Computational approaches to decipher splicing regulatory network of the *RON* proto-oncogene

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

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

von
Samarth Thonta Setty
aus Bengaluru (Indien)

Frankfurt (2019)
(D 30)

Vom Fachbereich Biowissenschaften (FB15) der Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr. Sven Klimpel

Gutachter: Dr. Kathi Zarnack, Prof. Dr. Enrico Schleiff

Datum der Disputation: June 03 2020

Table of Contents

Page

List of Figures

Figure Number Page

List of Tables

Table Number Page

Acknowledgments

"My acknowledgments are to those people and ideas that pushed me through this arduous journey"

Dedication

"I wish to dedicate this thesis to all the amazing and awesome women in my life, especially my mother."

Zusammenfassung

Primäre Transkripte oder prä-mRNAs (Precursor-Messenger-RNAs, Pre-mRNAs) werden von der DNA abgeschrieben und erfordern mehrere Verarbeitungsschritte, bevor sie Boten-RNA (mRNAs) oder ‚mature transcripts' genannt werden. In Eukaryonten wird die mRNA von der Transkription im Kern bis zum Export in das Zytoplasma weiterverarbeitet und modifiziert, wie z.B. durch (alternatives) Spleißen (Ben-Yishay et al., 2019). Spleißprozesse entstehen indem die Transkripte, die diskontinuierlich in nicht-kodierende Introns und kodierende Exons organisiert sind, mit den Spleißkomponenten in regulatorischen Bereichen interagieren, um die Introns aus der prä-mRNA zu entfernen. Das resultierende, durchgehende Transkript des Genes kann dann in die Proteinsynthese eingehen bzw. bei fehlerhaft prozessierten Transkripten in dessen Abbau resultieren. Alternatives Spleißen ist ein co- oder posttranskriptioneller Prozess, bei dem ein einzelnes Gen zu unterschiedlichen Transkript-Isoformen führt, indem Exons selektiv eingefügt oder übersprungen werden. Dieser "Teilen-und-Kombinieren"-Schritt multipliziert die eukaryotische Proteom-Diversität um ein Vielfaches und ist aufgrund seiner fundamentalen Auswirkungen an einer Vielzahl von Krankheiten beteiligt (Yang et al., 2016). Hinzu kommt, dass alternatives Spleißen gewebe- und entwicklungsstadien-spezifisch determiniert ist und damit an wichtigen biologischen Prozessen wie Gewebehomöostase, Zell- und Organdifferenzierung beteiligt ist (Baralle

et al., 2017). Alternatives Spleißen wird in einer mehrstufigen Spleißreaktion mithilfe einer Vielzahl von Proteinen des Spleißosoms ermöglicht, was eine präzise Steuerung und Modulation ermöglicht (Johnson et al., 2012). Um diese präzise Steuerung zu gewährleisten, benötigt die Spleißmaschine zusätzliche *trans*-agierende Faktoren oder RNA-Bindeproteine (RBP). Diese RBPs binden an RNA-Sequenzelemente oder *cis*-regulatorische Elemente innerhalb der prä-mRNA, steuern die Spleißosomenaktivität und ermöglichen somit kontextabhängiges alternatives Spleißen (Fu et al., 2014). Im Einzelnen werden *cis*-regulatorische Elemente in exonische oder intronische Spleißverstärker (ESEs oder ISEs) sowie Spleißhemmer (ESSs oder ISSs) eingeteilt, je nachdem, ob sie den Exon-Einbau verstärken oder unterdrücken (Kornblihtt et al., 2013). *Cis*-regulatorische Elemente bestehen aus RNA-Sequenzen oder aus Motiven, wie z.B. ”G-Runs”, die Strukturen wie G-Quadruplexe ausbilden (Oberstrass et al., 2005; Warf et al., 2009). Zusätzlich zur regulatorischen Rolle lenken *trans*-agierende Faktoren die Spleißosomanordnung über Exons hinweg, um die sogenannte Exon-Definition zu vermitteln (Conti et al., 2013). Auffallend ist, dass die repressive oder aktivierende Rolle von RBPs von ihren jeweiligen Bindungsorten in der prä-mRNA abhängt (Erkelenz et al., 2013). So erfolgt z.B. eine positionsabhängige Regulation, entweder als Repressor oder Aktivator des Spleißens, durch RBPs wie HNRNP- oder SR-Proteine. In verschiedenen Studien konnte gezeigt werden, wie HNRNP-Proteine das Spleißen bei Bindung an ein Exon hemmen (Rothrock et al., 2005; Mauger et al., 2008), jedoch Spleißen bei Bindung an ein Intron aktivieren (Hui et al., 2003; Xiao et al., 2009). Umgekehrt wurde gezeigt, dass exongebundene SR-Proteine das Spleißen aktivieren (Cartegni et al., 2002; Shen et al., 2004) und introngebundene SR-Proteine das Spleißen unterdrücken (Ibrahim et al., 2005; Buratti et al., 2007; Shen et al., 2012). Zusätzlich werden Spleißereignisse durch ein Zusammenspiel mehrerer *cis*-regulatorischer Elemente in Kombination mit ihren korrespondierenden RBPs gesteuert (Nasrin et al., 2014; Qian et al., 2014).

Die modulare Organisation dieser Interaktionen deutet auf die Existenz eines ”Spleißcodes" hin, d.h. eines Codes von prä-mRNA-Merkmalen, der von spezifischen RBPs dekodiert wird. Umfangreiche bioinformatische Studien ermöglichen *in silico* Vorhersagen über das Ergebnis von Spleißprozessen, durch die Integration von Informationen über RNA-Merkmale mit RNA-Sequenzierungsdaten (RNA-Seq), mit dem Ziel, den "Spleißcode" zu entschlüsseln (Barash et al., 2013; Xiong et al., 2015). Die resultierenden Algorithmen sind nützlich, um prinzipielle Trends zu erkennen und um die Einführung von Mutationen in experimentellen Analysen vorzuschlagen. Ihre Vorhersagekraft ist jedoch gering, wenn es darum geht, die Folgen einzelner Nukleotidaustausche auf alternatives Spleißen zu bewerten (Soukarieh et al., 2016). Obwohl einige dieser regulatorischen Wechselwirkungen bereits im Detail beschrieben sind, fehlt uns ein umfassendes Verständnis des regulatorischen Codes, der einer Spleißentscheidung zugrunde liegt. Um ein bestimmtes Spleißereignis im Detail zu untersuchen, ist außerdem eine experimentelle Validierung der *in silico* Vorhersagen erforderlich, was häufig die Mutagenese von Minigen-Reportern beinhaltet (Cooper, 2005). Jüngste Hochdurchsatzstudien trugen dazu bei, ein besseres Verständnis des Spleißcodes zu erreichen, indem sie die Minigene-Reporter-Mutagenese mit Hochdurchsatz-Sequenzierung ("Deep Sequencing") kombinierten. Dies ermöglichte eine genaue Vorhersage des Einflusses von Sequenzvarianten auf das Spleißen, lieferte ein nahezu vollständiges Mutationsscreening eines menschlichen Exons und erlaubte die Herleitung allgemeiner Spleißmechanismen (Rosenberg et al., 2015; Julien et al., 2016; Ke et al., 2018).

In all diesen Studien beschränkten sich die Mutationen jedoch auf exonische Regionen oder synthetische Sequenzen, während *trans*-agierende Faktoren nicht berücksichtigt wurden. Diese Einschränkungen führten zur Notwendigkeit alternativer Ansätze, um die Regulierung des Spleißens in mehr Detail zu untersuchen. Als Konsequenz haben wir in diesem Projekt einen Hochdurchsatz-Screeningansatz etabliert, um *cis*-regulatorische Elemente, die eine bestimmte Spleißentscheidung steuern, umfassend zu identifizieren und zu charakterisieren. Hierzu war es wichtig, das richtige Minigen-Reporter-System für die

Durchführung der Zufallsmutagenese auszuwählen. Zu diesem Zweck wurde aus dem genomischen Lokus des *RON*-Gens (Recepteur d'origine nantais) ein Minigen-Konstrukt generiert, das aus dem alternativ gespleißten Exon 11, seinen vollständigen flankierenden Introns sowie den Exons 10 und 12 besteht. Die ausschlaggebenden Kriterien für die Wahl des *RON* Minigen-Reporter-Systems war der Aspekt, dass es mit hoher Effizient richtig gespleißt wurde, die Existenz früheren Studien zu seiner Regulation und seine klinische Relevanz bei der Krebs-Pathogenese. RON ist eine Rezeptor-Tyrosin-Kinase, die durch das Proto-Onkogen *MST1R* (auch als *RON* bezeichnet; Gaudino et al., 1994) kodiert wird. Unter normalen physiologischen Bedingungen löst die Auto-Phosphorylierung der Protein-Tyrosin-Kinase-Domäne die nachgeschaltete Signalkaskade aus (Schlessinger, 2000), die biologische Funktionen wie die Kontrolle der Immunantwort bei Entzündungen und Wundheilung ermöglicht (Yao et al., 2013; Faham et al., 2016). Das Überspringen des alternativen *RON*-Exons 11 führt zu der Isoform RONΔ165, die bei metastasierendem Krebs hochreguliert wird und durch die Förderung des Übergangs von epithelialem zu mesenchymalem Gewebe zur Tumorinvasivität beiträgt (Collesi et al., 1996; Zhou et al., 2003). Darüber hinaus wird RONΔ165 häufig bei soliden Tumoren, einschließlich Eierstock-, Bauchspeicheldrüsen-, Brust- und Darmkrebs, hochreguliert (Ghigna et al., 2005; Mayer et al., 2015; Chakedis et al., 2016). Aus früheren Studien ging hervor, dass das Spleißen von Exon 11 weitgehend durch *trans*-agierende Faktoren wie HNRNPH reguliert wird (Lefave et al., 2011; Bonomi et al., 2013). Obwohl bereits einige wenige *cis*-regulatorische Elemente identifiziert wurden, die den Einbau von Exon 11 regulieren (Lefave et al., 2011), gibt es bisher keine umfassende und systematische Studie, die es ermöglicht hätte, sowohl zuvor entdeckte als auch noch unbekannte *cis*-regulatorische Elemente zu analysieren. Daher wurde das Proto-Onkogen *RON* als Minigen ausgewählt, um einen solchen Screening-Ansatz zu durchzuführen.

Hierfür wurde eine "mutagene PCR" durchgeführt, um eine Bibliothek mit mutagenisierten Varianten des *RON*-Minigen zu generieren. Eine "paired-end" Hochdurchsatz-

DNA-Sequenzierung identifizierte die resultierenden Kombinationen von Mutationen in der Minigen-Bibliothek. Hierauf wurde die Bibliothek als "Pool" in menschliche Zellen transfiziert; die alternativ gespleißten Isoformen wurden durch "paired-end" Hochdurchsatz-RNA-Sequenzierung quantifiziert. Wichtig war die Verwendung einer Barcodesequenz, um jede Minigen-Variante zu kennzeichnen und eindeutig in den resultierenden Isoformen identifizieren zu können. Dieser Ansatz ermöglichte die paarweise Zuordnung von mutierten Minigenvarianten zu den entsprechenden Spleißisoformen. Insgesamt wurden 5.791 Minigen-Varianten erfasst, darunter 5.200 mit Mutationen und 591 mit der Wildtyp (wt)-Sequenz. Die Identifikation der Mutationen ("Mutation calling") zeigte 18.948 Punktmutationen, die gleichmäßig über alle Positionen der *RON*-Minigensequenz verteilt waren, womit Mutationsereignisse über die gesamte Länge des Minigens beobachtet wurden. Um die verschiedenen Spleißisoformen der Minigen-Bibliothek zu erfassen, wurde die Bibliothek zunächst als "Pool" in menschliche HEK293T-Zellen transfiziert. Anschließend wurden die resultierenden alternativ gespleißten Transkripte mittels RNA-Sequenzierung mit einer Primerkombination quantifiziert, die eine eindeutige Identifizierung aller kanonischen Isoformen ermöglichte. Die kanonischen Isoformen umfassten den Einbau des alternative Exons (AE) („exon inclusion"), den AE-Ausschluss („exon skipping"), die vollständige Intronretention (IR), sowie Retention des ersten Introns (erste IR) bzw. des zweiten Introns (zweite IR) und machten 94% aller Spleißprodukte aus. Die restlichen 6% der Isoformen resultierten aus der Verwendung von kryptischen 3'- und 5'-Spleißstellen. Sanger-Sequenzierung und RT-PCR von zufällig ausgewählten Minigenvarianten bestätigten die Genauigkeit der Mutationsbestimmung und der Quantifizierungen der RNA-Sequenzierungen. Zusammenfassend ermöglichte das Hochdurchsatz-Mutagenese-Screening die Analyse von Mutationseffekten über die gesamte Sequenzlänge des RON-Minigens.

Die resultierenden Minigene-Varianten enthielten im Durchschnitt 3,6 Mutationen. Das bedeutet, dass das beobachtete Spleißmuster einer Minigen-Variante häufig das Ergebnis mehrerer Mutationen darstellte. Weiterhin kann dieselbe Mutation zu unterschiedlichen Spleißeffekten führen, in Abhängigkeit von weiteren, im gleichen Minigen auftretenden Mutationen. Um die Effekte der einzelnen Mutationen zu erfassen, wurde eine lineare Regressionsmodellierung angewendet, bei der ein dynamisches Spleißmodell formuliert wurde, das die Berechnung der linearen Regression auf der Grundlage von Spleißisoform-Verhältnissen erlaubt (Isoform-Entstehung versus Isoform-Abbau im Verhältnis zur Rate der AE-Einschlußisoform als Referenz). Die Betrachtung der Verhältnisse der Isoformen anstelle der absoluten Frequenzen bei der linearen Regression berücksichtigte die Nicht-linearität von Mutationseffekten. Unter der Annahme, dass Mutationseffekte additiv sind, berechnete die lineare Regression einzelne Mutationseffekte („log-fold changes" relativ zum Wildtyp) aus der Summe mehrerer Mutationseffekte in einem Minigen. Um das Regressionsmodell zu validieren, wurden hieraus abgeleitete Einzelmutationseffekte mit RT-PCR-abgeleiteten Spleißmessungen von neu erstellten Minigenen mit Einzelmutationen verglichen. Zum Vergleich mit der Modellierung wurde zusätzlich die mittlere Isoform-Frequenz über alle Minigenvarianten, die eine bestimmte Mutation aufwiesen, als vereinfachter Ansatz zur Abschätzung einzelner Mutationseffekte berechnet. Diese medianbasierte Schätzung wurde jedoch durch die von der linearen Regression abgeleiteten Einzelmutationseffekte übertroffen. Insbesondere bei Mutationen, die mit niedriger Frequenz in der Bibliothek auftraten, d.h. im Falle, dass nur wenige Minigenvarianten eine bestimmte Mutation teilten, konnten diese Mutationen den Median leicht vom tatsächlichen Einzelmutationseffekt verschieben. Zusammenfassend ermöglichte die Verwendung eines linearen Regressionsmodells unter der Annahme additiver Mutationseffekte die genaue Abschätzung der Effekte einzelner Mutationen aus den gemessenen, kombinierten Mutationseffekten der ca. 1.800 Mutationen auf die fünf kanonischen Isoformen. Innerhalb aller Mutationen wurde das Spleißen von mindestens einer Isoform durch 778 Mutationen signifikant verändert, welche als spleißwirksame Mutationen bezeichnet wurden (Signifikanz definiert durch ≥ 5% Änderung der Isoform-Frequenz bei 5% Falscherkennungsrate, FDR). Die höchste Dichte an spleißwirksamen Positionen, d.h. Positionen mit mindestens einer spleißwirksamen Mutation, wurde im alternativen Exon gefunden. Bemerkenswert war, dass 91% aller Positionen innerhalb des *RON*-Exons 11 (134 von 147 Nukleotiden) spleißwirksam waren. Dennoch enthielten auch die vor- und nachgelagerten Introns 77% bzw. 82% spleißwirksame Positionen, und ebenso waren etwa die Hälfte der Positionen in den flankierenden konstitutiven Exons an der Spleißregulation beteiligt. Dies zeigte, dass der untersuchte Sequenzbereich dicht mit Spleißregulationsregionen besetzt war und dass Spleißregulation nicht nur von den Nukleotiden innerhalb des alternativen Exons kontrolliert wird, sondern auch von den benachbarten Introns und konstitutiven Exons. Insbesondere das häufige Auftreten von Spleißpositionen in Exons unterstreicht die doppelte Rolle der Proteinkodierungs- und Spleißregulierungsfunktion.

Wie wir zeigten, sind spleißwirksame Mutationen evolutionär sowohl in Exons als auch in Introns konserviert. In Introns unterscheiden sich spleißwirksame Positionen durch eine höhere evolutionäre Konservierung von solchen Positionen, die das Spleißen nicht beeinflussen, was den evolutionären Selektionsdruck zur Aufrechterhaltung der spleißwirksamen Positionen belegt (Xing et al., 2006). Im Gegensatz dazu wurde keine größere Konservierung der spleißwirksamen Positionen in Exons festgestellt, was darauf hindeutet, dass der durch die Spleißfunktion eines Nukleotids erzeugte Selektionsdruck durch proteinkodierende Einschränkungen außer Kraft gesetzt wird. Der Vergleich mit den Mutationswirkungsvorhersagen der *in silico* Applikation SPIDEX (SPANR; Xiong et al., 2015) zeigte, dass die Mutationseffekte mit Ausnahme weniger starker spleißwirksamer Mutationen durch unseren Screening-Ansatz genauer bestimmt wurden. Um zu testen, inwieweit synonyme Mutationen an der Spleißregulation beteiligt sind, wurden die Effekte von synonymen und nicht-synonymen Mutationen verglichen. In Übereinstimmung mit früheren Ergebnissen für *FAS*/*CD95*-Exon 6 (Julien et al., 2016) vermitteln synonyme Mutationen in den *RON*-Exons 10, 11 und 12 eine Spleißregulation mit ähnlichen Effektgrößen wie nichtsynonyme Mutationen. Darüber hinaus verändern die 135 synonymen spleißwirksamen Mutationen nicht die RON-Proteinsequenz, sondern tragen zu pathogenen Auswirkungen durch eine Veränderung von Proteinmenge und -funktion durch alternatives Spleißen bei (Xing et al., 2005; Shabalina et al., 2013). Zusammenfassend lässt sich sagen, dass das alternative Spleißen von *RON*-Exon 11 durch zahlreiche Intron- und Exon-Positionen reguliert wird und dass die Bedeutung dieser regulatorischen Positionen innerhalb von Introns durch ihre evolutionäre Konservierung widergespiegelt wird. Darüber hinaus konnten wir zeigen, dass die spleißwirksamen Mutationen häufig mit synonymen Mutationen zusammenfielen. Da diese Mutationen die Sequenz des kodierten Proteins nicht verändern, könnten sie fälschlicherweise als nicht-pathogen interpretiert werden, während sie tatsächlich die Proteinfunktion durch alternatives Spleißen beeinträchtigen können.

