsen *et al.*, 2007] Mikkelsen T S, Ku M, Jaffe D B, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T K, Koche R P, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander E S, Bernstein B E (2007); Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, URL http://dx.doi.org/10.1038/nature06008.

[Mizielinska *et al.*, 2013] Mizielinska S, Lashley T, Norona F E, Clayton E L, Ridler C E, Fratta P, Isaacs A M (2013); C9orf72 frontotemporal lobar degeneration is characterised by frequent neuronal sense and antisense RNA foci. *Acta Neuropathologica*, 126(6):845–857, URL http://dx.doi.org/10.1007/s00401-013-1200-z.

[Molitor *et al.*, 2021] Molitor L, Bacher S, Burczyk S, Niessing D (2021); The Molecular Function of PURA and Its Implications in Neurological Diseases. *Frontiers in Genetics*, 12:638217, URL http://dx.doi.org/10.3389/fgene.2021.638217.

[Molitor *et al.*, 2023] Molitor L, Klostermann M, Bacher S, Merl-Pham J, Spranger N, Burczyk S, Ketteler C, Rusha E, Tews D, Pertek A, Proske M, Busch A, Reschke S, Feederle R, Hauck S M, Blum H, Drukker M, Fischer-Posovszky P, König J, Zarnack K, Niessing D (2023); Depletion of the RNA-binding protein PURA triggers changes in posttranscriptional gene regulation and loss of P-bodies. *Nucleic Acids Research*, 51(3):1297–1316, URL http://dx.doi.org/10.1093/nar/gkac1237.

[Mölder *et al.*, 2021] Mölder F, Jablonski K P, Letcher B, Hall M B, Tomkins-Tinch C H, Sochat V, Forster J, Lee S, Twardziok S O, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J (2021); Sustainable data analysis with Snakemake. *F1000Research*, 10:33, URL http://dx.doi.org/10.12688/f1000research.29032.1.

[Ninomiya *et al.*, 2020] Ninomiya K, Adachi S, Natsume T, Iwakiri J, Terai G, Asai K, Hirose T (2020); LncRNA-dependent nuclear stress bodies promote intron retention through SR protein phosphorylation. *The EMBO Journal*, 39(3):e102729, URL http://dx.doi.org/10.15252/embj.2019102729.

[Nirschl *et al.*, 2017] Nirschl J J, Ghiretti A E, Holzbaur E L F (2017); The impact of cytoskeletal organization on the local regulation of neuronal transport. *Nature Reviews Neuroscience*, 18(10):585–597, URL http://dx.doi.org/10.1038/nrn.2017.100.

[Nystrom, 2023] Nystrom S (2023); memes: motif matching, comparison, and de novo discovery using the MEME Suite. *Bioconductor*, URL http://dx.doi.org/10.18129/b9.bioc.memes.

[Ohashi *et al.*, 2000] Ohashi S, Kobayashi S, Omori A, Ohara S, Omae A, Muramatsu T, Li Y, Anzai K (2000); The Single-Stranded DNA- and RNA-Binding Proteins Pur α and Pur β Link BC1 RNA to Microtubules Through Binding to the Dendrite-Targeting RNA Motifs. *Journal of Neurochemistry*, 75(5):1781–1790, URL http://dx.doi.org/10.1046/j.1471-4159.2000.0751781.x.

[Ohashi *et al.*, 2002] Ohashi S, Koike K, Omori A, Ichinose S, Ohara S, Kobayashi S, Sato T A, Anzai K (2002); Identification of mRNP complexes containing pur alpha, mStaufen, fragile X protein and myosin Va, and their association with rough endoplasmic reticulum equipped with a kinesin motor. *Journal of Biological Chemistry*, 277(40):37804–37810, URL http://dx.doi.org/10.1074/jbc.m203608200.