Weiterhin korrelierten die aus dem Screening-Ansatz quantifizierten Mutationseffekte mit dem alternativen *RON*-Spleißen bei Krebspatienten. Da *RON* ein Proto-Onkogen ist und das Überspringen von Exon 11 für die Krebspathophysiologie relevant ist, wurden Mutationseffekte in der nicht-invasiven menschlichen Brustkrebs-Zelllinie MCF7 quantifiziert. Die klinische Relevanz der Mutationen wurde anschließend durch die Analyse von Krebspatientendaten bewertet. Zunächst wurde die Datenbank "Catalogue of Somatic Mutations in Cancer" (COSMIC) auf Mutationen innerhalb der *RON*-Minigenregion untersucht. COSMIC enthält manuell annotierte, somatische Mutationen aus begutachteten Publikationen und Datenbanken. Die Region des *RON*-Minigenes enthielt 33 COSMIC-Einträge, von denen 20 mit spleißwirksamen Mutationen zusammenfielen. Sieben dieser Mutationen waren synonyme Mutationen, was darauf hindeutet, dass ihre Rolle bei Krebs aus ihren Auswirkungen auf die alternative Spleißregulation und nicht aus ihrer Proteinkodierungsfunktion resultieren könnte. Um die Rolle von Mutationen, die das *RON*-Spleißen bei Krebs beeinflussen, weiter zu untersuchen, wurden als nächstes Patientendaten aus dem "The Cancer Genome Atlas" (TCGA) analysiert. Die stärksten Mutationseffekte wurden durch die Mutationen G370T und G297A, gefunden bei einem Patienten mit Kopf-Hals-Plattenzellkarzinom bzw. Schilddrüsenkarzinom, verursacht. Beide Mutationen führten zu einem erhöhten Grad an AE-Ausschlüssen, entweder aufgrund der Störung der 3′-Spleißstelle des AE (G297A) oder durch Änderung einer putativen Bindungsstelle im *RON*-Exon 11 (G370T). Da es sich bei G297A um eine Mutation der Spleißstelle handelt, hätte ihr starker Einfluss auf den erhöhten AE-Ausschluss bereits ohne die Informationen aus der Mutageneseanalyse abgeschätzt werden können. Im Gegensatz dazu ist die Mutation G370T eine nichtsynonyme Mutation, die auf Aminosäureebene ein vorzeitiges Stoppcodon („premature stop codon", PTC) im alternativen Exon einführt. Für PTC-haltige mRNAs wurde zuvor gezeigt, dass sie durch "nonsense-mediated mRNA decay" abgebaut werden (Nicholson et al., 2010). Unsere Arbeit zeigte jedoch, dass der AE-Ausschluss aufgrund der Mutation erhöht ist und somit das alternative Exon nicht in der reifen mRNA enthalten ist. Dies kehrt die physiologischen Folgen dieser Mutation um: Statt eines weniger funktionellen Rezeptors weisen die Zellen einen erhöhten RONΔ165 Pegel auf, was zuvor als Ursache für eine konstitutiven Rezeptoraktivierung gezeigt wurde (Collesi et al., 1996; Zhou et al., 2003). Zusammengenommen ermöglicht das Mutagenese-Screening die Bewertung der pathophysiologischen Relevanz von Krebsmutationen und deutet darauf hin, dass die Auswirkungen von nichtsynonymen Mutationen bei Krebs nicht nur durch Aminosäureveränderungen, sondern auch durch Spleißveränderungen entstehen können.

Um das Set von *trans*-regulierender Faktoren, die potenziell an der *RON*-Spleißregulation beteiligt sind, möglichst vollständig zu erfassen, wurde die "ATtRACT"-Datenbank für *in silico* Bindestellenvorhersagen auf die Sequenz des *RON*-Minigens angewandt (Giudice et al., 2016). Um die Ergebnisse des vorherigen Screens nutzen zu können und sich auf Positionen zu konzentrieren, die aktiv an der Spleißregulation beteiligt sind, wurden RBP-Motiv-Vorhersagen auf das Vorhandensein von mindestens 60% spleißwirksamen Positionen gefiltert (sogenannte spleißregulatorische Bindestellen, SRBS). Diese Analyse ergab 76 mutmaßliche Regulatoren des *RON*-Spleißens und bestätigte zuvor veröffentlichte Bindestellen von HNRNPH und SRSF1 (Ghigna et al., 2005; Lefave et al., 2011). Für die Auswahl der Kandidaten-RBPs wurde auf eine veröffentlichte RBP-„Knockdown" (KD)-Studie zurückgegriffen (Papasaikas et al., 2015), in der die KD-Effekte von ca. 250 RBPs auf verschiedene Spleißereignisse gemessen wurden, darunter auch auf das *RON*-Exon 11, gemessen in HeLa-Zellen. Eine Auswahl von putativen Regulatoren aus dieser Studie erfolgte anhand starker und in Replikaten konsistent auftretenden Auswirkungen auf das *RON*-Spleißen. Insgesamt wurde die durch KD induzierten Spleißveränderungen von 14 ausgewählten RBPs zwischen endogenem und *RON*-Minigen-Spleißen verglichen, und erfolgreiche KDs wurden durch die RT-qPCR Analyse bestätigt. Um neue Regulatoren des *RON*-Spleißens zu identifizieren, die in dem vorhergehenden RBP-KD-Screen nicht berücksichtigt wurden (Papasaikas et al., 2015), wurden *in vivo* RBP-Expression und *RON*-Spleißen in den Expressionsdaten aus TCGA analysiert. Die stärkste Assoziation zwischen RBP-Expressionsniveau und *RON*-Exon 11-Einbau wurde für *HNRNPH2* beobachtet. Zusammengenommen deuten diese Ergebnisse darauf hin, dass das *RON*-Spleißen durch mehrere Spleißaktivatoren und -repressoren weitgehend reguliert wird, wobei SRSF2 und HNRNPH der stärkste Aktivator bzw. Repressor sind. Darüberhinaus entdeckten wir zahlreiche bisher unbekannte *cis*-regulatorische Elemente sowohl in Introns als auch in Exons und fanden heraus, dass das heterogene nukleare Ribonukleoprotein H (HNRNPH) das alternative *RON*-Spleißen auf mehreren Ebenen sowohl in Zelllinien als auch in Krebs weitgehend reguliert. Weiterhin weist die große Anzahl der am Prozess beteiligten RBPs auf ein komplexes Spleißregulationsnetzwerk hin, das an der Steuerung des *RON*-Spleißens beteiligt ist.

Eingehender betrachtet sind hnRNP-Proteine funktionell vielfältig und regulieren unterschiedliche Schritte des RNA-Stoffwechsels, einschließlich prä-mRNA-Verarbeitung, mRNA-Transport und Translationskontrolle. Von diesen Proteinen weisen HNRNPH1 und HNRNPH2 96% Aminosäureidentität auf und werden im Folgenden als HNRNPH zusammengefasst. Das RNA-Bindemotiv von HNRNPH-Proteinen besteht aus Guanin-reichen Regionen ("G-runs") (Romano et al., 2002; Marcucci et al., 2007). Spleißänderungen aufgrund reduzierter Expression von HNRNPH korrelieren mit der Länge des "G-runs" in der Bindestelle (Xiao et al., 2009). RBPs aus der HNRNPH-Proteinfamilie können mit G-Quadruplex-Faltungen der RNA interagieren (Hacht et al., 2014; Liu et al., 2017). Desweiteren wurden HNRNPH-assoziierte Spleißdefekte mit mehreren Krankheiten einschließlich Krebs in Verbindung gebracht (Rauch et al., 2010; Lefave et al., 2011). Insgesamt fanden wir heraus, dass das *RON*-Minigen 22 SRBS für HNRNPH enthält, die sich in allen Regionen des Minigens befinden. Diese kann man in fünf verschiedene Cluster zusammenfassen, mit drei bis fünf SRBS. Um die positionspezifischen Interaktionen zwischen HNRNPH und dem Minigen zu untersuchen, wurde als Methode Einzel-Nukleotid auflösende UV-Kreuzvernetzung und Immunopräzipitation („individual-nucleotide resolution UV crosslinking and immunoprecipitation", iCLIP) in HEK293T-Zellen durchgeführt. Veränderte Spleißergebnisse resultierten aus Unterbrechungen bzw. Verlängerungen der "G-runs", was als reduzierte bzw. verbesserte HNRNPH-Bindung interpretiert werden konnte. Wie bereits erwähnt, hing die Art der durch HNRNPH vermittelten Spleißänderungen von der Position der HNRNPH-Bindestelle innerhalb des Minigenes ab. HNRNPH-Bindung im alternativen Exon vermittelte Spleißunterdrückung, während die Bindung von HNRNPH im stromaufwärts gelegenen Intron Spleißen förderte. Mutationen in Cluster 1 und 5, die sich in den vor- und nachgelagerten konstitutiven Exons befinden, reduzierten die Intronretention und in Cluster 5 wurde dies von einem erhöhten AE-Ausschluss begleitet. Eine solche kontextabhängige Regulierung wurde bereits für HNRNPH und andere Spleißregulatoren beschrieben (Xiao et al., 2009; Katz et al., 2010). Die stärksten Spleißwirkungen wurden durch Mutationen im HNRNPH SRSBS Cluster 3, das sich im alternativen Exon befindet, vermittelt. Auffallend ist, dass *HNRNPH* KD ein ähnliches Spleißmuster induzierte wie Mutationen im Cluster 3, was die Bedeutung des HNRNPH SRBS Clusters 3 in der HNRNPH-vermittelten Regulation des *RON*-Spleißens unterstreicht. Damit lässt sich zusammenfassend sagen, dass HNRNPH gleichzeitig als

Aktivator und Repressor des *RON*-Spleißens fungiert und somit eine komplexe Regulation verschiedener Isoformen über mehrere Bindungsstellen in der Minigeneregion vermittelt.

Um die funktionell relevantesten Stellen der HNRNPH-Regulierung zu identifizieren, wurde das Spleißen der Minigen-Bibliothek unter HNRNPH KD-Bedingungen analysiert. Es wurde erwartet, dass Mutationen mit Einfluss auf die HNRNPH-Bindung eine positive oder negative Synergie mit dem HNRNPH KD zeigen würden, d.h. die kombinierten Effekte von Mutation und Knockdown sich als größer oder kleiner zeigen würden als die einzelnen Effekte. Synergistische Effekte akkumulierten innerhalb des alternativen Exons, bemerkenswerterweise mit einem Prozentsatz von 93% an synergistischen Positionen innerhalb des SRBS-Clusters 3. Dementsprechend stellt das alternative Exon die Schlüsselregion für die HNRNPH-vermittelte Spleißregulation von *RON* dar. So ermöglichte ein KD Experiment mit darauffolgender RNA-Sequenzierung der mutierten Minigen-Bibliothek die anschließende Berechnung der Synergie zwischen Mutationen und HNRNPH KD, und damit eine quantitative Bewertung der funktionellen Bedeutung der Bindestellen. Die Beobachtung, dass einzelne Punktmutationen im Cluster 3 ausreichen, um die HNRNPH-vermittelte Spleißunterdrückung fast vollständig aufzuheben, deutete darauf hin, dass die HNRNPH-Bindungsereignisse voneinander abhängig sind und Mutationen, die die Bindung von HNRNPH beeinträchtigen, auch die HNRNPH-Bindung bei benachbarten SRBS beeinträchtigen. Um diese Idee zu testen, wurden iCLIP-Experimente im Rahmen von Minigenen mit Punktmutationen im Cluster 3 wiederholt. Im Einklang mit dem Modell einer kooperativen Bindung war die Reduktion des iCLIP-Signals nicht auf den Ort der Punktmutation beschränkt, sondern erstreckte sich über das gesamte alternative Exon. Eine kooperative Regulierung des *RON*-Spleißens durch HNRNPH würde bedeuten, dass die Menge an *RON*-Isoformen empfindlich auf Veränderungen des HNRNPH-Proteinspiegels reagierte, was in einer steilen, sigmoiden Dosis-Wirkungskurve resultieren würde. Tatsächlich wurde eine schalterartige Spleißreaktion des *RON*-Exons 11 aus dem Minigen sowie dem endogenen *RON*-Gen für die Veränderung der HNRNPH-Konzentrationen beobachtet, die mit der Hill-Gleichung quantitativ ausgewertet werden kann. Darüber hinaus zeigte *HNRNPH2* die steilste Regressionssteigung unter den 190 RBPs, die auf Expressionskorrelation in den TCGA-Daten getestet wurden, was mit der kooperativen Regulierung auch *in vivo* übereinstimmte. Zusammengenommen fungierte HNRNPH als Spleißschalter des *RON*-Exons 11 durch kooperative Bindung, was zu großen Spleißänderungen führt, die durch kleine Veränderungen des HNRNPH-Proteingehalts verursacht werden. So wurden durch iCLIP- und Synergieanalysen zwischen Mutationen und HNRNPH-Knockdown-Daten die relevantesten HNRNPH-Bindestellen im gesamten *RON*-Minigen ermittelt. Schließlich wurde gezeigt, dass die kooperative HNRNPH-Bindung einen Spleißschalter des *RON*-Exons 11 ermöglicht.

Zusammenfassend haben wir in diesem PhD-Projekt ein Hochdurchsatz-Screening auf Basis zufälliger Mutagenese etabliert, um die regulatorische Landschaft des Exon 11-Spleißens im *RON*-Minigen umfassend zu charakterisieren. Da das Minigene nicht nur *RON*-Exon 11, sondern auch die benachbarten Introns sowie die flankierenden, konstitutiven Exons enthält, ermöglichte das Screening die Bewertung des Beitrags der Regionen um das alternative Exon und erweiterte damit die bisherigen Studien. Klinische Daten halfen, die ermittelten Mutationseffekte des Screeningansatzes mit dem *RON*-Spleißen bei Krebspatienten mit den gleichen Mutationen zu vergleichen. Dementsprechend werden die Ergebnisse bei der Beurteilung von Mutationseffekten helfen, die für die klinische Diagnostik potenziell relevant sind. Neben der umfassenden Charakterisierung von *cis*-regulatorischen Elementen identifizierten Knockdown-Experimente die an der *RON*-Exon 11-Regulation beteiligten Spleißfaktoren. Weiterführend lässt sich feststellen, dass dieser neuartige Screening-Ansatz ein Werkzeug darstellt, um die Wechselwirkung von RNA-Sequenzvarianten mit *trans*-regulierenden Regulatoren zu untersuchen. Das ermöglicht Einblicke in alternative Spleißregulationsmechanismen und die Relevanz von Mutationen bei menschlichen Erkrankungen. Weiterhin bieten die Ergebnisse eine einzigartige Sichtweise auf die Komplexität der Spleißregulierung eines alternativen Exons.

Abstract

Alternative splicing (AS) is a co- or post-transcriptional process by which one gene gives rise to multiple isoforms. This 'split and combine' step multiplies eukaryotic proteome diversity several fold and is implicated in several diseases given its pervasive impact. Control of alternative splicing is brought about by *cis*-regulatory elements, such as RNA sequence and structure, which recruit *trans*-acting RNA-binding proteins (RBPs). Although several of these interactions are already described in detail, we lack a comprehensive understanding of the regulatory code that underlies a splicing decision.

Here, we have established a high-throughput screen to comprehensively identify and characterise *cis*-regulatory elements that control a specific splicing decision. A cancer-relevant splicing event in proto-oncogene *RON* was picked as a minigene prototype for initialising the screening approach. Then, we transfected a library of thousands of randomly mutagenised minigene variants as a pool into human cells, and subsequently quantified the spliced isoforms by RNA sequencing. Importantly, we used a barcode sequence to tag the minigene variants and thereby linked mutations to their corresponding spliced products. By using a linear regression-based modelling approach, we were able to determine the effects of single mutations on *RON* AS. In total, more than 700 mutations were found to significantly affect the splicing regulation of the *RON* alternative exon. In addition, mutation effects quantified from the screening approach correlate with *RON*

alternative splicing in cancer patients. We discovered numerous previously unknown *cis*-regulatory elements in both introns and exons, and found that the RBP heterogeneous nuclear ribonucleoprotein H (HNRNPH) extensively regulates *RON* AS at multiple levels in both cell lines and cancer. Furthermore, the large number of RBPs involved in the process, point to a complex splicing regulatory network involved in the control of *RON* splicing. iCLIP and synergy analysis between mutations and HNRNPH knockdown data pinpointed the most relevant HNRNPH binding sites across *RON*. Finally, cooperative HNRNPH binding was shown to mediate a splicing switch of *RON* alternative exon. In summary, our results provide an unprecedented view on the complexity of splicing regulation of an alternative exon. The novel screening approach introduces a tool to study the relationship of RNA sequence variants along with *trans*-acting regulators to their impact on the splicing outcome, offering insights on alternative splicing regulation and the relevance of mutations in human disease.

Preface

The *RON* mutagenesis screening project (Braun et al., 2018) constituted of three sections; experimental, bioinformatics and mathematical modelling and was supervised by Dr. Julian König (IMB, Mainz), Dr. Kathi Zarnack (BMLS, Frankfurt), Dr. Stefanie Ebersberger (IMB, Mainz) and Dr. Stefan Legewie (IMB, Mainz), respectively. In my Phd study, I performed most of the bioinformatics analyses of the project. In this thesis, I have presented the topics that were exclusively analysed by myself, along with analyses performed by the other collaborators to present a complete and cohesive picture of the project for better comprehension. The results of the project have been published in the following article -

Braun, S.; Enculescu, M.; Setty, S. T.; Cortés-López, M.; Almeida, B. P. de; Sutandy, F. X. R.; Schulz, L.; Busch, A.; Seiler, M.; Ebersberger, S.; Barbosa-Morais, N. L.; Legewie, S.; König, J.; Zarnack, K., 2018: Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. *Nature Communications.*, 9, 3315.

These collaborators and their corresponding work in brackets are as follows -

Simon Braun (*RON* high throughput mutagenesis screen and library generation and prep.), F. X. Reymond Sutandy (iCLIP experiments), Mariela Cortés-López (RBP site annotation and analyses), Laura Schulz (RT-PCR validation experiments), Dr. Markus Seiler (Annotation of RNA G-quadruplex sequences), Dr. Anke Busch (iCLIP and RNA-seq data processing and splice isoform quantification), Bernardo P. de Almeida and Dr. Nuno L.

Barbosa-Morais (TCGA and GTEx analyses; iMM, Lisbon). The project was conceived by Dr. Julian König and experimental analyses were performed under his supervision. The bioinformatics analyses were performed under the supervision of Dr. Katharina Zarnack with additional supervision by Dr. Stefanie Ebersberger. Dr. Mihaela Enculescu and Dr. Stefan Legewie designed and performed the mathematical modelling approach.

1

Introduction

Pre-mRNA splicing (spliceosomal catalysis)

The messenger RNA (mRNA) is an essential macromolecule, involved in the tissue-specific process of transcription in cells. Eukaryotic genes are transcribed into precursor mRNA or pre-mRNA, which are made up of discrete units called exons (coding regions) and introns (non-coding regions). The process of removal of introns from the pre-mRNA and joining of exons in order to form a functional mRNA transcript is called splicing. Splicing is realised by a multi-step catalysing reaction involving the spliceosome, a complex made up of huge protein assemblies which interacts with other splicing factors to give rise to the mature mRNA transcript (Wahl et al., 2009). Such a multi-step reaction involving many proteins allows opportunities for increased control and modulation, such as the choice to include exons selectively via a process known as alternative splicing (Johnson et al., 2012). This process forms the basis for tissue and proteome diversity and occurs in ~95% of multi-exonic human genes (Pan et al., 2008). Furthermore, structural images from yeast cryo-electron microscopy showed that the underlying mechanism of spliceosomal catalysis is conserved across eukaryotes (Plaschka et al., 2019) suggesting the universality of the process among higher organisms.

The spliceosomal catalytic reactions are mediated by the spliceosome complex, which is an intricate biological machine consisting of a higher order assembly of 5 small nuclear RNAs (snRNAs) and ~200 related proteins (Will et al., 2011; Fica et al., 2013). During or immediately after transcription, the spliceosomal complex mediates the catalytic excision of introns and stitching together of exons (splicing) in the pre-mRNA in a series of steps as detailed below.

The assembly of the spliceosome is initiated with the recognition of the signals at the splice sites, the branch point and the polypyrimidine tract (**Figure 1.1**, top left corner). Firstly, the U1 small nuclear RNP (snRNP) binds to the 5′ splice site and the U2 auxiliary factor (U2AF) binds to the polypyrimidine tract along with the binding of the splicing factor 1 (SF1) to the branch point, thus forming the complex E. Subsequent recruitment of the U2 snRNP is initiated at the 3′ splice site region. The resulting interaction between U1 and U2 snRNPs depends on the helicases PRP5 and UAP56 and together forms the pre-spliceosomal complex A.

Next, the pre-catalytic spliceosome complex (complex B) is formed by the recruitment of U4/U6-U5 tri-snRNP, and is followed by the release of U1 and U4 snRNPs. This necessitates a remodelling step (B activated complex) and activates the catalytic centre of the spliceosome (complex B*), thus enabling the subsequent splicing reactions. In detail, the splicing catalysis consists of two transesterification steps; starting with, the 2′-OH of the

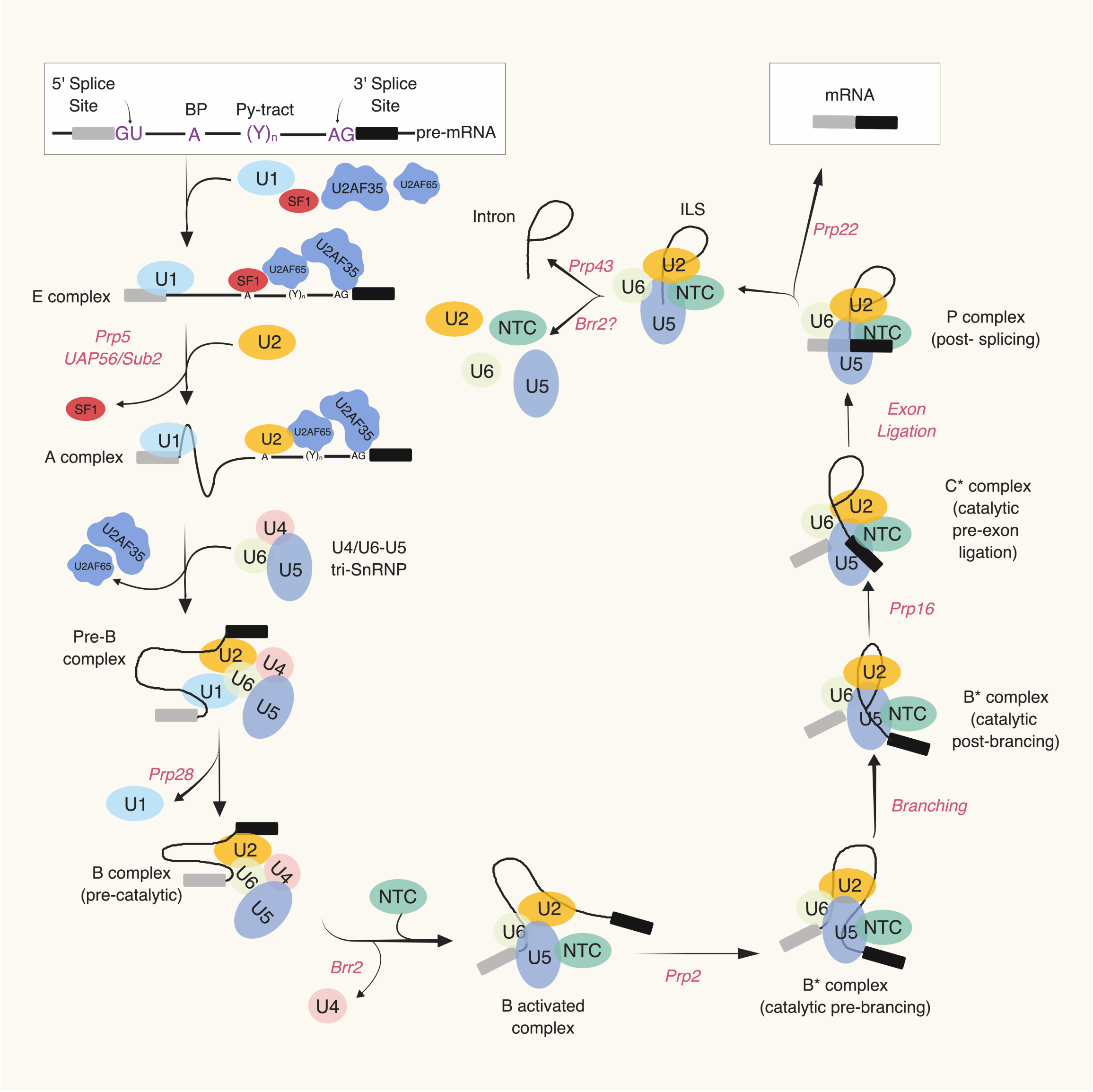


Figure 1.1: The spliceosome mediates alternative splicing in a multi-step cycle. Core-splicing signals within introns of pre-mRNAs comprise the 5′ splice site (5′ SS), the branch point (BP), the polypyrimidine-tract (Py-tract), and the 3′ splice site (3′ SS). Early splicing factors U1 snRNP, SF1 and the U2AF heterodimer, consisting of U2AF65 and U2AF35, target these signals. During the splicing reaction, the splicing factors U2 snRNP and U4/U6-U5 tri-snRNP are required to excise the intron, while numerous additional factors transiently interact with the pre-mRNA during the different steps of splicing catalysis. Boxes represent exons while the connecting introns are depicted with lines. Adapted from (Fica et al., 2017).

adenine, which is a part of the branch point consensus signal, attacking the phosphate of the guanine at the 5′ consensus splice site. This results in a cleaved product between the exon and the guanine in the intron and in a free 3′-OH at the upstream exon. Additionally, the intron lariat is released and is attached to the downstream exon. In the second step, the free 3′-OH of the upstream exon reacts with the phosphate at the 5′ end of the downstream exon to covalently join both exons and release the intron lariat via the C* complex and then forms the post-spliceosomal P complex (Wahl et al., 2009; Fica et al., 2017; Shi, 2017). Finally, the ligated exons and intron lariat are released from the spliceosome by the PRP22 helicase (Will et al., 2011), leading to the formation of the mature mRNA transcript.

Alternative splicing

The selective skipping or inclusion of exons leads to increased mRNA diversity in a process called alternative splicing (AS). Alternative splicing is quite prevalent, given that at least 95% of human genes are alternatively spliced (Kornblihtt et al., 2013; Manning et al., 2017) resulting in an increased phenotypic diversity from a smaller set of genes (Yang et al., 2016). Alternative splicing determines homeostasis and development of cell, tissue and organ fate (Baralle et al., 2017) and may result in more than 1,000 different isoforms from a single human gene (Treutlein et al., 2014). Alternative splicing is also a hallmark of evolutionary complexity as more splicing is seen in higher vertebrates (Barbosa-Morais et al., 2012; Chen et al., 2014). Such extensive alternative splicing has been observed in higher organisms as a result of a higher presence of weak splice sites (Lee et al., 2015), which vary from the consensus strong splice sites. Recently, studies on AS have increasingly shown its essential role in cellular differentiation, stress responses, disease development and cancer cell survival (Coltri et al., 2019). There are five different events of alternative splicing observed, which are based on the choice of splice sites, which in turn involve splice site strength and context (**Figure 1.2 and 1.3**).

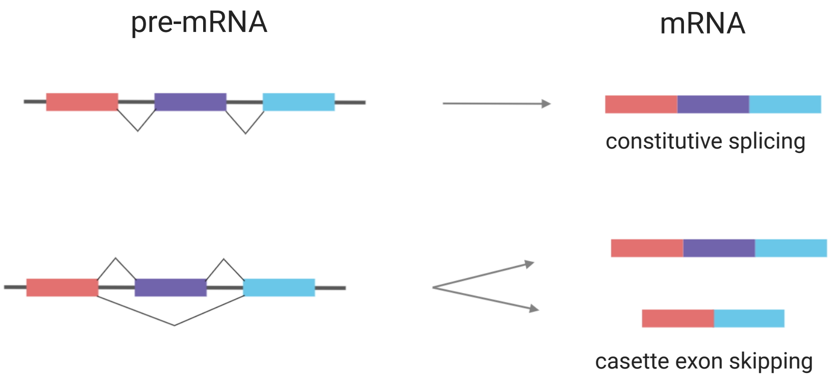


Figure 1.2: Alternative splicing and canonical isoforms. Canonical splicing isoforms include exon inclusion and skipping isoforms. Exons are shown as boxes, while the solid lines connecting them are introns. Other solid lines indicate the splice junctions between exons. Adapted from (Wang et al., 2015).

Cassette exons, can either be present (exon inclusion) or absent (exon skipping) in the mature mRNA (**Figure** 1.2). Unspliced introns (intron retention or IR) and alternative splice site usage (alternative 3' and alternative 5' splice site) result in altering the exonic sequences (Kornblihtt et al., 2013) (**Figure** 1.3). While exon skipping is the most common alternative splicing event (Wang et al., 2008), IR has recently been highlighted as an important mechanism to regulate gene expression and was shown to occur in more than half of all human introns (Braunschweig et al., 2014; Jacob et al., 2017). In some cases, one exon among two or more possible cassette exons is included in the mRNA resulting in exon inclusion in a mutually exclusive manner (**Figure** 1.3). Furthermore, previous studies showed that most genes express multiple isoforms simultaneously ('non-minimalistic pattern') (Djebali et al., 2012). Recent studies on single nucleotide variants (SNVs) of 1,812 tumour samples, revealed that exon skipping and intron retention were among the most frequently observed alternative splicing events in cancer (Jung et al., 2015).

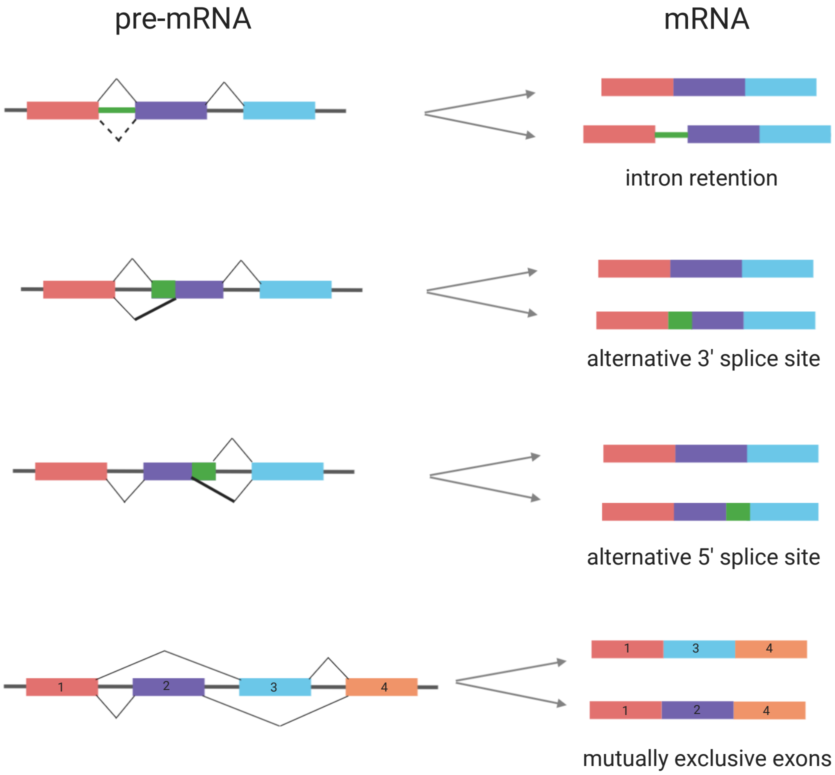


Figure 1.3: Alternative splicing and non-canonical isoforms. Other types of alternative splicing result in different mRNA isoforms. Exons are shown as boxes, while the solid lines connecting them are introns. Other solid lines indicate the splice junctions between exons. Adapted from (Wang et al., 2015).

Regulation of alternative splicing

The regulation of alternative splicing is a multi-layered fine-tuned mechanism where the information content in RNA sequences is decoded by the cognate RBP proteins. The factors that determine regulation of AS are detailed below.

1. Splice site strength

Constitutive splicing occurs in strong splice sites, which harbour conserved consensus sequences at 5′ (CAG | GUAAGU) and 3′ (NYAG | G) splice sites in mammals (Mount, 1982; Black, 2003). If the sequences at the splice sites match these consensus sequences, then it is immediately recognised by the early spliceosome and splicing occurs (Garg et al., 2007). However, many additional alternative splice sites of weak or intermediate strength may exist along with stronger sites, which may complicate the splicing decision. In order to ensure proper splicing, the splice sites must be selected in a correct manner. This is accomplished by the recruitment of RNA binding proteins (RBPs) in a context-dependent manner. In particular, this competitive selection of splice sites among stronger constitutive and weaker sites decides alternative splicing (Baralle et al., 2017).

2. *Cis*-regulatory elements

Control of alternative splicing is achieved through *cis*-regulatory elements, such as RNA sequence and structure, which recruit *trans*-acting RBPs. The RBPs interact with pre-mRNA sequence elements, called Exonic and Intronic Splicing Enhancers or Silencers (ESE, ESS, ISE and ISS; **Figure** 1.4). These sequences enhance or inhibit the usage of weak alternative splice sites in conjunction with the spliceosome (Kornblihtt et al., 2013; Wang et al., 2015). Moreover, *cis*-regulatory elements can be highly context-dependent (Goren

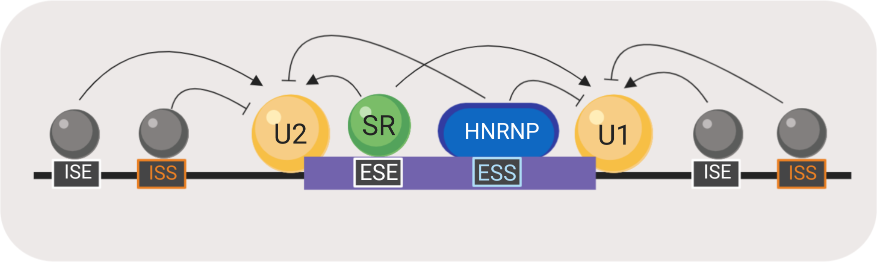


Figure 1.4: Regulation of alternative splicing is a multi-layered process involving splicing enhancers and silencers. SR and HNRNP proteins interact with splicing enhancers and silencers either at introns or exons to mediate splicing regulation, in conjunction with splicing proteins. ESE-exon splicing enhancer, ESS-exon splicing silencer, ISE-intron splicing enhancer and ISS-intron splicing silencer. Grey circles indicate various splicing factors. Adapted from (Bartys et al., 2019).

et al., 2006) and splicing events are controlled by combinatorial interplay of multiple *cis*-regulatory elements (Nasrin et al., 2014; Qian et al., 2014). Moreover, it was shown that variation in *cis*-regulatory elements majorly drives splicing evolution (Hsiao et al., 2016).

3. *Trans*-acting factors

Trans-acting factors are splicing-related proteins that target *cis*-regulatory elements present in the pre-mRNA (Fu et al., 2014). RBPs work in conjunction with the spliceosome to mediate exon definition (De Conti et al., 2013) and inhibit or activate cognate splice site usage. Moreover, these proteins are ubiquitously expressed across normal human tissues (Gerstberger et al., 2014), are evolutionarily conserved (Anantharaman et al., 2002) and are known to be involved in all aspects of RNA processing (Glisovic et al., 2008). The group of *trans*-acting splicing factors include Serine-Arginine repeat (SR) splicing factors, heterogeneous nuclear ribonucleoproteins (HNRNPs) and other splicing factors (Cordin et al., 2013) (**Figure 1.4**). HNRNPs are a highly conserved family of RNA-binding proteins which bind to nascent pre-mRNA and are involved in diverse cellular functions such as localization, maturation, translation among other roles. SR proteins are essential splicing factors that regulate both constitutive and alternative splicing. *Cis*-regulatory elements are sometimes bound by multiple RNA binding domains of a single *trans*-acting factor and hence require complex interactions to determine better specificity of RBPs (Maris et al., 2005; Auweter et al., 2006; Valverde et al., 2008; Dominguez et al., 2010). The *trans*-acting splicing factors usually bind to 4-7 bp of target RNA sequence (Daubner et al., 2013), but even longer *cis*-regulatory elements have been shown.

4. Positional regulation

The activity of *trans*-acting factors depends on their binding site position within the pre-mRNA (Erkelenz et al., 2013). As a general rule, HNRNP proteins binding at the exon repress splicing (Rothrock et al., 2005; Mauger et al., 2008), but display opposite behaviour and activate splicing, when they bind to the introns (Hui et al., 2003; Xiao et al., 2009). In the case of bound SR proteins, a role reversal takes place and splicing is activated at the exons (Cartegni et al., 2002; Shen et al., 2004) and repressed at introns (Ibrahim et al., 2005; Buratti et al., 2007; Shen et al., 2012). Hence both HNRNP and SR proteins perform a position-dependent regulatory function in opposing directions as splicing repressors or activators, similar to other *trans*-acting factors such as NOVA1 (Licatalosi et al., 2008) or RBFOX1 (Sun et al., 2012).

5. RNA structure

In addition to sequence and proteins, alternative splicing regulation is mediated through pre-mRNA secondary structure (Oberstrass et al., 2005; Warf et al., 2010). For instance, RBPs such as MBNL1, PTBP1, SRSF1 and HNRNPA1 can modulate RNA secondary structures and thus mediate exon skipping or inclusion (Blanchette et al., 1999; Oberstrass et al., 2005; Pascual et al., 2006; Liu et al., 2010) (**Figure 1.5**). The arrangement and number of RBP binding sites and a structure-dependent competitive behaviour for instance; hairpin

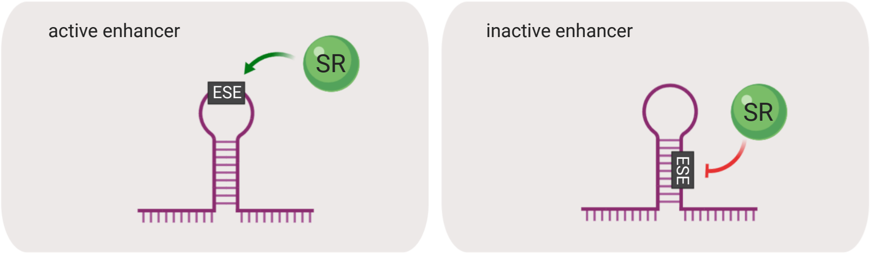


Figure 1.5: Pre-mRNA secondary structures mediate splicing regulation. The influence of a representative exonic hairpin structure on the enhancer activity of SR proteins is shown here. ESE - Exonic splicing enhancer. Adapted from (Bartys et al., 2019).

or stem loop presence, also determines splicing regulation (Taylor et al., 2018). Furthermore, RNA G-quadruplexes represents a secondary structure element that is shown to be frequently involved in splicing regulation (Conlon et al., 2016; Huang et al., 2017; Weldon et al., 2018). These secondary structures are formed by multiple stacks of four guanines, organised into a planar arrangement via Hoogsteen hydrogen bonding (Cammas et al., 2017). Moreover, multiple RBPs were shown to interact with RNA G-quadruplexes (Hacht et al., 2014; Liu et al., 2017) including HNRNPH proteins (Decorsière et al., 2011; Fisette et al., 2012; Conlon et al., 2016). G-quadruplexes have been shown to be widely prevalent in the human transcriptome in addition to being evolutionarily conserved (Kwok et al., 2016). However, *in-vitro* and *in-vivo* data show contradictory evidences, limiting the interpretation of the role of RNA G-quadruplexes (Biffi et al., 2014; Laguerre et al., 2015; Guo et al., 2016 ; Kwok et al., 2016), and thus require more functional studies. Furthermore, there are studies showing many more determinants involving chromatin modifications, transcriptional kinetics, developmental stages and cell type specificities, and binding specificities like allelle specific binding that may modulate alternative splicing.