[Ohn *et al.*, 2008] Ohn T, Kedersha N, Hickman T, Tisdale S, Anderson P (2008); A functional RNAi screen links O-GlcNAc modification of ribosomal proteins to stress granule and processing body assembly. *Nature Cell Biology*, 10(10):1224–1231, URL http://dx.doi.org/10.1038/ncb1783.

[Oksuz *et al.*, 2023] Oksuz O, Henninger J E, Warneford-Thomson R, Zheng M M, Erb H, Vancura A, Overholt K J, Hawken S W, Banani S F, Lauman R, Reich L N, Robertson A L, Hannett N M, Lee T I, Zon L I, Bonasio R, Young R A (2023); Transcription factors interact with RNA to regulate genes. *Molecular Cell*, 83(14):2449–2463.e13, URL http://dx.doi.org/10.1016/j.molcel.2023.06.012.

[Ortiz-González, 2021] Ortiz-González X R (2021); Mitochondrial Dysfunction: A Common Denominator in Neurodevelopmental Disorders? *Developmental Neuroscience*, 43(3-4):222–229, URL http://dx.doi.org/10.1159/000517870.

[Panas *et al.*, 2012] Panas M D, Varjak M, Lulla A, Eng K E, Merits A, Hedestam G B K, McInerney G M (2012); Sequestration of G3BP coupled with efficient translation inhibits stress granules in Semliki Forest virus infection. *Molecular Biology of the Cell*, 23(24):4701–4712, URL http://dx.doi.org/10.1091/mbc.e12-08-0619.

[Panas *et al.*, 2016] Panas M D, Ivanov P, Anderson P (2016); Mechanistic insights into mammalian stress granule dynamics. *Journal of Cell Biology*, 215(3):313–323, URL http://dx.doi.org/10.1083/jcb.201609081.

[Pandey *et al.*, 2020] Pandey P R, Yang J H, Tsitsipatis D, Panda A C, Noh J H, Kim K M, Munk R, Nicholson T, Hanniford D, Argibay D, Yang X, Martindale J L, Chang M W, Jones S W, Hernando E, Sen P, De S, Abdelmohsen K, Gorospe M (2020); circSamd4 represses myogenic transcriptional activity of PUR proteins. *Nucleic Acids Research*, 48(7):3789–3805, URL http://dx.doi.org/10.1093/nar/gkaa035.

[Penberthy *et al.*, 2004] Penberthy W T, Zhao C, Zhang Y, Jessen J R, Yang Z, Bricaud O, Collazo A, Meng A, Lin S (2004); Pur alpha and Sp8 as opposing regulators of neural gata2 expression. *Developmental Biology*, 275(1):225–34, URL http://dx.doi.org/10.1016/j.ydbio.2004.08.007.

[Perez-Riverol *et al.*, 2019] Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu D J, Inuganti A, Griss J, Mayer G, Eisenacher M, Pérez E, Uszkoreit J, Pfeuffer J, Sachsenberg T, Yılmaz S, Tiwary S, Cox J, Audain E, Walzer M, Jarnuczak A F, Ternent T, Brazma A, Vizcaíno J A (2019); The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research*, 47(Database issue):D442–D450, URL http://dx.doi.org/10.1093/nar/gky1106.

[Proske *et al.*, 2023] Proske M, Janowski R, Bacher S, Kang H S, Monecke T, Köhler T, Hutten S, Tretter J, Crois A, Molitor L, Varela-Rial A, Fino R, Donati E, Fabritiis G D, Dormann D, Sattler M, Niessing D (2023); PURA Syndrome-causing mutations impair PUR-domain integrity and affect P-body association. *BioRxiv*, URL http://dx.doi.org/10.1101/2023.09.19.558386.

[Proudfoot *et al.*, 2002] Proudfoot N J, Furger A, Dye M J (2002); Integrating mRNA Processing with Transcription. *Cell*, 108(4):501–512, URL http://dx.doi.org/10.1016/s0092-8674(02)00617-7.