Mechanisms of RBP function in alternative splicing regulation

Core splicing factors such as U1 interact with *trans*-acting proteins such as SR and HNRNP proteins, in conjunction with other transcript-specific splicing regulators in order to modulate splicing regulation. Both SR and HNRNP proteins assist in the splice site recognition and can recruit the spliceosome (Howard et al., 2015). Importantly, the interplay between SR and HNRNP proteins and recruitment and stabilisation of the splicing machinery defines the exon boundaries ('exon definition') before splicing activation can occur (Graveley, 2000; Caputi et al., 2002) (**Figure 1.6** panel a). Additionally, a competition between splicing repressors and either SR or HNRNP proteins (Zhu et al., 2001; Paradis et al., 2007) can activate splicing (**Figure 1.6** panel b). The direct competition of HNRNP proteins with core splicing factor U2AF (Heiner et al., 2010; Zarnack et al., 2013) can repress splicing (**Figure 1.7** panel a).

By simultaneously interacting with U1 snRNP and U2AF splicing proteins, SR proteins position the 5′ and 3′ splice site close to each other and activate splicing (Wu et al., 1993; Graveley, 2000) (**Figure 1.6** panel c). However, HNRNP proteins enable splicing repression by looping and hiding the exons (Martinez-Contreras et al., 2006; König et al., 2010) (**Figure 1.7** panel b). Moreover, cooperative binding by polymeric HNRNP proteins across the gene shows contrasting behaviour as it can both promote exon skipping (Okunola et al., 2009) and exon inclusion depending on the context, although the mechanism remains undiscovered (Gueroussov et al., 2017) (**Figure 1.8**).

The heterogeneous nuclear ribonucleoprotein H (HNRNPH)

Many studies have elucidated the important role of HNRNPH proteins in driving splicing regulation (Martinez-Contreras et al., 2007; Chaudhury et al., 2010; Katz et al., 2010). Based on structural differences, the HNRNPH family consists of five members;

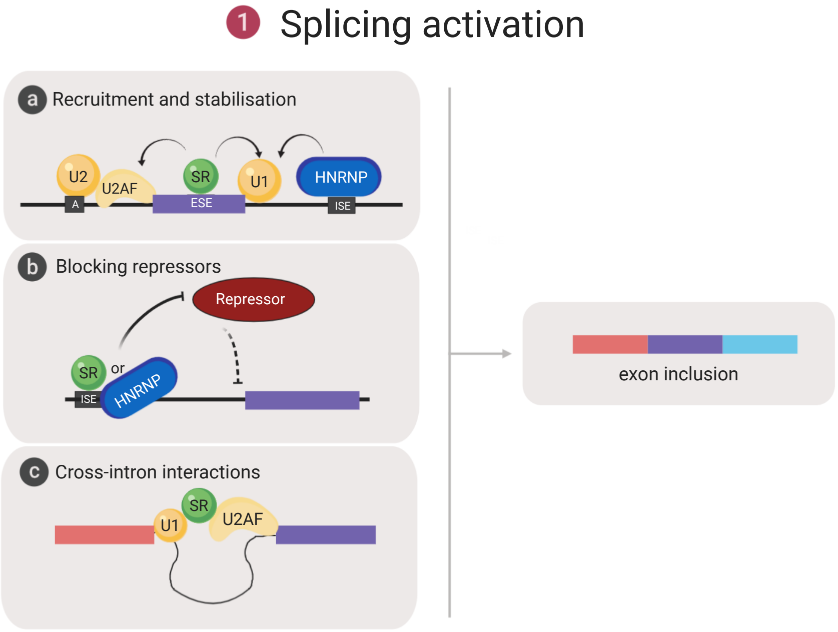


Figure 1.6: SR and HNRNP proteins mediate splicing activation. (a) Recruitment of core splicing factors by SR and HNRNP proteins mediate splicing regulation. (b) Competition of SR or HNRNP proteins with splicing repressors activates exon inclusion. (c) SR proteins mediate cross-intron interactions and activate splicing. Exons and introns are depicted as boxes and lines, respectively. Adapted from (Graveley, 2000; Martinez-Contreras et al., 2006; Wachter et al., 2012) and from Simon Braun.

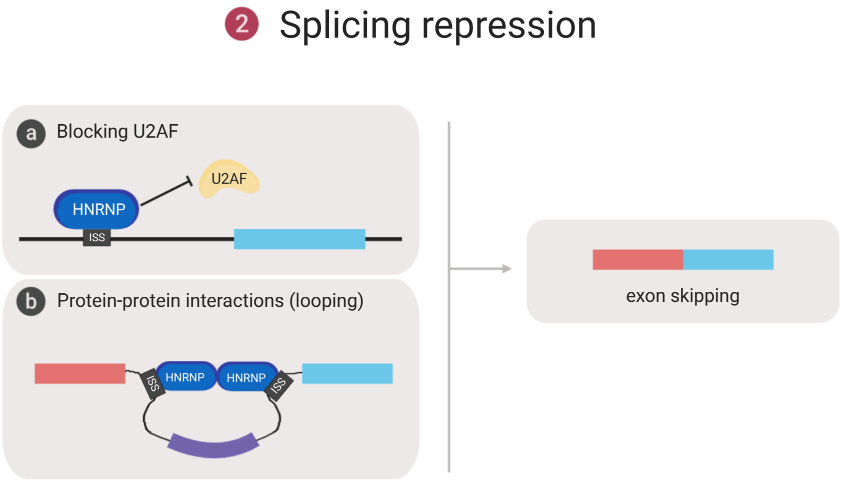


Figure 1.7: SR and HNRNP proteins mediate splicing repression. (a) Competition of HNRNP proteins with core splicing factors represses exon inclusion. (b) HNRNP protein mediated looping of pre-mRNA at ISS sites represses exon inclusion. Exons and introns are depicted as boxes and lines, respectively. Adapted from (Graveley, 2000; Martinez-Contreras et al., 2006; Wachter et al., 2012) and from Simon Braun.

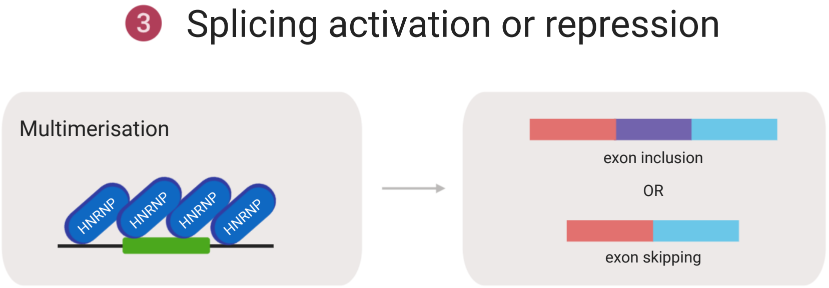


Figure 1.8: Multimerised HNRNP proteins mediate splicing activation and repression. Multimerisation of HNRNP proteins across pre-mRNA can both activate and repress exon inclusion. Exons and introns are depicted as boxes and lines, respectively. Adapted from (Gueroussov et al., 2017).

HNRNPH1, HNRNPH2, and HNRNPF, HNRNPH3a and HNRNPH3b. HNRNPH proteins contain RRMs (RNA recognition motifs)-related domains that are loosely similar to the consensus RRMs and are called quasi-RRMs (qRRMs) and regions rich in glycine, tyrosine and arginine (GYR) or glycine- and tyrosine-rich (GY) (Martinez-Contreras et al., 2007). In our study, the protein HNRNPH is referred to both the proteins HNRNPH1 and HNRNPH2 as they share 96% amino acid identity (Martinez-Contreras et al., 2007). While HNRNPF is similar to HNRNPH on the amino acid level, HNRNPH3a and HNRNPH3b are less conserved and lack a quasi RNA recognition motif (qRRM) (Nazim et al., 2017). **Figure** 1.9 shows the conservation between the HNRNPH proteins.

HNRNPH binds to G-rich sequences or G-runs and consequently, mediates splicing regulation via its qRRM domains (Ding et al., 1999; Dominguez et al., 2010). Depending on the location of the binding motif in the different regions of the gene, HNRNPH either activates or represses exon inclusion, respectively (Xiao et al., 2009; Katz et al., 2010). Accordingly, the length of these G-runs also becomes important in mediating the regulatory activity (Romano et al., 2002; Marcucci et al., 2007; Masuda et al., 2008; Rahman et al., 2015; Uren et al., 2016; Nazim et al., 2017). Given the importance of HNRNPH in crucial cellular processes, several diseases are linked to its deregulation (Rauch et al., 2010; Lefave et al., 2011; Stark et al., 2011). In particular, deregulation mediated by RBP HNRNPH leads to enhanced metastatic response in gliomas (Lefave et al., 2011).

Alternative splicing in disease

Alternative splicing is crucial in the generation of biological complexity, and its deregulation leads to several human diseases (Daguenet et al., 2015). Deregulation of splicing has been attributed to mutations, which may result in perturbing a *cis*-regulatory element. Consequently, exon inclusion or exon skipping may lead to irregular protein activity. For instance, mutations that alter a splice site, result in loss of function of that site and may

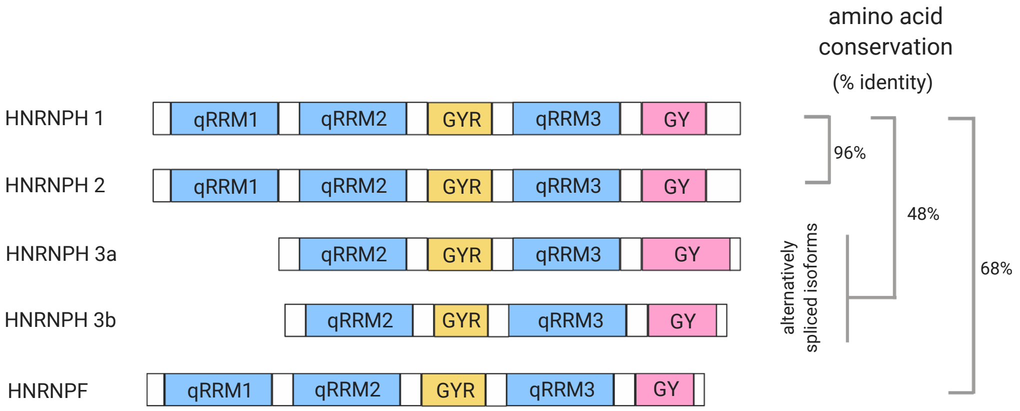


Figure 1.9: HNRNPH proteins show conservation of domains. The quasi-RNA recognition motifs (qRRM) mediate RNA binding, while the glycine-tyrosine-arginine-rich (GYR) domains control nuclear localization. A glycine-tyrosine-rich (GY) domain at the C-terminus mediates protein-protein interactions. Adapted from (Nazim et al., 2017).

lead to changing the splice location. This can cause insertion or deletion of unwanted sequences and in turn, longer or shorter exons in the mature mRNA. Although, many splicing errors are safeguarded by nonsense-mediated mRNA decay (NMD), a quality control step in cells (Danckwardt et al., 2002; Hug et al., 2016), several splicing-related diseases still prevail. It is estimated that 10% of disease-causing exonic mutations and that up to 1/3 of all disease-associated alleles alter splicing (Havens et al., 2013; Soemedi et al., 2017).

Given the prevalence of splicing deregulation especially in cancer, the study of the role of alternative splicing in cancers has been a primary concern. Splicing misregulation occurs in transcripts coding for genes associated with all hallmarks of cancer including invasion, metastasis and angiogenesis (Amin et al., 2011; Bonomi et al., 2013; Oltean et al., 2014; Sveen et al., 2016). Moreover, a 30% rise in alternative splicing events is observed in tumour samples when compared to normal samples. Many tumours consist of thousands of alternative splicing events not detectable in normal samples; i.e. on average, ~930 new exon-exon junctions (‘neojunctions’) were observed in tumours and not typically found in GTEx normal samples. Strikingly, tumours with aberrant splicing isoforms tend to exhibit decreased expression levels of immune system-related genes and increased expression levels of cell-cycle marker genes. These tumour-specific splicing events create a novel immune response and could be exploited in immunotherapy; for instance in personalised tumour vaccines (Mo et al., 2017; Kahles et al., 2018). Furthermore, RBP expression is a tightly controlled process under normal circumstances, and any significant changes may affect a plethora of transcripts and thus result in an imbalance in cellular homeostasis (Mittal et al., 2009). Indeed, altered expression of some RBPs has been linked to diseases including cancer (Musunuru, 2003; Lukong et al., 2008; Castello et al., 2013; Correa et al., 2016; Sebestyén et al., 2016). Notably, many splicing factors moonlight in cancer either as tumour suppressors or proto-oncogenes and drastically alter expression levels of genes involved in cancer at a system-wide scale (Dvinge et al., 2016; Seiler et al., 2018). A study revealed that RBPs have an average of approximately 3 mutations per Mb across 26 cancer types and 281 RBPs to be enriched for mutations in at least one cancer type (Neelamraju et al., 2018). Moreover, cancer related RBPs are predominantly downregulated in tumour when compared to normal tissues (Wang et al., 2018), however Zhang et al. (2018) showed that several RBPs are upregulated in cancers. In the 16 TCGA cancer types studied, 109 RBPs were consistently upregulated while only 41 RBPs were consistently downregulated (Zhang et al., 2018). Furthermore, expression of many HNRNP RBPs is altered in neurodegenerative diseases and cancer (Boukakis et al., 2010). The importance of HNRNP proteins in cancer is underlined by the fact that the RBP levels correlate to prognosis of patients (Karni et al., 2007; Huelga et al., 2012).

Recepteur d'origine nantais (RON) receptor tyrosine kinase

Receptor tyrosine kinases (RTKs) discovered in 1960s, are transmembrane cell-surface receptors that bind to ligands, which promotes receptor dimerisation and stimulation of the tyrosine kinase activity (phosphorylation) and subsequent signal transduction (Schlessinger, 2000, 2014). RTKs are key regulators of homeostasis and normal development, but also have a critical role in the development and progression of many types of cancer. One of the most well studied example in physiological conditions is insulin binding a multimeric RTK complex (Meyts et al., 1973). Proteins in the receptor tyrosine kinase family consist of the following domains -

1. extracellular ligand-binding domain.
2. intracellular protein tyrosine kinase domain.
3. transmembrane helix.

Recepteur d'origine nantais (RON) is a receptor tyrosine kinase (Ronsin et al., 1993) encoded by the proto-oncogene *MST1R* (also referred to as *RON*) and is a member of the MET proto-oncogene family (Benvenuti et al., 2007). *RON* undergoes transcription in epithelial cells, however it is lowly expressed (Gaudino et al., 1994, 1995). In contrast to highly expressed receptor tyrosine kinases which express 70,000 molecules per keratinocyte, only 600 – 1,000 RON receptor proteins were estimated to be present per cell (Roque et al., 1992; Wang et al., 1996). In physiological conditions, RON exists as a single-chain precursor, which is cleaved enzymatically into a 185 kDa heterodimer of a 35 kDa (α) and a 150 kDa (β) disulfide-linked chain (Gaudino et al., 1994). In addition, the RON ligand macrophage-stimulating protein (MSP or MST1) requires proteolytic maturation before it can be bound to RON receptor and activated (Wang et al., 1994a; b).

Under normal conditions, RON signalling regulates inflammatory immune response and wound healing in humans (Yao et al., 2013; Faham et al., 2016). Although RON and MSP are evolutionary highly conserved, their function varies in different species (Gaudino et al., 1995; Quantin et al., 1995; Aoki et al., 1996; Bassett, 2003). Since homozygous knockout of RON is embryonically lethal in mice, RON is essential to life (Muraoka et al., 1999). In relation to disease, the RON receptor kinase plays an active role through its splicing isoforms in cancers of the breast, pancreas, lung and colon (Ghigna et al., 2005; Mayer et al., 2015; Chakedis et al., 2016). Among the different RON isoforms produced aberrantly in cancer, the isoform formed by the skipping of the *RON* alternative exon 11 'RONΔ165', is the focus of the current study. The skipping isoform is translated into a protein with an in-frame deletion of 49 amino acids and consequently, the receptor persists as a single-chain precursor molecule in the cytoplasm. Previous studies have shown that spontaneous multimerisation of RONΔ165 occurs, finally leading to constitutive activation of RON, meaning an activation which is independent of binding of a ligand (Collesi et al., 1996; Zhou et al., 2003). In contrast to other *RON* splicing isoforms, RNA-seq studies showed RONΔ165 overexpression leads to similar gene expression changes to those produced by the activation of the full-length protein, suggesting that RONΔ165 promotes tumour formation by aberrant RON activation (Chakedis et al., 2016). Moreover, RONΔ165 is consistently upregulated in metastatic cancer, and contributes to tumour invasiveness by promoting epithelial-to-mesenchymal transition (Collesi et al., 1996; Zhou et al., 2003; Wang et al., 2004; Ghigna et al., 2005; Mayer et al., 2015). Targeted strategies for cancer therapies involving RON are based on therapeutic antibodies. Moreover, antibody-drug conjugates (ADCs), which are combinations of anti-RON monoclonal antibodies (mAbs) with cytotoxic drugs, particularly Zt/gr-DM1 and H-Zt/g4-MMAE have been shown to eradicate tumours *in-vivo* by inducing RON internalisation (Yao et al., 2006, 2019; Li et al., 2010). Additionally, mAbs that bind to RON receptors and block MSP-RON signalling pathways have been assessed in clinical trials (ClinicalTrials.gov Identifier: NCT01119456; antibody RON8, Narnatumab, ImClone; phase-I discontinued) (O'Toole et al., 2006). However, even though the RON-antibody binding affects MSP-RON signalling, constitutive activation of RONΔ165 is not affected as it is ligand independent and tumours expressing this isoform can escape such therapies (Chakedis et al., 2016). Therefore, more information on the splicing impact of mutations on RONΔ165 pathogenic isoform in patients is required to advance personalised or targeted therapy.

Decoding the regulation of alternative splicing

In contrast to high-throughput strategies, previous strategies employed minigene reporters with limited number of targetted mutations to decipher the role of splicing related mutations (Cooper, 2005). These strategies still remain the most reliable and direct way to assess mutation and its consequent phenotypic effects. However, they are time consuming and impractical for approaches where a large scale or novel splicing study is necessary. Additionally, they fail to detect more complex regulation events such as those involving epistasis between two or more mutations (Nasrin et al., 2014; Ahsan et al., 2017). Hence, recent *in-silico* studies in combination with experimental tools helped better predict genome-wide splice-altering mutations (Soemedi et al., 2017; Adamson et al., 2018; Cheung et al., 2019). To study the role of RBP regulators, a genome-wide siRNA robotic screen catalogued the splicing regulators of a specific endogenous gene (Tejedor et al., 2015). With the advent of NGS, high-throughput studies benefited from combining deep sequencing techniques with minigene reporter mutagenesis screens. These studies enabled a more accurate prediction of the impact of sequence variants on splicing, provided mutational screening of a human exon, or allowed insights into general splicing mechanisms (Rosenberg et al., 2015; Julien et al., 2016; Ke et al., 2018, respectively). However, mutagenesis approaches in these studies were targetted to specific exonic regions or restricted to the use of short synthetic spacer regions. Additionally, regulatory layers like *trans*-acting factors or secondary structures were not considered. Given these constraints, there is a need to study the comprehensive landscape of splicing regulation using a combination of *in-silico* and *in-vivo* approaches.

Motive and objectives of the project

In light of limitations in the previous approaches, the motive of this PhD project is to establish a high-throughput screen to comprehensively identify and characterise the complete *cis*-regulatory landscape controlling the splicing regulation of an alternatively spliced exon. In order to study a cancer-relevant splicing decision, we chose a minigene derived from the proto-oncogene *RON* (derived from the *MST1R* genomic locus). The aim is to assess the splicing regulatory effects of the exonic and intronic mutations of *RON* minigene in both physiological and diseases states and to expand pre-existing knowledge on the mechanisms of alternative splicing regulation.

The objectives of the project are as follows -

1. To establish a high-throughput screening method to identify *cis*-regulatory elements by developing bioinformatics pipelines for analysing the *RON* mutagenised minigene library.

2. To identify key *cis*-regulatory elements in the *RON* minigene from the high-throughput screen.

3. To identify important *trans*-acting factors regulating splicing of *RON* minigene by analysing RNA-seq data resulting from KD of known splicing factors.

2

Materials and methods

Experimental methods

The methods presented below were performed by the following people under supervision by Dr. Julian König - Simon Braun (*RON* high throughput mutagenesis screen and library generation and prep.), F. X. Reymond Sutandy (iCLIP experiments) and Laura Schulz (RT-PCR validation experiments).

Cloning of recombinant plasmid DNA

The *RON* wt plasmid resulted from PCR amplification of a segment of the *MST1R* gene by polymerase chain reaction using Phusion DNA polymerase (NEB) with the forward primer oJ303 and the reverse primer oS111 at an annealing temperature of 65 °C with human genomic DNA (Promega) as a template (**Figure 4.5**). The 779 bp DNA product was gel-purified with the QIAquick Gel Extraction Kit (QIAGEN) and then digested using HindIII and XbaI restriction endonucleases (NEB). The cut DNA fragment was purified using a PCR purification kit (QIAGEN) prior to ligation into the pcDNA 3.1 (+) vector (Invitrogen). To match AE inclusion in the *RON* wt minigene to endogenous levels, the first nucleotide of the alternative exon was replaced by a guanine. Plasmids containing point mutations were generated using the Q5 Site-Directed Mutagenesis Kit (NEB) according to the manufacturer's instructions.

The *HNRNPH1* open-reading-frame was PCR-amplified from a plasmid (kindly donated by Dr. Davor Lessel, Institute of Human Genetics, University Medical Center Hamburg-Eppendorf) using oS428 and oS429 (**Figure 4.5**) and cloned into the pcDNA 3.1 (+) vector (Invitrogen) to generate an overexpression construct. Successful cloning of plasmids was monitored by diagnostic restriction digest of the target vector followed by analysis of resulting fragments via agarose gel electrophoresis. In addition, relevant sequences of all recombinant plasmids were validated by Sanger sequencing.

SDS PAGE and Western blotting

Protein samples were first separated by SDS-PAGE using precast 4-12% NuPAGE Bis-Tris gels (Invitrogen) and MOPS SDS running buffer (Invitrogen). Subsequently, proteins were transferred to a 0.45 μm pore size nitrocellulose membrane (GE Healthcare). Before Western Blot analysis, loading was controlled by staining membranes using Ponceau red staining solution (**Supplementary Table 2**). Antibodies used for Western Blot analysis are listed in (**Table 4.9**).

Synthesis of cDNA and semiquantitative RT-PCR

Extraction of RNA for subsequent analysis was carried out using the RNeasy Plus Mini Kit (QIAGEN). Semiquantitative RT-PCR was used for quantification of isoform ratios of individual plasmids and endogenous *RON* mRNA. Towards this, reverse transcription was achieved in a volume of 20 μl using 500 ng of total RNA, 1 μl $(dT)_{18}$ primer (100 μM, Thermo Scientific), 1 μl dNTPs (10 mM, NEB), and 1 μl RevertAid reverse transcriptase (Fermentas) by heating 70 °C for 5 min, 25 °C for 5 min, 42 °C for 60 min, 45 °C for 10 min, and 70 °C for 5 min. Subsequently, 1 μl of the cDNA was used as a template for the PCR reaction with the following condition: 94 °C for 30 s, 24 cycles (minigene) or 35 cycles (endogenous) of [94 °C for 20 s, 52 °C (minigene) or 62 °C (endogenous) for 30 s, 68 °C for 30 s] and final extension at 68 °C for 5 min. The primers oS066 (forward primer) and oS067 (reverse primer) were used to amplify the minigene derived isoforms. These primers annealed to the upstream constitutive exon and a region located downstream of the random barcode but upstream of the polyadenylation site. The primers oS044 (forward primer) and oS045 (reverse primer) were used to amplify endogenously derived isoforms. To analyse endogenous *RON* mRNA in *RON* minigene transfected cells, the same reverse primer but a different forward primer spanning the splice junction between exon 12 and 13 was used (oS237) with the condition as follows: 94 °C for 30 s, 32 cycles of [94 °C for 20 s, 61 °C for 30 s, 68 °C for 60 s] and final extension at 68 °C for 5 min. This region is not part of the minigene and endogenous *RON* transcripts are thus exclusively amplified in the PCR reaction. The TapeStation 2200 capillary gel electrophoresis instrument (Agilent) was used for isoform quantification of the PCR products (**Figure** 4.5).

Quantification of mRNA levels by RT-qPCR

To quantify mRNA levels of siRNA treated cells, cDNA was analysed by real-time quantitative reverse transcription PCR (RT-qPCR) using the Luminaris HiGreen Low ROX (Thermo Scientific) and the ViiA 7 Real-Time PCR System (Thermo Scientific). PCR amplification efficiency of each RT-qPCR primer pair was assessed by the dilution method and relative mRNA expression levels were quantified by efficiency corrected calculation (Dorak, 2006, 2007). Beta-Actin was used as a reference transcript for normalization. All measurements were carried out in technical triplicates. Primers used for RT-qPCR are listed in (**Figure 4.5**).

Cell culture

HEK293T, MCF7, and MDA-MB-435S cells were grown in Dulbecco's modified Eagle medium (DMEM; Invitrogen) supplemented with 10% foetal bovine serum (Invitrogen) at 37 °C with 5% CO_2. MCF10A cells were cultured in DMEM/F12 (Invitrogen), supplemented with 20ng/ml epidermal growth factor (Peprotech), 0.5 mg/ml hydrocortizone (Sigma), 100 ng/ml cholera toxin (Sigma), 10 μg/ml insulin (Sigma), 5% horse serum (Invitrogen) and 1x penicillin/ streptomycin (Invitrogen) at 37 °C with 5% CO_2.

siRNA-mediated knockdown of target mRNA

RNA interference induced knockdown of target mRNAs was performed using single small interfering RNAs (siRNAs) at a final concentration of 20 nM synthesized by Sigma-Aldrich. For gradual *HNRNPH* KD, the siRNA concentration ranged from 0.05 nM to 10 nM. One day before transfection, 2 x 10^5 HEK293T cells were seeded in a 6-well plate to result with approximately 20% confluence at the day of transfection. MCF7 cells were seeded three days prior to transfection with 0.5 x 10^5 cells per well of a 6-well plate. The transfection mix was prepared by incubating 3 μl of RNAiMax (Invitrogen) with 2 μl of siRNA (20 μM) in 200 μl OPTI-MEM (Invitrogen) for 20 min and then the mixture was added dropwise to the cells. The cells were collected 48 hours after the knockdowns. Knockdown efficiencies were assessed by RT-qPCR or Western blot analyses. siRNA sequences are listed in (**Table 4.8**).

Transfection of minigene plasmids

Transfection of minigene plasmids in six-well plates was carried out by mixing 2 μg of minigene plasmid DNA, 100 μl OPTI-MEM (Invitrogen), and 10 μg or 20 μg polyethylenimine MW ~ 2500 transfection reagent (Polysciences, Inc.) for HEK293T or MCF7 cells, respectively. Following incubation for 20 min, the mixture was added to the cells. For overexpression of *HNRNPH1*, the respective plasmid was transfected using 5 μl of Lipofectamine2000 (Invitrogen) and 1 μg or 2.5 μg of plasmid DNA according to the manufacturer's instructions. Cells were harvested another 24 hours after the transfection.

Preparation of mutated *RON* minigene library

Mutagenesis was done by error-prone PCR amplification of the wild-type *RON* minigene using the GeneMorph II Kit (Agilent). The target mutation frequency was controlled via adjusting the input DNA concentration and the number of PCR cycles i.e. when lower input DNA amounts are used, higher number of PCR cycles are required to reach saturated PCR product levels, resulting in increasing mutation frequencies.

Subsequently, a mutation rate of three to four mutations per minigene was achieved by mutagenic PCR using the forward primer oJ303 and the reverse primer oS111 using 8 μg, 4 μg, or 0.8 μg of *RON* wt plasmid DNA (corresponding to 1 μg, 0.5 μg, or 0.1 μg PCR amplicons of 776 bp) and 30x, 30x, or 20x PCR cycles, respectively. The reverse primer contains a 15 N random region as a molecular barcode to uniquely assign each of the generated minigene variants. PCR products were separated via agarose gel electrophoresis using a 0.8% TAE gel and subsequently purified using the QIAquick Gel Extraction Kit (QIAGEN). Next, PCR products were prepared for ligation by restriction digest using HindIII (NEB) and XbaI (NEB) restriction endonucleases. Using 3:1 molar excess of insert to vector, the PCR products were then ligated to dephosphorylated and HindIII (NEB) and XbaI (NEB) digested pcDNA 3.1 (+) vector (Invitrogen) using a Quick Ligation Kit (QIAGEN). Subsequently, *Escherichia coli* DH5α were chemically transformed with 2 μl of the ligation mixture. We generated three independent libraries using the same process.

For single bacterial colonies of equal size, 5-10% of the cell-ligation mixture was spread on multiple selection plates to yield 150-200 transformants per 10 cm plate. Following overnight incubation at 37 °C, the number of transformants per plate was counted and then 2000 transformants per library (each transformant corresponds to a single mutant minigene variant) were harvested using LB + ampicillin selection medium and a Drigalski spatula for detachment of the colonies. Finally, the plasmid DNA of the collected cells was extracted using the Plasmid Plus Midi Kit (QIAGEN). In addition, 200 wild type plasmids were generated to be used as a spike-in to the abovementioned libraries by using the same primers and template wild type plasmid but non-mutagenic PCR amplification with Phusion DNA Polymerase (NEB) and the following conditions: 98 °C for 30 s, 30 cycles of [98 °C for 10 s, 61 °C for 20 s, 72 °C for 20 s] and final extension at 72 °C for 5 min. Following confirmation of similar mutation frequencies between the three mutant minigene libraries, they were mixed in 1:1:1 molar ratio with 3.5% of the wild type spike-in library to result with a final library containing approximately 6,200 minigenes in equimolar ratio.

Library quality controls

Mutation frequencies of individual minigene variants from the libraries were assessed by Sanger sequencing. To this end, re-transformation of *E. coli* DH5alpha with *RON* library DNA and subsequent plasmid DNA isolation was carried out to obtain single minigene constructs for sequencing. Similarly, wild type minigenes were obtained by re-transformation of *E. coli* DH5alpha with the wild type library and following plasmid DNA isolation. To test whether the 15 N random barcode region affects splicing, splicing of these wild type minigenes was tested in semiquantitative RT-PCR analysis.

Library amplification

To amplify pre-existing library, 36 ng of the mutant minigene library were electroporated using One Shot™ TOP10 Electrocomp™ *E. coli* (Invitrogen). Following outgrowth, a dilution series ranging from 1:10^2 to 1:10^7 was spread on selection plates to estimate the number of successful transformants (obtaining approximately 7x10^6 transformants; corresponding to 1,000-fold coverage of the 6,200 plasmid library). The remaining transformation mixture was then transferred directly from outgrowth to 300 ml LB + ampicillin selection medium overnight. The next day, the grown bacteria cultures were split in six 50 ml aliquots and plasmid DNA was extracted using the Plasmid Plus Midi Kit (QIAGEN). The integrity of the amplified library was confirmed by RT-PCR analysis and RNA-sequencing (RNA-seq).

Emulsion PCR amplification of DNA fragments for high-throughput sequencing

To prevent chimeric amplicon formation during PCR-amplification of DNA and cDNA fragments intended for next-generation DNA- and RNA-seq, respectively, a water-in-oil emulsion PCR was carried out according to a previously published protocol (Williams et al., 2006). In brief, 400 μl of oil-surfactant mixture (**Supplementary Table 9**) was added to a 1.8 ml round bottom CryoTube vial (Nunc, Thermo Scientific) and stirred using a 3x8 mm stirring bar on a magnetic stirrer at 1,000 rpm. Next, a PCR mixture was made containing the following ingredients: 52 μl of 5x Phusion HF-Buffer, 26 μl of 100g/l BSA, 7.8 μl of each 10 μM forward and reverse primer (**Figure 4.5**), 5.2 μl of 10 mM dNTPs, 5.2 μl of Phusion DNA-polymerase (NEB), 1.65 μl of 5 ng/μl template DNA (DNA-seq) or 12 μl template cDNA (RNA-seq), and filled to 260 μl with water. Then 200 μl of the PCR mixture was added dropwise to the oil-surfactant mixture over a period of 1.5 min to generate a water-in-oil emulsion. After the addition of the PCR mixture was complete, the solution was stirred for additional 5 min. Next, the mixture was dispensed in 12x 50 μl PCR tubes and each reaction was overlaid with 10 μl mineral oil (Sigma Aldrich). PCR amplification was carried out using the following conditions: 98 °C for 30 s, 18 cycles (DNA-seq) or 15 cycles (RNA-seq) of [98 °C for 10 s, 61 °C (DNA-seq) or 56 °C (RNA-seq) for 20 s, 72 °C for 20 s (DNA-seq) or 1 min (RNA-seq)] and final extension at 72 °C for 5 min. To control the PCR amplification with a non-emulsified control reaction, 50 μl of the aqueous PCR mixture was amplified in addition to the 12x 50 μl emulsion PCR reactions. Following amplification, the emulsion PCR reactions were pooled in a 1.5 ml Eppendorf tube and oil and water phases were separated by centrifugation at 13,000xg for 5 min. Next, the upper (oil) phase was discarded and the water phase was extracted twice using 1 ml of water-saturated diethyl ether (Sigma Aldrich). Remaining solvent was removed from the broken emulsion by vacuum centrifugation for 10 min at 30 °C using a Vacufuge™ Concentrator (Eppendorf). The samples were then spun at 20,000xg for 3 min to pellet precipitated BSA and the supernatants were purified using the GeneRead Size selection kit (QIAGEN) for DNA-seq samples or Agencourt AMPure XP beads (Beckman Coulter) for samples intended for RNA-seq as follows: Two volumes of Agencourt AMPure XP beads were mixed with the sample and incubated for 15 min. Following three washing steps each using 200 μl of 75% ethanol, the beads were air-dried for 15 min.

Library preparation and sequencing of high-throughput DNA-seq libraries

Plasmid DNA from the *RON* minigene library was amplified by emulsion PCR and five overlapping amplicons were generated using five different forward primers oS118, oS119, oS120, oS138, and oS105, and oS106 as a common reverse primer. The purified products were first analysed with the TapeStation 2200 capillary gel electrophoresis instrument (Agilent) and then fluorimetrically quantified using a Qubit fluorimeter (Thermo Scientific). Samples were multiplexed in equimolar ratios prior to high-throughput sequencing. Sequencing was carried out on the Illumina MiSeq platform using 2 x 300 bp paired-end reads (600-cycle MiSeq Reagent Kit v3) and a 10% PhiX spike-in to increase sequence complexity.

Library preparation and sequencing of high-throughput RNA-seq libraries

For preparation of high-throughput RNA-seq libraries, the total RNA obtained from transfected HEK293T cells or MCF7 cells was enriched for mRNA by performing polyA selection using Dynabeads® Oligo $(dT)_{25}$ beads (Invitrogen) as follows: 50 μl beads were first equilibrated with 1 ml binding buffer and then resuspended in 50 μl binding buffer. 20 μg of total RNA was mixed with an equal volume of binding buffer and heated for 2 min at 65 °C. After heating, the RNA was immediately placed on ice and then the beads were incubated with the RNA shaking at 1,500 rpm for 10 min at RT. Next, the supernatant was removed and the beads were washed twice with 200 μl Washing Buffer B (**Supplementary Table 6**). Polyadenylated RNAs were finally eluted with 15 μl of 10 mM Tris-HCl for 2 min at 77 °C.

Reverse transcription was carried out using 500 ng of enriched mRNA under the above-mentioned conditions. Next, emulsion PCR using the following primers containing Illumina sequencing adaptors were used: oS119 (forward primer) and oS106 (reverse primer). Purified amplification products were analysed using the TapeStation 2200 capillary gel electrophoresis instrument (Agilent) and fluorimetrically quantified with a Qubit fluorimeter (Thermo Scientific). High-throughput sequencing was performed using 2 x 300 bp paired-end reads (600-cycle MiSeq Reagent Kit v3) on an Illumina MiSeq system and a 10% PhiX spike-in to increase sequence complexity.

iCLIP experiment

Individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) was used to capture the binding pattern of HNRNPH on the *MST1R* transcript. iCLIP was performed according to a previously published protocol (Sutandy et al., 2016). The iCLIP libraries were made from HEK293T cells 24 h after transfection of the *RON* wt minigene (in triplicates) or mutated *RON* minigenes carrying point mutations G305A (in triplicates), G331C or G348C (both in duplicates). The cells were irradiated with 150 mJ/cm^2 UV light at 254 nm. For the immunoprecipitation step, 7.5 μg of a polyclonal rabbit anti-HNRNPH antibody from Abcam (AB10374) were used. RNase digestion was performed by adding 10 μl of 1/100 diluted RNase I (Ambion) to the sample of the wt minigene experiment or 1/300 diluted RNase I (Ambion) to each sample of the experiment comparing the iCLIP landscape of the *RON* wt minigene with the *RON* G305A point mutation minigene. Next-generation sequencing was performed on an Illumina HiSeq2500 for the *RON* wt minigene (51-nt single-end reads) and MiSeq or NextSeq 500 for the RON wt/ point mutant minigene comparison (75-nt single-end reads).

Bioinformatics methods

In addition to below mentioned bioinformatics methods performed by myself, the methods and results presented below were performed by the following people under supervision by Dr. Katharina Zarnack with additional supervision by Dr. Stefanie Ebersberger-

Mariela Cortés-López (RBP site annotation), Dr. Markus Seiler (Annotation of RNA G-quadruplex sequences), Dr. Anke Busch (iCLIP and RNA-seq data processing and splice isoform quantification). Furthermore, we collaborated with Bernardo P. de Almeida and Dr. Nuno L. Barbosa-Morais for the TCGA and GTEx analyses.

Bioinformatics programs

The following list compiles the programs used in the current study.

- **Python v2.7.9: Anaconda v2.2.0 (64-bit)**

The Anaconda environment was used for scripting and development of pipelines in Python.

- **SLURM**

Slurm workload manager is an open source cluster management and job scheduling system running at the Goethe University Bioinformatics cluster facility. In order to run computationally intensive programs on the cluster, SLURM commands were used. The scripts were first tested on a 64-bit macOS Sierra 10.12.3 system and then run on the Bioinformatics computing cluster facility.

- **Bash**

Wrapper scripts were written as Bash scripts. SED and AWK were also used wherever simpler and faster functionalities were required, for instance; to calculate read lengths.

- **R**

Computing Cluster: v3.1.3 MacOS: v3.1.3 to v3.4.1 (R framework: RStudio v1.0 onwards)

R is a scripting language extensively used for statistical analyses and has excellent package ecosystem for bioinformatic analyses. Scripts written in R programming language were frequently used for computational analyses, with functionalities from Biostrings, ShortRead, Knitr, ggplot2, dPlyr, ggpubr among many others. More details on the functionalities used from these packages are described in the following sections.