[Quinlan and Hall, 2010] Quinlan A R, Hall I M (2010); BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, URL http://dx.doi.org/10.1093/bioinformatics/btq033.

[Ramos *et al.*, 2000] Ramos A, Grünert S, Adams J, Micklem D R, Proctor M R, Freund S, Bycroft M, Johnston D S, Varani G (2000); RNA recognition by a Staufen double-stranded RNA-binding domain. *The EMBO Journal*, 19(5):997–1009, URL http://dx.doi.org/10.1093/emboj/19.5.997.

[Ray *et al.*, 2013] Ray D, Kazan H, Cook K B, Weirauch M T, Najafabadi H S, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat L H, Dale R K, Smith S A, Yarosh C A, Kelly S M, Nabet B, Mecenas D, Li W, Laishram R S, Qiao M, Lipshitz H D, Piano F, Corbett A H, *et al.* (2013); A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, URL http://dx.doi.org/10.1038/nature12311.

[Reijnders *et al.*, 2018] Reijnders M R F, Janowski R, Alvi M, Self J E, Essen T J v, Vreeburg M, Rouhl R P W, Stevens S J C, Stegmann A P A, Schieving J, Pfundt R, Dijk K v, Smeets E, Stumpel C T R M, Bok L A, Cobben J M, Engelen M, Mansour S, Whiteford M, Chandler K E, Douzgou S, Cooper N S, Tan E C, Foo R, Lai A H M, *et al.* (2018); PURA syndrome: clinical delineation and genotype-phenotype study in 32 individuals with review of published literature. *Journal of Medical Genetics*, 55(2):104, URL http://dx.doi.org/10.1136/jmedgenet-2017-104946.

[Reiter *et al.*, 2021] Reiter T, Brooks† P T, Irber† L, Joslin† S E K, Reid† C M, Scott† C, Brown C T, Pierce-Ward N T (2021); Streamlining data-intensive biology with workflow systems. *GigaScience*, 10(1):giaa140, URL http://dx.doi.org/10.1093/gigascience/giaa140.

[Renton *et al.*, 2011] Renton A E, Majounie E, Waite A, Simón-Sánchez J, Rollinson S, Gibbs J R, Schymick J C, Laaksovirta H, Swieten J C v, Myllykangas L, Kalimo H, Paetau A, Abramzon Y, Remes A M, Kaganovich A, Scholz S W, Duckworth J, Ding J, Harmer D W, Hernandez D G, Johnson J O, Mok K, Ryten M, Trabzuni D, Guerreiro R J, *et al.* (2011); A Hexanucleotide Repeat Expansion in C9ORF72 Is the Cause of Chromosome 9p21-Linked ALS-FTD. *Neuron*, 72(2):257–268, URL http://dx.doi.org/10.1016/j.neuron.2011.09.010.

[Reuter *et al.*, 2015] Reuter J, Spacek D V, Snyder M (2015); High-Throughput Sequencing Technologies. *Molecular Cell*, 58(4):586–597, URL http://dx.doi.org/10.1016/j.molcel.2015.05.004.

[Riggs *et al.*, 2020] Riggs C L, Kedersha N, Ivanov P, Anderson P (2020); Mammalian stress granules and P bodies at a glance. *Journal of Cell Science*, 133(16):jcs242487, URL http://dx.doi.org/10.1242/jcs.242487.

[Roach *et al.*, 2022] Roach M J, Pierce-Ward N T, Suchecki R, Mallawaarachchi V, Papudeshi B, Handley S A, Brown C T, Watson-Haigh N S, Edwards R A (2022); Ten simple rules and a template for creating workflows-as-applications. *PLOS Computational Biology*, 18(12):e1010705, URL http://dx.doi.org/10.1371/journal.pcbi.1010705.