- **FASTQC v0.11.3**

FastQC tool is used for checking the quality (Phred scale) of the sequenced reads in the study (https://www.bioinformatics.babraham.ac.uk/projects/fastqc).

- **Trimmomatic v0.33**

Trimmomatic is a tool used in trimming the sequencing reads. The parameters used and their explanations are given below (Bolger et al., 2014).

• HEADCROP: Cut the specified number of bases from the start of the read • MINLEN: Drop the read if it is below a specified length • SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.

- **Picard v1.131**

Picard is a tool used in the alignment of sequenced reads from the DNA-seq data (http://broadinstitute.github.io/picard/).

- **Samtools v0.1.19**

Samtools is a tool used in the manipulation of data formats and alignment of sequenced reads from the DNA-seq data (Li et al., 2009).

- **NextGenMap v0.4.12**

NextGenMap (http://cibiv.github.io/NextGenMap/) is a fast read alignment tool, using a hash-table based algorithm, and is used to align DNA-seq reads (Sedlazeck et al., 2013).

- **STAR v2.4.2a**

STAR (https://github.com/alexdobin/STAR) is a splice-aware aligner used in aligning the RNA-seq reads of the library (Dobin et al., 2013).

- **PhyloP**

PhyloP conservations scores retrieved from UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgTables; table: Mammal Cons, PhyloP46wayPlacental) are a measure of evolutionary conservation at individual alignment sites and are interpreted in elation to expected evolution under neutral drift:

Positive scores – Measure conservation, which is slower evolution than expected, at sites that are predicted to be conserved.

Negative scores – Measure acceleration, which is faster evolution than expected, at sites that are predicted to be fast-evolving (Pollard et al., 2010).

- **MaxEntScan**

MaxEntScan algorithm models the short splice site sequences accounting for non-adjacent as well as adjacent dependencies between positions and is based on the Maximum Entropy Principle and comapres other models as well (Yeo et al., 2004).

- **G4hunter**

G4hunter is a tool that takes into account G-richness and G-skewness of a given sequence and gives a quadruplex propensity score as output (Bedrat et al., 2016).

- **SPIDEX v1.0**

SPIDEX predicts splicing changes based on multiple sequence features with deep learning models (https://www.deepgenomics.com/spidex) (Xiong et al., 2015).

- **Adobe Illustrator v22**

Illustrator was used to reduce image size, improve colouring scale, increase visibility and adding annotations. Care was taken to avoid any manipulation of data in images.

Bioinformatics databases

The following list compiles the databases used in the current study.

- **ATtRACT** (A daTabase of RNA binding proteins and AssoCiated moTifs)

ATtRACT (https://attract.cnic.es/) is a database of experimentally validated RNA binding proteins and associated motifs. The scan sequence tool of the database was used to identify potential RBP regulator binding sites along the *RON* minigene sequence (Giudice et al., 2016).

- **ExAC** (The Exome Aggregation Consortium)

ExAC database catalogues variations and respective allele frequencies obtained from sequencing the protein-coding regions of 60,706 individuals of varied human ancestries (Lek et al., 2016).

- **COSMIC** v82 (Catalogue Of Somatic Variants In Cancer)

COSMIC is a comprehensive resource for exploring the impact of somatic mutations in human cancer (https://cosmic-blog.sanger.ac.uk/cosmic-release-v82/) (Bamford et al., 2004).

- **TCGA** (The Cancer Genome Atlas)

Database of molecular characterisations (DNA, RNA, proteins) of 20,000 primary cancer and matched normal samples across 33 cancer types (Tomczak et al., 2015).

- **GTEX** (Genotype-Tissue Expression) project v7

GTEX is a comprehensive public resource of tissue-specific gene expression and regulation, comprising of sequenced samples from 53 non-diseased tissue sites across nearly 1000 individuals (Carithers et al., 2015; DeLuca et al., 2015).

Bioinformatics methods: Data retrieval, sample processing via pipelines and further analyses

iCLIP data processing

The iCLIP data was processed at the core bioinformatics facility at IMB, Mainz. Initially, the sequencing reads were filtered by checking the Phred quality scores of experimental and random barcode regions and thereafter adapter trimming was performed. The trimmed reads were mapped to the human genome (hg19/GRCh37) using STAR (Dobin et al., 2013) resulting in ~49 million (HiSeq 2500), ~10 million (MiSeq), or ~178 million (NextSeq 500) uniquely mapping reads. The resulting crosslink events were normalised to the total number of crosslink events within the minigene region excluding *RON* exon 11. normalised counts were averaged between replicates, counted into 5-nt sliding windows and then subtracted between the conditions to determine differences in HNRNPH crosslinking. HNRNPH iCLIP data of the *RON* wt was then compared with those of the minigenes containing the selected point mutations and was used to determine differential binding of HNRNPH.

DNA-seq data processing and mutation calling

The DNA-seq library was sequenced using Illumina MiSeq (300-nt paired-end) technology with a total of 40 million reads (merge of two libraries) and analysed with a custom built pipeline (Python version 2.7.9: Anaconda 2.2.0, 64-bit) consisting of custom Python, R and bash scripts (**Figure 3.7 A**).

In detail, FastQC (fastqc_v0.11.3; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for quality control, and Trimmomatic (version 0.33; parameters HEADCROP: 20 SLIDINGWINDOW: 7:10 MINLEN: 0) (Bolger et al., 2014) was used for removal of additional bases like sequencing adapters (HEADCROP) and trimming of low-quality bases (SLIDINGWINDOW corresponding to an average Phred score < 10 in 7-nt window).

To test the optimal combination of listed trimming parameters, average Phred quality scores of 5, 10, 20 and 30 were checked. For the sliding window; the parameters were set at 3-nt, 5-nt, 7-nt and none, before deciding on an afforementioned optimal parameters based on balancing sufficient good quality reads being kept against those that were filtered out.

Then, the resulting reads were filtered for a minimum length of 130 nt (read #1) and 90 nt (read #2). This was done in a way to capture maximal sequence content along with sufficient barcode diversity. In order to extract the 15-nt barcode (read #1) which assigns the read pairs to an individual minigene variant, the function ‘matchLRPatterns()’ from the R/Bioconductor package ‘Biostrings’ was used to search for the flanking restriction sites or patterns (Lpattern=”TCTAGA”, Rpattern=”GGATCC”, allowing for one mismatch). Read pairs were only retained when they passed a quality threshold of Phred score ≥ 30 at every single barcode position. The barcodes ranged from 14- to 16-nt with most being 15-nt as intended.

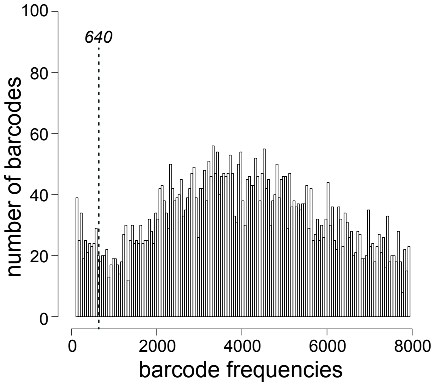


Figure 2.1: Coverage of reads for plasmids in the DNA-seq library. The bar plot shows the occurence of reads per plasmid counts of the *RON* minigene DNA-seq library. The distribution shows a marked increase at the threshold of 640 reads. The data of barcodes below a frequency of 2 plasmids, corresponding to the first peak, and above 8000 plasmids, are not shown.

Barcode filtering: The frequency of these barcode occurrences ranged from 1 to 8000+ and showed a bimodal distribution. It was postulated that the first peak showed rare barcode occurrences as a result of mutations or indels in the barcode sequence and the second smaller peak results from barcodes which are highly frequent and are true representatives of the clonal population of minigene variants shown in **Figure 2.1**.

Aligning the minigene variants

Each minigene variant was assigned read pairs based on its matching barcode sequence and had at least 640 read pairs for optimal coverage (threshold chosen as per **Figure 2.1**). Then, the matched reads for each mingiene variant were mapped to the *RON* wt minigene sequence using NextGenMap (version 0.4.12) (Sedlazeck et al., 2013). A read was reported as mapped if $>$ 50% of its bases were mapped, the alignment had an identity $>$ 65%, and at least one stretch of 13 bp was identical to the reference. Samtools (Li et al., 2009) and Picard (`http://broadinstitute.github.io/picard/`) were used to process, sort and index the resultant BAM files. Bedgraph files were created from these BAM files to calculate coverage maps of the *RON* minigene.

Read coverage from DNA-seq across the minigene variants

The read pairs that mapped across the minigene sequence, are a product of the amplification of five different forward primers and a common reverse primer, as previously explained in the library generation protocol. This resulted in a non-uniform but sufficient coverage across the whole minigene as in (**Figure 3.6 A and B**). The dips in coverage are a result of the trimming at the ends of reads amplified by the different forward-reverse primer combinations.

Mutation calling and filtering

Mutations were called using the HaplotypeCaller tool (version 3.4.0) of the Genome Analysis Toolkit (GATK) (Van der Auwera et al., 2013) with -dt NONE to override the downsampling function. This was done to include those mutations which remain unseen as downsampling leads to a loss of reads at certain positions. Overlapping reads were recounted using bam-readcount (`https://github.com/genome/bam-readcount`) and then manually filtered against single nucleotide variants (SNV) with low penetrance based on reference (Ref) and alternative (Alt) allele frequencies: (i) Alt / (Alt + Ref) > 0.8, and (ii) (Alt + Ref) / total > 0.5 taking into account all other isoforms. Since GATK was not built for such custom applications and is mostly used for human genomic variation datasets, we filtered manually using the mentioned filtering criteria.

The mutations included 18,948 point mutations and 608 short insertions and deletions. Indels were separately considered as independent sequence variants in the mathematical splicing model, but were not taken forward in the subsequent analyses as the focus was on SNVs. The final library contained 5,791 minigene variants, including 591 wt and 5,200 mutated minigenes. The accuracy of mutation calling was validated by Sanger sequencing of 59 randomly selected minigene variants, confirming the presence of all 169 GATK-called mutations without any further false-negatives.

RNA-seq data processing

RNA-seq libraries were sequenced on Illumina MiSeq (300-nt paired-end), yielding 17-22 million reads per sample and analysed with a custom Python pipeline similar to DNA-seq (**Figure** 3.7 **B**). As before, low-quality sequences were removed first (average Phred score < 20 in 6-nt window) and then the 15-nt barcode (read #1) was extracted as described in the previous section. Only reads originating from the 5,791 minigene variants that were recovered from the DNA-seq library were considered for further analyses.

Aligning the minigene variants from RNA-seq

Read pairs for each minigene variant were aligned to the *RON* wt minigene sequence using the splice-aware alignment algorithm STAR (version 2.5.1b) (Dobin et al., 2013), allowing up to 10 mismatches without adding any prior information on existing splice junctions. Only read pairs pertaining to splice isoform boundaries i.e. both mates extended at least 10 nt beyond the constitutive exon boundaries were considered. Furthermore, all improperly or inconsistently mapped read pairs were removed from the analysis. Read pairs are referred to as improperly mapped if they map with a wrong orientation, while inconsistently mapped read pairs overlap and show a disagreement in their mapping patterns. Finally, only minigene variants which were covered by at least 100 read pairs were used further, resulting in 5,697, 5,645 and 5,623 minigene variants detected in RNA-seq replicates 1, 2 and 3 from HEK293T cells, respectively.

Reconstruction and quantification of splicing isoforms

For each read pair, the underlying splicing isoform was reconstructed based on the CIGAR strings of the two mates. Isoforms, which were supported by less than 1% of the read pairs or less than two read pairs in any plasmid, were removed from the analysis. The frequency of each isoform for each minigene variant was calculated as the number of read pairs supporting this particular isoform in comparison to the total read pairs for all detected isoforms for one particular minigene variant. Finally, all thus derived non-canonical isoforms (cryptic splice site activation products) were categorised in the isoform group ‘other’.

Dynamic model of splicing

Splicing dynamics were modelled using a set of ordinary differential equations, in which concentrations of RNA intermediates are determined by production and degradation terms. The pre-mRNA precursor x_0 is produced at a constant rate c and spliced into five splice products with linear kinetics and rates r_i. All non-canonical isoforms are included in the model as one additional species produced at rate r_6. This leads to $\frac{\mathrm{dx}_0}{\mathrm{dt}} = c - (r_1 + r_2 + r_3 + r_4 + r_5 + r_6)\, x_0$. Six additional differential equations describe the dynamics of the canonical (AE inclusion, AE skipping, full IR, first IR, and second IR) and non-canonical ('other') splice isoforms. The concentration x_i of isoform i is described by $\frac{\mathrm{dx}_i}{\mathrm{dt}} = r_i x_0 d_i x_i$, where d_i are RNA degradation rates.

The measured isoform frequencies in the model correspond to the concentration of transcripts x_i normalised by the total RNA concentration. These fractions are calculated analytically from the consideration of the system being in a steady state. As a result, the frequency p_i of a certain isoform i has the form $p_i = \frac{K_i}{(K_1+K_2+K_3+K_4+K_5+K_6)}$.

Here, the splicing rates $K_j = \frac{r_j}{d_j}, j = 1, 2, 4, 5, 6$ are the ratios of production and degradation rates for the isoforms involving splicing, and $K_3 = 1 + \frac{r_3}{d_3}$ reflects sum of the unspliced pre-mRNA (x_0) and full intron retention (x_3) isoforms, which cannot be discriminated experimentally. Thus, due to normalization, a change in the production rate of one isoform due to a particular mutation will affect all isoform frequencies, and this effect depends in a nonlinear manner on the values of all splice rates K_i (i.e., on the mutational background). To infer the mutation effects from the data, the isoform ratio relative to the inclusion isoform ($\frac{p_i}{p_1} = \frac{K_i}{K_1}$) was used as a representative as this no longer depends on all splice rates, and relates to K_i in a linear fashion.

Calculation of single mutation effects by linear regression

To calculate single mutation effects in HEK293T cells, the combined log fold-changes of multiple mutations on a splice isoform ratio were assumed to be the sum of individual log fold-changes. One such equation was formulated for each minigene, resulting in a system of 5,621-5,697 equations for each splice isoform ratio, depending on the number of minigene variants obtained from each RNA-seq replicate.

To test the assumption of additive mutation effects, single mutation effects were analysed interact in minigenes containing several mutations.

To this end, a subset of mutations that is contained in the library as single-mutation minigenes (~600 minigene variants), and furthermore occur within double/triple-mutation minigenes together with other mutations from the list was analysed. For the majority of these mutations, the combined mutational effects on the splicing rates K_i were multiplicative, e.g. $K_i\frac{\left(m_1,m_2\right)}{K_i}\left(\mathrm{WT}\right) = K_i\frac{\left(m_1\right)}{K_i}\left(\mathrm{WT}\right) * K_i\frac{\left(m_2\right)}{K_i}\left(\mathrm{WT}\right)$, where $K_i\left(\mathrm{WT}\right), K_i\left(m_1\right), K_i\left(m_2\right)$ and $K_i\left(m_{1,}m_2\right)$ are the splicing rates of the wt minigene and of the minigenes including mutation m_1 or mutation m_2 or both mutations m_1 and m_2, respectively. In practice, the mutational effects $K_i\frac{\left(m_1,\ldots m_n\right)}{K_i}\left(\mathrm{WT}\right)$ were calculated as a mutation-induced fold-change of the splice isoform ratios p_i/p_1 (see above). By a log-transformation, the above multiplicative relationship transforms to a linear one that connects the measured cumulative mutation effects with the predominantly unknown single mutation effects. For the whole pool of measured minigene variants, this constitutes a system of linear equations that can be solved for the single mutation effects in a least-square sense. Additionally certain minigenes were excluded (**Figure 4.1**) from the linear regression because they deviated from linear behaviour:

(i) Minigenes simultaneously harbouring two splice site mutations as apparent lack-of-effect of secondary splice site mutations at non-zero inclusion frequencies contradicts their strong effect as isolated splice site mutations, and introduces strong inconsistencies and biases in linear regression. This is probably due to leaky sequencing reads originating from other minigenes where inclusion is the predominant isoform.

(ii) Minigenes with strong activation of cryptic splice sites: The activation of cryptic splice sites by mutations leads to the generation of a plethora of new splicing prod-

> ucts ('other') which behave heterogeneously and cannot be considered in our model. Therefore, we performed first the regression on the complete dataset and subsequently excluded the minigenes containing mutations that were predicted to exhibit an increased 'other' isoform frequency.

As an alternative approach to estimate the mutation effects of the excluded mutations, the median isoform frequency was calculated using isoform frequencies from all minigene variants harbouring the given mutation (**Table** 4.5). Moreover, the median isoform frequency was then compared to the prediction of the regression model. If enough minigene variants with the mutation are present in the library, this procedure should average out the effect of accompanying mutations. The median isoform frequency for a mutation was independently calculated for each isoform category and treated as a representative measure of the splicing effect of that particular mutation.

Estimation of the inference accuracy of the model

The training dataset comprised of single point mutations in individual minigene variants with their measured splicing effects (~600 mutations). These single-mutation minigenes were used to estimate the inference accuracy of the model, and to assess the dependency of the prediction accuracy on the occurrence of a mutation in the dataset. For each such mutation, the following cross-validation procedure was repeated: The single-mutation minigene was removed from the dataset before fitting the regression model, and kept for the evaluation of the regression results. The remaining minigenes containing the particular mutation were removed from the dataset successively (and in different permutations), and each time the effect of the mutation was assessed by regression and the prediction compared to the single-mutation minigene value. In this way, estimates for the inference error were obtained based on 1 up to $n - 1$ minigenes containing a particular mutation, where n is the total occurrence of the mutation in the dataset. In some cases, estimation of mutational effects was not possible from a reduced dataset, e.g. the inference error for a particular mutation was estimated only for occurrences between m and $n - 1$, with1 $< m \leq n-1$. Finally, the standard deviation of the prediction errors for all mutations was estimated for each measured frequency.

Definition of significant single mutation effects and synergistic interactions

The estimated single mutation effects on splice isoform ratios, obtained by linear regression were used to predict single mutation effects on each splice isoform frequency (p_i). To quantify the effects of each individual mutation on each isoform frequency, a z-score value was calculated from the model-derived single mutation effects, using the mean and standard deviation of the 591 wt minigene variants: $\frac{\left(p_i{}^{\text{mutation}} - \, mean(p_i{}^{\text{wt}})\right)}{standard \ deviation \ (p_i{}^{\text{wt}})}$. The z-scores were independently calculated per replicate and later averaged. Only mutations present in all three replicates were kept for further analyses.

In order to combine the evidence from the three replicate experiments, Stouffer's test was applied to combine the z-scores (Whitlock, 2005). The resulting metric (under standard normal distribution) was converted into a p-value and subjected to multiple testing correction (Benjamini-Hochberg). A mutation was considered as significant for a given isoform if it displays (i) ≥5% change in isoform frequency compared to the mean of the 591 minigene variants (ΔIF ≥ 5%), and (ii) less than 5% false discovery rate (FDR, adjusted p-value < 0.05). Combining all six isoform categories, this approach identified 778 and 1,022 splicing-effective mutations in HEK293T and MCF7 cells, respectively. These accumulated into 469 and 550 splicing-effective positions, i.e. nucleotide positions in the *RON* minigene where at least one out of three possible mutations shows a significant effect on at least one isoform.

To calculate z-scores for synergistic interactions between mutations and *HNRNPH* knockdown from the model-derived isoform ratios, the log-transformed fold-change was divided in isoform ratios (KD over control condition) by the wt variation (standard deviation). Then, the z-scores were calculated by replicates and then averaged, removing mutations that were not present in the three replicates under KD conditions. Stouffer's test and multiple testing correction was then applied as above. To identify significant synergistic interactions, a cutoff at 0.1% FDR (adjusted p value < 0.001) was applied. Additionally, a consistent directionality of the synergistic effects was required in all three replicates. Combining the five different isoform ratios, this approach identified 358 significant synergistic interactions (|z-score| > 2) on 281 positions between mutations and *HNRNPH* knockdown in MCF7 cells. Applying more stringent cutoffs at |z-score| > 3 or > 5 identified 227 or 71 significant synergistic interactions, respectively.

Characterisation of splicing-effective positions

Splice site strengths were predicted using the sequence analysis software MaxEntScan (Yeo et al., 2004) for all mutations in the positions considered by MaxEntScan (278-300 nt and 442-450 nt for the 3' and 5' splice site, respectively. PhyloP scores were retrieved from the UCSC Genome Browser for the genomic coordinates corresponding to the *RON* minigene (chr3:49933134 – 49933840, human genome version hg19). Gentoyped mutations with their respective allele frequency data were retrieved from Exome Aggregation Consortium (ExAC) for the exons within the *RON* minigene region.

Annotation of splice-regulatory RBP binding sites (SRBS)

The 'Scan Sequence' tool of the ATtRACT database was used (Giudice et al., 2016) to identify potential RBP binding sites along the *RON* wt minigene sequence. Duplicated records, e.g. due to overlapping database entries from different experimental methodologies, were removed. Only those binding sites were retained, for which ≥60% of positions were identified as splicing-effective in the screen. This step was independently performed for each splice isoform. Within each RBP, these binding sites were then collapsed if they shared an overlap of ≥2 nt and still harboured ≥60% splicing-effective positions for at least one isoform after collapsing, if they did not fulfil this condition, they were kept unmerged. For the comparison in **Figure** 3.28, the HNRNPH SRBS within each cluster were extended by 2 nt. Nucleotide positions in the two isolated SRBS in the constitutive exons were excluded from this analysis.

In order to connect mutation effects to the sequence specific binding of HNRNPH, G-run-disrupting mutations were defined as a G-to-H mutation at any position of the G-run, while the two possible H-to-G mutations in immediately neighbouring positions were counted as G-run-extending. **Figure 3.27 A** compares the median splicing effect (average of three biological replicates) of all G-run disrupting versus extending mutations for the 22 predicted HNRNPH SRBS. In detail, **Figure 3.30** gives the G-run disrupting mutation effects of the 5 SRBS at the alternative exon, contributing together to form the effects of cluster 3.

Comparison of screen-derived splicing-effective mutations to *in-silico* G-quadruplex propensity scores- and splicing effects predictions at the *RON* AE

Along with linear organisation of elements, secondary structures like G-quadruplexes (G-rich elements) also play a role in splicing regulation (Marcel et al., 2011). G-quadruplex propensity scores were retrieved for the positions 303-332 via the tool G4hunter (Bedrat et al., 2016). **Figure 3.18 A** indicates the effects of predicted G hunter scores in comparison to Delta Inclusion IF values

To compare SPIDEX (SPANR) predictions to the screening method, mutation effects of *RON* Exon 11 were compared to their counterpart mutation effects (dPSI) in SPIDEX dataset. The SPIDEX genome-wide dataset contained scores for predictions of the functional effects of 650,000 intronic and exonic variants on cassette splicing (Xiong et al., 2015). The variants predicted in the genomic coordinates corresponding to the *RON* minigene (chr3:49933134 - 49933840, human genome version hg19) were retrieved. **Figure 3.18 B** shows how SPIDEX predictions compare to AE Inclusion outcome of the screening.

Clinical relevance of *RON* mutations in Cancer

To analyse gene expression and alternative splicing profiles of healthy (control) and cancerous (case) human tissues, normalised gene expression datasets were collected from two different sources i.e. for case; The Cancer Genome Atlas (TCGA) (`https://cancergenome.nih.gov/`) and control; Genotype-Tissue Expression (GTEx) project (v7). Firstly, to establish the expression profiles of the control dataset, 11,688 *post-mortem* samples comprising of 30 human tissues from 714 healthy human donors were collected (Carithers et al., 2015; DeLuca et al., 2015). Then, normalised gene expression data of tumour samples comprising the case, were retrieved via Firebrowse (http://firebrowse.org/). Alternative splicing profiles for both these datasets were quantified using the newly developed *psichomics* tool (version 1.2.1, https://github.com/nuno-agostinho/psichomics, Nuno Saraiva-Agostinho and Nuno L Barbosa-Morais 2018). Consequently, with a minimum coverage threshold, *RON* exon 11 percent spliced-in (PSI) values were calculated.

In total, *RON* gene expression profiles and *RON* exon 11 PSI quantifications were derived from 2,743 control samples, from 24 healthy human tissues, and 4,514 tumour samples, from 27 cancer types. Normalised TPM values for comparison of HNRNPH2 expression between tumour and healthy tissues were obtained using Toil (Vivian et al., 2017).

Calculation of single mutation effects in TCGA tumour samples

Exome sequencing data from TCGA tumour samples were downloaded from Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/) after obtaining the required access permissions. A total of 153 patients consisting of 55 different mutations within the *RON* minigene were taken for analysis.

To quantify the splicing effects of these mutations, the difference of *RON* exon 11 skipping (calculated as 1 - PSI) between mutated and non-mutated tumour samples in each cohort was considered. Then, *RON* exon 11 skipping differences of these mutations were correlated with the skipping isoform frequencies (AE skipping) derived from the screen (of the same mutations). With a minimum read coverage threshold (average of greater than 24 on splice junctions), the correlation analysis was performed using data from 117 patients from 14 cohorts harbouring a total of 36 different mutations. The variability of *RON* exon 11 inclusion levels in TCGA patient samples was considering by taking into account the standard deviation of *RON* exon 11 in ‘unmutated’ TCGA tumour samples (i.e. without a given mutation) in the same cohorts.

Identification of *trans*-acting factors involved in *RON* alternative splicing regulation

Papasaikas et al. (2015) tested a large-scale RBP KD screen on various alternative exon models whereby, the KD effects of >200 RBPs on splicing of *RON* exon 11 and other alternative exons in HeLa cells were reported. In detail, the KD effect in terms of z-scores were calculated from percent spliced-in (PSI) upon siRNA treatment and the median absolute PSI deviation, divided by its standard deviation. A positive z-score implied more AE inclusion upon RBP KD and vice versa. Consequently, among 125 RBPs affecting *RON* exon 11 splicing ($|$z-score$| > 1.5$), 17 RBPs were identified to have predicted SRBS in the *RON* minigene. In order to identify potential regulators of *RON* exon 11 splicing in humans, mRNA expression profiles of RBPs were correlated with *RON* exon 11 splicing in cancer. The correlation analysis (Spearman correlation) was performed with a final list of 190 identified RBPs; consisting of 65 identified via ATtRACT, 108 identified in the RBP KD screen (Papasaikas et al., 2015), and 17 in common to both afforementioned approaches.

This selection included RBPs with different functions in splicing, such as HNRNPH, a known repressor of exon 11 inclusion, SRSF2, a global splicing activator, and PRPF6, a core component of the U5 snRNP subunit of the spliceosome. An initial selection of five RBPs was done for high-throughput screening approach in the context of the minigene library, to learn about their interplay with the cis-regulatory landscape of *RON*. The *RON* minigene library was transfected into HEK293T cells, in which HNRNPH, SRSF2, PRPF6, PUF60 or SMU1 had been depleted with siRNAs (RBP KDs) along with a mock siRNA control. Following splicing quantification in both the control and KD libraries, the AE inclusion levels of the respective libraries were compared to identify changes in splicing behaviour.

In addition to this, the mRNA expression levels of the RBPs were correlated with *RON* exon 11 inclusion levels across TCGA tumour samples. The significance of those correlations (ranking of minus base-10 logarithm of the associated p-value) was tested against those of all RBPs obtained from (Sebestyén et al., 2016) and of all protein-coding genes using Gene Set Enrichment Analysis (GSEA) tool (Mootha et al., 2003; Subramanian et al., 2005). RBPs and protein-coding genes were initially filtered to the ones showing at least the same average expression value as the least expressed pre-selected RBP, known to be highly expressed in cancer, so that GSEA was not biased by varying gene expression ranges. Then, to assess the relative magnitude of association between each RBP and *RON* exon 11 splicing, linear regression approaches were performed between the following datasets; the expression profile of each of the 190 pre-selected RBPs and *RON* exon 11 PSI in TCGA tumour samples.

Analysis of the splicing switch response and cooperativity

Differences in percent spliced-in (ΔPSI) data for *RON* exon 11 inclusion from the endogenous *RON* gene and the wt *RON* minigene measured at different *HNRNPH* knockdown (KD) and overexpression (OE) levels (**Figure 3.36**) were fitted using the following Hill function -

$y\left(x\right) = y_{\max} - \frac{\left(y_{\max} - y_{\min}\right) x^{n_H}}{x^{n_H} + EC50^{n_H}}$,

with x and y being vectors of experimentally determined HNRNPH levels and corresponding splicing outcomes (ΔPSI), respectively (**Figure 3.36 A**). $y_{\min}, y_{\max}, EC50, n_H$ are fitted parameters. Fitting was done by minimising the residual cost function

$$\chi^2 = \frac{\left(\Delta PSI - y\left(\text{HNRNPH}\right)\right)}{\sigma_{\Delta\text{PSI}}},$$

where σ_{PSI} denotes the standard deviation of the PSI measurement. Minimisation was done using the Matlab non-linear least-squares solver *lsqnonlin*. The parameter ranges used during fitting were $y_{\min} \in \left[-0.5, 0\right], y_{\max} \in \left[0, 0.5\right], EC50 \in \left[0.1, 2\right], n_H \in \left[1, 20\right]$. The optimal parameter values that were found were

for the endogenous *RON* gene: $y_{\min} = -0.11, y_{\max} = 0.36, \text{EC50} = 0.93, n_H = 17.4$

for the wt *RON* minigene: $y_{\min} = -0.11, y_{\max} = 0.3, \text{EC50} = 0.94, n_H = 13.8$

Confidence intervals were determined for all parameters by using a profile likelihood approach. For each fitted parameter θ, the following workflow was repeated: The parameter was assigned successively a number of values around its optimal value θ_0 listed above. While keeping this parameter at the fixed value, the remaining parameters were optimised and the value of the corresponding cost function was determined. Thus, the dependence of the cost function $\chi^2(\theta)$ on the parameter value around the minimum corresponding to the optimal value θ_0 was determined. The likelihood-based confidence interval for this parameter is defined by

$$\left[\theta, \chi^2(\theta) - \chi^2(\theta_0) < \chi^2(\alpha, 1)\right],$$

where α is the confidence level and $\chi^2(\alpha, 1)$ is the χ^2 distribution with degree of freedom 1. For each parameter, the 95% confidence intervals were found by determining the values θ on both sides of θ_0, for which the likelihood $\chi^2(\theta)$ crosses the threshold $\chi^2(\theta_0) + \chi^2(0.95, 1)$. The 95% confidence intervals found were for the endogenous *RON* gene:

$$y_{\min} \in [-0.12, -0.1], y_{\max} \in [0.28, 0.43], EC50 \in [0.89, 0.95], n_H \in [10.8, 35.2],$$

and for the wt *RON* minigene:

$$y_{\min} \in [-0.14, -0.08], y_{\max} \in [0.3, 0.31], EC50 \in [0.93, 0.95], n_H \in [10.4, 17.7]$$

3

Results

The *RON* minigene construct

In this study, a high-throughput screening approach is used to identify the landscape of mutations that regulate splicing in a minigene context. Firstly, it was imperative that the right minigene reporter system was chosen for performing random mutagenesis. Towards this, a minigene construct was obtained from the genomic locus of the *MST1R* gene which comprised of the exon 11 (alternatively spliced) together with its complete flanking introns and the exons 10 and 12 (**Figure** 3.1). In the context of the minigene construct, the upstream (exon 10) and downstream exons (exon 12) are constitutively spliced.

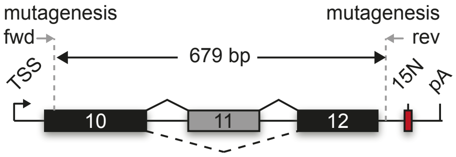


Figure 3.1: The *RON* minigene model focusses on the alternative splicing of exon 11 of the *MST1R* gene. The *RON* minigene comprises of the AE 11 and the upstream and downstream constitutive exons 10 and 12, respectively. A 679 bp region was chosen for mutagenesis using error-prone PCR with the indicated forward and reverse primers. TSS, transcriptional start site; pA, polyadenylation site; 15 N, the 15-nt barcode.

After the *RON* minigene was obtained, mutagenic PCR was performed resulting in a library of mutagenised variants of the minigene. High-throughput DNA-seq and following analyses led to the identification of mutation sets in the minigene library (**Figure 3.2**). Next, the library was transfected as a pool into human cells and the alternatively spliced isoforms were quantified by high-throughput RNA-seq. Importantly, a unique barcode sequence was used to tag each minigene variant. In addition, a 50-nt intronic spacer sequence was introduced to avoid any barcode effects. The barcode is recapitulated in the resulting splice-isoforms allowing the one-to-one assignment of mutated minigene variants to their corresponding splicing isoforms.

In order to study the role of alternative splicing regulation, the *RON* minigene reporter system was chosen for the following reasons -

1. *Limitations of Sequencing*: In order to sequence a minigene of interest, it is required to be of a permissible length due to sequencing technology limitations. This is because all the mutations and tagged barcodes must be characterised via paired-end DNA-seq. The resultant reads should cover the whole minigene without any gaps and with good quality phred scores. In the case of RNA-seq, the experimental barcode and all relevant splice junctions must be part of the paired-end RNA-seq read to allow proper assignment of minigenes and their corresponding splicing products. In light

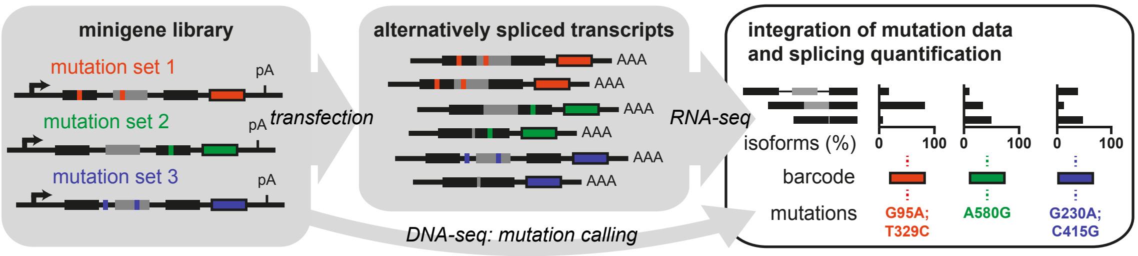


Figure 3.2: Overview of the High-throughput screening approach to decode *RON* exon 11 splicing. A library of mutated minigene variants resulting from mutagenic PCR; left panel. After transfection into human cells, the minigenes are transcribed and are alternatively spliced; middle panel. Mutations in the library and corresponding splicing isoforms are characterised by next-generation DNA-seq and RNA-seq, respectively. A unique 15nt barcode, that labels each minigene variant, links mutations to their splicing effects. Black and grey boxes indicate constitutive and alternative exons, respectively.

of these requirements, Illumina next-generation sequencing technology was used for both DNA-seq and RNA-seq. The longest read lengths available at the time for Illumina MiSeq platform were 2 x 300 bp with the 600-cycle kit. However, trimming the read ends for quality resulted in shortening of read length and necessitated the use of several primers along the minigene. Thus, sequencing becomes the limiting factor towards the choice of the minigene system allowing for a theoretical maximum distance of 300 bp between the 3′ end of barcode and relevant splicing information (i.e. junctions) containing sites (**Figure 3.3**). However, the quality of the reads reduce with increasing read lengths (Schirmer et al., 2015) and all of the 300 bp sequence of the read cannot be used for downstream analyses.

2. *Endogenous splicing of the minigene reporter*: Minigene splicing patterns may differ from the endogenous splicing outcome given its artificial nature. Evidently, increased AE inclusion levels were observed from the endogenous *RON* gene compared to the minigene reporter system (**Figure 3.4 A**). To offset this in the *RON* minigene, the thymidine at the alternative exon start was replaced with a guanine to in-

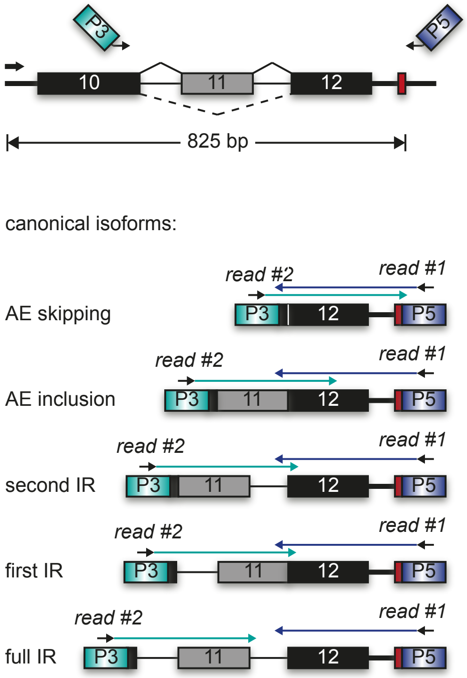


Figure 3.3: The five canonical isoform types are correctly identified by paired-end RNA-seq. Read 1 starting from the P5 adaptor provides the 15-nt barcode information and the splice junction upstream of exon 12, while read 2 from P3 reads the splice junction downstream of exon 10. For partial or full IR isoforms, both reads extend into the respective intron.

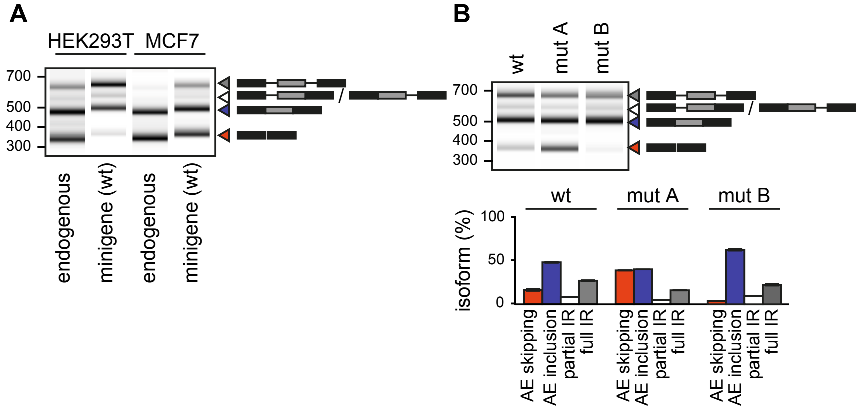


Figure 3.4: The *RON* minigene resembles its endogenous counterpart in function. RT-PCR analysis shows the *RON* minigene gives rise to the same splicing isoforms as the endogenous *MST1R* gene in HEK293T and MCF7 cells; in panel A. Gel-like representation of capillary electrophoresis. A 52 bp size difference between isoforms stems from different primer combinations used to differentiate between splicing products from endogenous *RON* and *RON* minigene as in (**Figure 4.5**). Published mutations mut A (Bonomi et al., 2013) and mut B (Lefave et al., 2011) trigger expected splicing changes towards increased or decreased AE skipping, respectively. panel B shows Gel-like representation of RT-PCR products from HEK293T cells. Bar diagram below shows the average isoform frequencies (in %) for AE inclusion and skipping, as well as partial and full IR from biological triplicates. Error bars denote the standard deviation. Partial IR refers to the sum of first IR and second IR isoforms that cannot be discriminated by RT-PCR analysis.

crease the 3′ splice site strength of the alternative exon (Yeo et al., 2004). In addition, RT-PCR based splicing analyses in different cell lines confirmed that the minigene splicing patterns are comparable to the endogenous locus as the same isoforms were produced. Therefore, the minigene reporter system allows the proper study of the regulation of *RON* exon 11 alternative splicing.