[Rodnina, 2023] Rodnina M V (2023); Decoding and Recoding of mRNA Sequences by the Ribosome. *Annual Review of Biophysics*, 52(1):161–182, URL http://dx.doi.org/10.1146/annurev-biophys-101922-072452.

[Rosenfeld *et al.*, 2011] Rosenfeld J A, Drautz J M, Clericuzio C L, Cushing T, Raskin S, Martin J, Tervo R C, Pitarque J A, Nowak D M, Karolak J A, Lamb A N, Schultz R A, Ballif B C, Bejjani B A, Gajecka M, Shaffer L G (2011); Deletions and duplications of developmental pathway genes in 5q31 contribute to abnormal phenotypes. *American Journal of Medical Genetics Part A*, 155(8):1906–1916, URL http://dx.doi.org/10.1002/ajmg.a.34100.

[Salcedo-Arellano *et al.*, 2020] Salcedo-Arellano M J, Dufour B, McLennan Y, Martinez-Cerdeno V, Hagerman R (2020); Fragile X syndrome and associated disorders: Clinical aspects and pathology. *Neurobiology of Disease*, 136:104740, URL http://dx.doi.org/10.1016/j.nbd.2020.104740.

[Sanders *et al.*, 2020] Sanders D W, Kedersha N, Lee D S, Strom A R, Drake V, Riback J A, Bracha D, Eeftens J M, Iwanicki A, Wang A, Wei M T, Whitney G, Lyons S M, Anderson P, Jacobs W M, Ivanov P, Brangwynne C P (2020); Competing Protein-RNA Interaction Networks Control

Multiphase Intracellular Organization. *Cell*, 181(2):306–324.e28, URL http://dx.doi.org/10.1016/j.cell.2020.03.050.

[Sanger *et al.*, 1977] Sanger F, Nicklen S, Coulson A R (1977); DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, URL http://dx.doi.org/10.1073/pnas.74.12.5463.

[Sartor *et al.*, 2006] Sartor M A, Tomlinson C R, Wesselkamper S C, Sivaganesan S, Leikauf G D, Medvedovic M (2006); Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics*, 7(1):538, URL http://dx.doi.org/10.1186/1471-2105-7-538.

[Shah *et al.*, 2016] Shah A, Qian Y, Weyn-Vanhentenryck S M, Zhang C (2016); CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. *Bioinformatics*, 33(4):566–567, URL http://dx.doi.org/10.1093/bioinformatics/btw653.

[Shi *et al.*, 2021] Shi X, Ren S, Zhang B, Guo S, He W, Yuan C, Yang X, Ig-lzevbekhai K, Sun T, Wang Q, Cui J (2021); Analysis of the role of Purα in the pathogenesis of Alzheimer's disease based on RNA-seq and ChIP-seq. *Scientific Reports*, 11(1):12178, URL http://dx.doi.org/10.1038/s41598-021-90982-1.

[Shin and Brangwynne, 2017] Shin Y, Brangwynne C P (2017); Liquid phase condensation in cell physiology and disease. *Science*, 357(6357):eaaf4382, URL http://dx.doi.org/10.1126/science.aaf4382.

[Singh *et al.*, 2015] Singh G, Pratt G, Yeo G W, Moore M J (2015); The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion. *Annual Review of Biochemistry*, 84(1):1–30, URL http://dx.doi.org/10.1146/annurev-biochem-080111-092106.

[Singh and Valcárcel, 2005] Singh R, Valcárcel J (2005); Building specificity with nonspecific RNA-binding proteins. *Nature Structural & Molecular Biology*, 12(8):645–653, URL http://dx.doi.org/10.1038/nsmb961.

[Sloggett *et al.*, 2013] Sloggett C, Goonasekera N, Afgan E (2013); BioBlend: automating pipeline analyses within Galaxy and CloudMan. *Bioinformatics*, 29(13):1685–1686, URL http://dx.doi.org/10.1093/bioinformatics/btt199.

[Smith *et al.*, 2017] Smith T, Heger A, Sudbery I (2017); UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, 27(3):491–499, URL http://dx.doi.org/10.1101/gr.209601.116.

[Sofola *et al.*, 2007] Sofola O A, Jin P, Qin Y, Duan R, Liu H, Haro M d, Nelson D L, Botas J (2007); RNA-Binding Proteins hnRNP A2/B1 and CUGBP1 Suppress Fragile X CGG Premutation Repeat-

Induced Neurodegeneration in a Drosophila Model of FXTAS. *Neuron*, 55(4):565–571, URL http://dx.doi.org/10.1016/j.neuron.2007.07.021.

[Song *et al.*, 2022] Song C, Zhang Y, Huang W, Shi J, Huang Q, Jiang M, Qiu Y, Wang T, Chen H, Wang H (2022); Circular RNA Cwc27 contributes to Alzheimer's disease pathogenesis by repressing Pur-α activity. *Cell Death & Differentiation*, 29(2):393–406, URL http://dx.doi.org/10.1038/s41418-021-00865-1.

[Spector, 2006] Spector D L (2006); SnapShot: Cellular Bodies. *Cell*, 127(5):1071.e1–1071.e2, URL http://dx.doi.org/10.1016/j.cell.2006.11.026.

[Stacey *et al.*, 1999] Stacey D W, Hitomi M, Kanovsky M, Gan L, Johnson E M (1999); Cell cycle arrest and morphological alterations following microinjection of NIH3T3 cells with Pur alpha. *Oncogene*, 18(29):4254–4261, URL http://dx.doi.org/10.1038/sj.onc.1202795.

[Sugimoto *et al.*, 2012] Sugimoto Y, König J, Hussain S, Zupan B, Curk T, Frye M, Ule J (2012); Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biology*, 13(8):R67, URL http://dx.doi.org/10.1186/gb-2012-13-8-r67.

[Sun *et al.*, 2020] Sun Y, Gao J, Jing Z, Zhao Y, Sun Y, Zhao X (2020); PURα Promotes the Transcriptional Activation of PCK2 in Oesophageal Squamous Cell Carcinoma Cells. *Genes*, 11(11):1301, URL http://dx.doi.org/10.3390/genes11111301.

[Sutandy *et al.*, 2018] Sutandy F X R, Ebersberger S, Huang L, Busch A, Bach M, Kang H S, Fallmann J, Maticzka D, Backofen R, Stadler P F, Zarnack K, Sattler M, Legewie S, König J (2018); In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. *Genome Research*, 28(5):699–713, URL http://dx.doi.org/10.1101/gr.229757.117.

[Swinnen *et al.*, 2020] Swinnen B, Robberecht W, Bosch L V D (2020); RNA toxicity in non-coding repeat expansion disorders. *The EMBO Journal*, 39(1):e101112, URL http://dx.doi.org/10.15252/embj.2018101112.

[Tian *et al.*, 2022] Tian L, Xie X, Das U, Chen Y, Sun Y, Liu F, Lu H, Nan P, Zhu Y, Gu X, Deng H, Xie J, Zhao X (2022); Forming cytoplasmic stress granules PURα suppresses mRNA translation initiation of IGFBP3 to promote esophageal squamous cell carcinoma progression. *Oncogene*, 41(38):4336–4348, URL http://dx.doi.org/10.1038/s41388-022-02426-3.

[Tommaso *et al.*, 2017] Tommaso P D, Chatzou M, Floden E W, Barja P P, Palumbo E, Notredame C (2017); Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, URL http://dx.doi.org/10.1038/nbt.3820.

[Ule *et al.*, 2003] Ule J, Jensen K B, Ruggiu M, Mele A, Ule A, Darnell R B (2003); CLIP Identifies Nova-Regulated RNA Networks in the Brain. *Science*, 302(5648):1212–1215, URL http://dx.doi.org/10.1126/science.1090095.

[Ule *et al.*, 2005] Ule J, Jensen K, Mele A, Darnell R B (2005); CLIP: A method for identifying protein–RNA interaction sites in living cells. *Methods*, 37(4):376–386, URL http://dx.doi.org/10.1016/j.ymeth.2005.07.018.

[Uren *et al.*, 2012] Uren P J, Bahrami-Samani E, Burns S C, Qiao M, Karginov F V, Hodges E, Hannon G J, Sanford J R, Penalva L O F, Smith A D (2012); Site identification in high-throughput RNA–protein interaction data. *Bioinformatics*, 28(23):3013–3020, URL http://dx.doi.org/10.1093/bioinformatics/bts569.

[Urlaub *et al.*, 2002] Urlaub H, Hartmuth K, Lührmann R (2002); A two-tracked approach to analyze RNA–protein crosslinking sites in native, nonlabeled small nuclear ribonucleoprotein particles. *Methods*, 26(2):170–181, URL http://dx.doi.org/10.1016/s1046-2023(02)00020-8.

[Van Nostrand *et al.*, 2016] Van Nostrand E L, Pratt G A, Shishkin A A, Gelboin-Burkhart C, Fang M Y, Sundararaman B, Blue S M, Nguyen T B, Surka C, Elkins K, Stanton R, Rigo F, Guttman M, Yeo G W (2016); Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6):508–514, URL http://dx.doi.org/10.1038/nmeth.3810.

[Van Nostrand *et al.*, 2020a] Van Nostrand E L, Freese P, Pratt G A, Wang X, Wei X, Xiao R, Blue S M, Chen J Y, Cody N A L, Dominguez D, Olson S, Sundararaman B, Zhan L, Bazile C, Bouvrette L P B, Bergalet J, Duff M O, Garcia K E, Gelboin-Burkhart C, Hochman M, Lambert N J, Li H, McGurk M P, Nguyen T B, Palden T, *et al.* (2020a); A large-scale binding and functional map of human RNA-binding proteins. *Nature*, 583(7818):711–719, URL http://dx.doi.org/10.1038/s41586-020-2077-3.

[Van Nostrand *et al.*, 2020b] Van Nostrand E L, Pratt G A, Yee B A, Wheeler E C, Blue S M, Mueller J, Park S S, Garcia K E, Gelboin-Burkhart C, Nguyen T B, Rabano I, Stanton R, Sundararaman B, Wang R, Fu X D, Graveley B R, Yeo G W (2020b); Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biology*, 21(1):90, URL http://dx.doi.org/10.1186/s13059-020-01982-9.

[Wang *et al.*, 2009] Wang Z, Gerstein M, Snyder M (2009); RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, URL http://dx.doi.org/10.1038/nrg2484.

[Wang *et al.*, 2010] Wang Z, Kayikci M, Briese M, Zarnack K, Luscombe N M, Rot G, Zupan B, Curk T, Ule J (2010); iCLIP Predicts the Dual Splicing Effects of TIA-RNA Interactions. *PLoS Biology*, 8(10):e1000530, URL http://dx.doi.org/10.1371/journal.pbio.1000530.

[Weber *et al.*, 2016] Weber J, Bao H, Hartlmüller C, Wang Z, Windhager A, Janowski R, Madl T, Jin P, Niessing D (2016); Structural basis of nucleic-acid recognition and double-strand unwinding by the essential neuronal protein Pur-alpha. *eLife*, 5:e11297, URL http://dx.doi.org/10.7554/elife.11297.

[White *et al.*, 2009] White M K, Johnson E M, Khalili K (2009); Multiple roles for Pur-α in cellular and viral regulation. *Cell Cycle*, 8(3):414–420, URL http://dx.doi.org/10.4161/cc.8.3.7585.

[Wilkinson *et al.*, 2016] Wilkinson M D, Dumontier M, Aalbersberg I J, Appleton G, Axton M, Baak A, Blomberg N, Boiten J W, Santos L B d S, Bourne P E, Bouwman J, Brookes A J, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C T, Finkers R, Gonzalez-Beltran A, Gray A J, Groth P, Goble C, Grethe J S, Heringa J, *et al.* (2016); The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, URL http://dx.doi.org/10.1038/sdata.2016.18.

[Witze *et al.*, 2009] Witze E S, Field E D, Hunt D F, Rothman J H (2009); C. elegans pur alpha, an activator of end-1, synergizes with the Wnt pathway to specify endoderm. *Developmental Biology*, 327(1):12–23, URL http://dx.doi.org/10.1016/j.ydbio.2008.11.015.

[Wratten *et al.*, 2021] Wratten L, Wilm A, Göke J (2021); Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods*, 18(10):1161–1168, URL http://dx.doi.org/10.1038/s41592-021-01254-9.

[Wright *et al.*, 2014] Wright C F, Fitzgerald T W, Jones W D, Clayton S, McRae J F, Kogelenberg M v, King D A, Ambridge K, Barrett D M, Bayzetinova T, Bevan A P, Bragin E, Chatzimichali E A, Gribble S, Jones P, Krishnappa N, Mason L E, Miller R, Morley K I, Parthiban V, Prigmore E, Rajan D, Sifrim A, Swaminathan G J, Tivey A R, *et al.* (2014); Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet (London, England)*, 385(9975):1305–14, URL http://dx.doi.org/10.1016/s0140-6736(14)61705-0.

[Xu *et al.*, 2013] Xu Z, Poidevin M, Li X, Li Y, Shu L, Nelson D L, Li H, Hales C M, Gearing M, Wingo T S, Jin P (2013); Expanded GGGGCC repeat RNA associated with amyotrophic lateral sclerosis and frontotemporal dementia causes neurodegeneration. *Proceedings of the National Academy of Sciences of the United States of America*, 110(19):7778–7783, URL http://dx.doi.org/10.1073/pnas.1219643110.

[Yeo *et al.*, 2009] Yeo G W, Coufal N G, Liang T Y, Peng G E, Fu X D, Gage F H (2009); An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature Structural & Molecular Biology*, 16(2):130–137, URL http://dx.doi.org/10.1038/nsmb.1545.

[Youn *et al.*, 2019] Youn J Y, Dyakov B J, Zhang J, Knight J D, Vernon R M, Forman-Kay J D, Gingras A C (2019); Properties of Stress Granule and P-Body Proteomes. *Molecular Cell*, 76(2):286–294, URL http://dx.doi.org/10.1016/j.molcel.2019.09.014.

[Zarnack *et al.*, 2013] Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, Reyes A, Anders S, Luscombe N, Ule J (2013); Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements. *Cell*, 152(3):453–466, URL http://dx.doi.org/10.1016/j.cell.2012.12.023.

[Zhang *et al.*, 2013] Zhang Y, Fonslow B R, Shan B, Baek M C, Yates J R (2013); Protein Analysis by Shotgun/Bottom-up Proteomics. *Chemical Reviews*, 113(4):2343–2394, URL http://dx.doi.org/10.1021/cr3003533.

[Zhu *et al.*, 2020] Zhu Y, Orre L M, Tran Y Z, Mermelekas G, Johansson H J, Malyutina A, Anders S, Lehtiö J (2020); DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis. *Molecular & Cellular Proteomics*, 19(6):1047–1057, URL http://dx.doi.org/10.1074/mcp.tir119.001646.

[Zipeto *et al.*, 2015] Zipeto M A, Jiang Q, Melese E, Jamieson C H (2015); RNA rewriting, recoding, and rewiring in human disease. *Trends in Molecular Medicine*, 21(9):549–559, URL http://dx.doi.org/10.1016/j.molmed.2015.07.001.