3. *Clinical relevance of RON AS regulation*: The AE skipping isoform involving *RON* exon 11 (RONΔ165) was previously shown to promote increased cell motility during epithelial to mesenchymal transition (EMT) (Ghigna et al., 2005). Therefore, this

aspect adds pathophysiological relevance to the study of alternative splicing regulation in the *RON* minigene.

4. *Prior knowledge of RON AS regulation*: Previous mutagenesis studies have identified the role of mutations in influencing *RON* alternative splicing. In line with these studies, two mutations were chosen that disrupt a HNRNPA1 (Bonomi et al., 2013) or HNRNPH binding site (Lefave et al., 2011) respectively for their effect in altering AE skipping in opposing directions (**Figure** 3.4 **B**). In summary, previous information was used in validating the regulation of the minigene reporter.

The *RON* minigene library

The library of *RON* minigene variants was generated with an optimised mutation rate of 3-4 mutations per minigene variant by varying amounts of template DNA and PCR cycles in the error-prone PCR amplification of the *RON* wt minigene (**Figure** 3.5 **A**). The PCR amplicons of each reaction mixture were used for cloning of mutated *RON* minigenes and upon transformation of *E. coli*, minigenes from ~2,000 colonies of each of the three transformations were collected (see methods; **Figure** 3.5 **B**). Sanger sequencing of ten minigene variants of each of the three libraries confirmed mutation frequencies with an average of 3-4 mutations per variant. Then, the libraries were pooled to form a single *RON* minigene library of ~6,000 minigenes. Moreover, the library was spiked-in with 200 *RON* minigenes with wt sequence which contained random barcodes, to act as an internal reference. RT-PCR analysis of a random selection of mutated minigene variants from the library showed highly variable splicing patterns, suggesting that the intended mutation frequency would suffice (**Figure** 3.5 **C**).

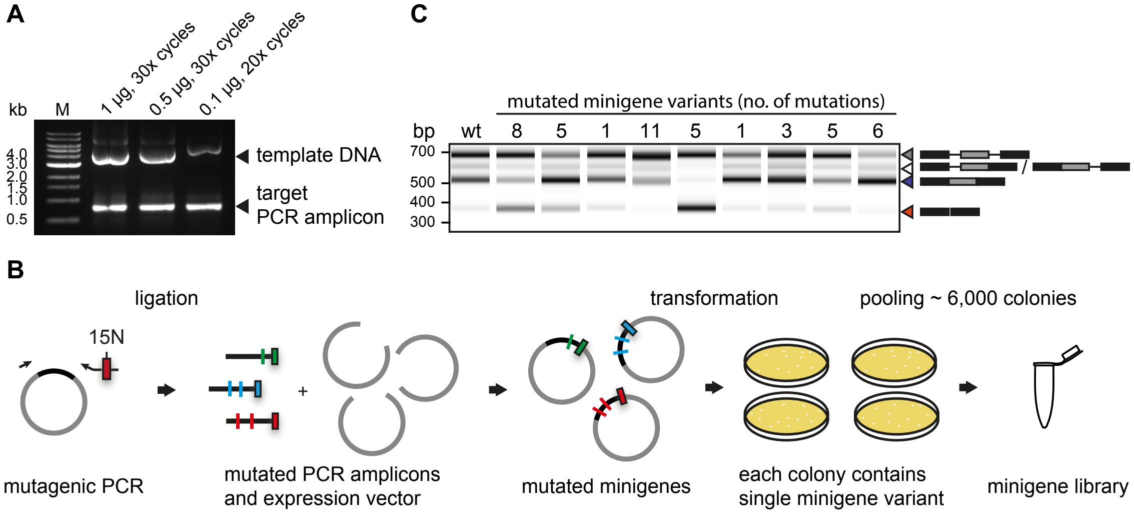


Figure 3.5: The library consists of thousands of mutated *RON* minigene variants. (A) Analytical agarose gel electrophoresis of fragments resulting from error prone PCR using varying amounts of *RON* wild type plasmid as template and indicated PCR cycles. The target PCR amplicon amount is saturated and equal for each condition and each condition results in minigenes with similar mutation frequency. (B) Schematic overview of the experimental steps performed to generate the *RON* minigene library. Mutagenic PCR generates mutated minigene fragments that are ligated with an expression vector to generate mutated minigenes. Following transformation in *E. coli*, transformants corresponding to single minigene variants are collected and the plasmid DNA is extracted, resulting in the minigene library. (C) Mutated minigene variants show different splicing outcomes. RT-PCR results displayed by gel-like representation of capillary electrophoresis from randomly selected clones from the *RON* minigene library in HEK293T cells. The number of mutations in each minigene variant is provided above.

Identifying mutations in the *RON* minigene library by high-throughput DNA-seq

To characterise the mutations of the minigene library, we performed next-generation sequencing with 300-nt paired-end reads (**Figure 3.6 A and B**). For sufficient coverage of the construct, five amplicons spanning the whole minigene were used. This was done for two reasons. Firstly, to offset the decrease in coverage due to bad quality at the ends of sequencing reads. It was previously was observed that read ends have a higher accumulation of errors during the sequencing process (Dohm et al., 2008; Schirmer et al., 2015). Thus, the 5 overlapping amplicons at the problematic areas helped increase the quality at the read ends (**Figure 3.6 C**). Secondly, amplicons which are larger in size compared to the rest, are preferred for primer extension during cluster formation. This resulted in in a biased library formation favouring the large amplicons. The overlapping amplicons helped reduce this bias in the library. Importantly, a 15-nt barcode was included in each read pair which enabled the complete reconstruction and assignment of all minigene variants in the library.

A custom-tailored Python pipeline was developed (**Figure 3.7 A**) to identify the mutations in the DNA-seq library. The library was sequenced on Illumina MiSeq (300-nt paired-end) with a total of 40 million reads. After initial quality control steps, the reads were filtered for a minimum length and the barcode which assigns the read pairs to an individual minigene variant, was extracted for each variant. Subsequently, read pairs for each minigene variant were aligned to the *RON* wt minigene sequence and the mutations were identified. Only those mutations that passed the rigorous filtering criteria were considered as the final set (refer materials and methods for details). In total, 5,791 unique minigene variants were captured including 5,200 with randomly introduced mutations and 591 with the wt sequence.

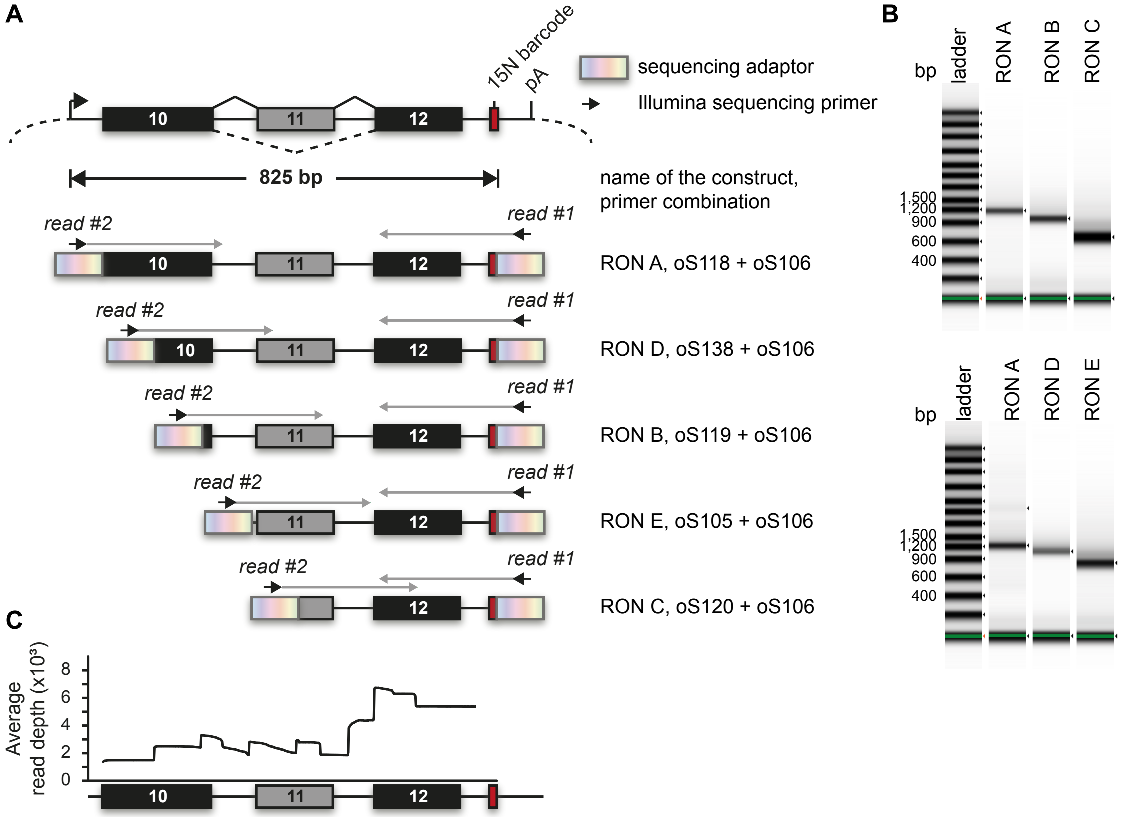


Figure 3.6: DNA-seq primer strategy of five overlapping amplicons ensures higher read coverage across the *RON* minigene sequence. (A) Overview of the five overlapping amplicons generated for paired-end DNA-seq. The reverse primer binds downstream of the 15-nt barcode (15N, red box) and introduces Illumina sequencing adaptor P5 (Read 1). Five variants of the forward primer bind to subsequent positions resulting in five overlapping amplicons of the minigene. The forward primers introduce P3 (Read 2). (B) Gel-like representation of the different PCR amplicons generated by emulsion PCR. The green line indicates the internal size standard. (C) Average read depth per position across the *RON* minigene after quality filtering and trimming. The graph is aligned with the overview of primer combinations in (A).

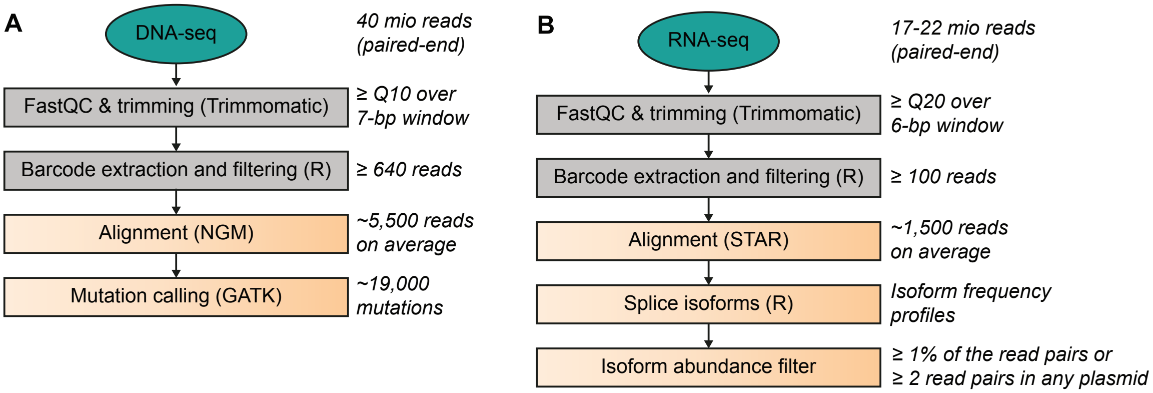

A
DNA-seq
40 mio reads (paired-end)
FastQC & trimming (Trimmomatic)
≥ Q10 over 7-bp window
Barcode extraction and filtering (R)
≥ 640 reads
Alignment (NGM)
~5,500 reads on average
Mutation calling (GATK)
~19,000 mutations
B
RNA-seq
17-22 mio reads (paired-end)
FastQC & trimming (Trimmomatic)
≥ Q20 over 6-bp window
Barcode extraction and filtering (R)
≥ 100 reads
Alignment (STAR)
~1,500 reads on average
Splice isoforms (R)
Isoform frequency profiles
Isoform abundance filter
≥ 1% of the read pairs or ≥ 2 read pairs in any plasmid

Figure 3.7: Custom bioinformatics workflows enable identification of mutations and corresponding splicing products from high-throughput DNA-seq and RNA-seq. (A) Bioinformatics workflow for DNA-seq analysis to characterise mutations. Quality control and trimming was performed with FastQC and Trimmomatic, respectively, followed by custom scripts to extract 15-nt barcode and filter for minigenes with ≥640 read pairs. Reads were aligned to wt *RON* minigene sequence using NextGenMap (NGM), and mutation calling was done using HaplotypeCaller tool from Genome Analysis Toolkit (GATK). (B) Bioinformatics workflow for RNA-seq analysis to quantify splice isoforms. Upon quality control and filtering similar to (A), reads were aligned to wt *RON* minigene using splice-aware alignment software STAR. All isoforms present in RNAseq library were reconstructed and filtered for minimum abundance using custom scripts. See methods for further details.

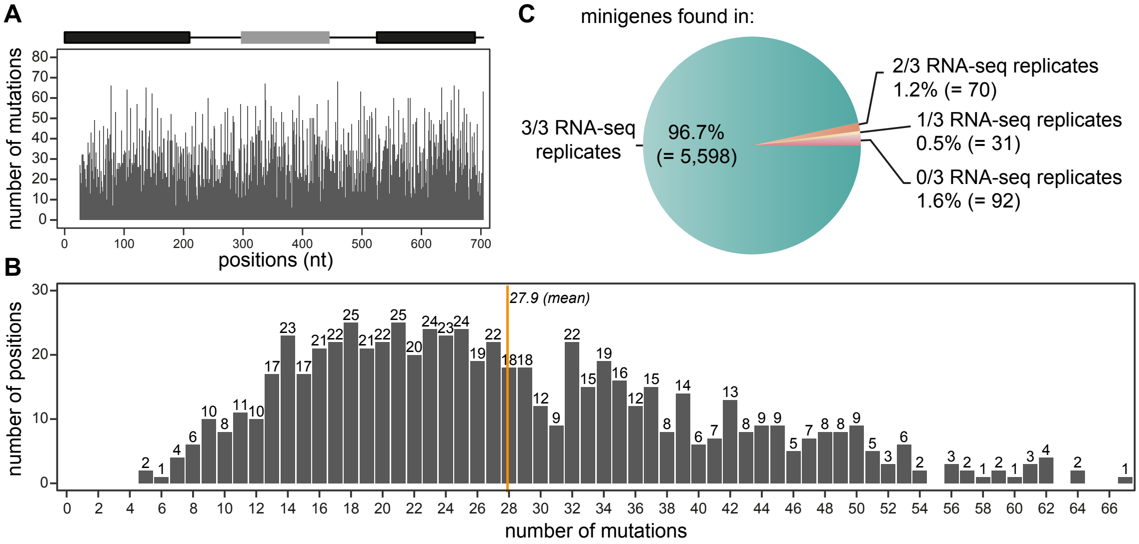


Figure 3.8: Mutations are distributed uniformly across the *RON* minigene and most variants are recovered in the three RNA-seq replicates. (A) Mutations equally distribute along the *RON* minigene positions. The bar chart shows the distribution of 18,948 point mutations across 5,791 minigenes. (B) Each position in the *RON* minigene is covered with at least two mutations in the *RON* minigene library. Positions are covered with an average of 28 mutations (orange line). (C) The majority of minigene variants identified by DNA-seq is present in all three RNA-seq replicates. Pie-chart representing the fraction of the minigenes present in the library (5,791) found in 0-3 RNA-seq replicates.

Mutation calling identified 18,948 point mutations with an average frequency of 3.6 mutations per minigene variant. The mutations were randomly spread across all positions of the *RON* minigene sequence, such that 97% of the positions were mutated at least ten times within the library (average 28 times per position; **Figure 3.8 A and B**). Sanger sequencing of 59 randomly selected minigene variants validated the accuracy of mutation calling since all 169 mutations were retrieved by the mutation calling without any additional false-positives. Thus, mutation effects could be screened across the entire sequence space of the *RON* minigene sequence.

Quantifying *RON* minigene alternative splicing by RNA-seq

To count the different splicing isoforms of the *RON* minigene library, the library was firstly transfected as a pool in HEK293T cells. Then, the resulting alternatively spliced transcripts were quantified by paired-end RNA-seq using a primer combination that allowed unambiguous identification of all canonical isoforms (**Figure** 3.3 and **Figure** 4.5). As previously done, each read pair included a 15-nt barcode to allow matching of the respective counterpart minigene variants.

Three independent RNA-seq replicates were generated and sequenced on Illumina MiSeq (300-nt paired-end), yielding 17–22 million reads per sample, and analysed with a custom Python pipeline similar to DNA-seq (**Figure** 3.7 **B**). Low-quality sequences were removed and the 15-nt barcode sequence was extracted similar to the steps followed in the DNA-seq pipeline.

Moreover, only reads originating from the 5791 minigene variants that were recovered from the DNA-seq library were considered for further analyses. Read pairs for each minigene variant were aligned to the RON wt minigene sequence. Only read pairs conferring splice isoform information (i.e., both mates extended at least 10 nt beyond the constitutive exon boundaries) were kept. Furthermore, all improperly or inconsistently mapped read pairs were removed from the analysis (refer materials and methods). Finally, only minigene variants which were covered by at least 100 remaining read pairs were used further, resulting in 5697, 5645 and 5623 minigene variants detected in RNA-seq replicates 1, 2 and 3 from HEK293T cells, respectively.

97% of minigene variants were found in all three replicates (**Figure** 3.8 **C**). The frequency of each isoform for each minigene variant was calculated as the number of read pairs supporting this particular isoform in relation to the total read pairs for all detected isoforms for this particular minigene variant. While the canonical isoforms of alternative exon (AE) inclusion, AE skipping, full intron retention (IR), first IR, and second IR accounted for 94%

of all splicing products, the remaining 6% were categorised as 'other' isoforms which originated from cryptic 3′ and 5′ splice site usage (**Figure** 3.9 **A and B**). These non-canonical 'other' isoforms can still be a major proportion of the splicing products of certain individual minigene variants. For instance, 3′ splice site disrupting mutations at the downstream constitutive exon 12 trigger activation of a cryptic downstream AG (marked by one asterisk in **Figure** 3.9 **A and F**). The wt minigenes showed considerably small splicing variance, thus indicating the barcode is not a confounding factor for quantifications (**Figure** 3.9 **C**). In contrast, 45% of mutated minigenes were spliced with more than 10% deviation from the wt average AE inclusion, showing that many of the introduced mutations can be studied for causing wide range of splicing effects.

In order to validate the splicing quantifications, splicing of minigene variants containing AE splice site mutations were checked (**Figure** 3.9 **C**). Splice sites abolishment by mutations disrupting the canonical splice signals at the AE, always led to severely decreased near null AE inclusion levels, illustrating that the screening allows reliable detection of strong splicing changes, in addition to small splicing changes. Furthermore, RT-PCR derived splicing measurements of 59 randomly selected minigene variants was performed and compared with the RNA-seq derived quantifications for variants containing the same mutations, respectively (**Figure** 3.9 **D**). The observed correlation confirmed that the accuracy of RNA-seq derived splicing quantification is comparable to individual RT-PCR measurements (Pearson correlation coefficient, $r = 0.95$, p-value < 2e-16). Taken together, the high-throughput mutagenesis screening allows analysis of mutation effects across the entire *RON* minigene region.

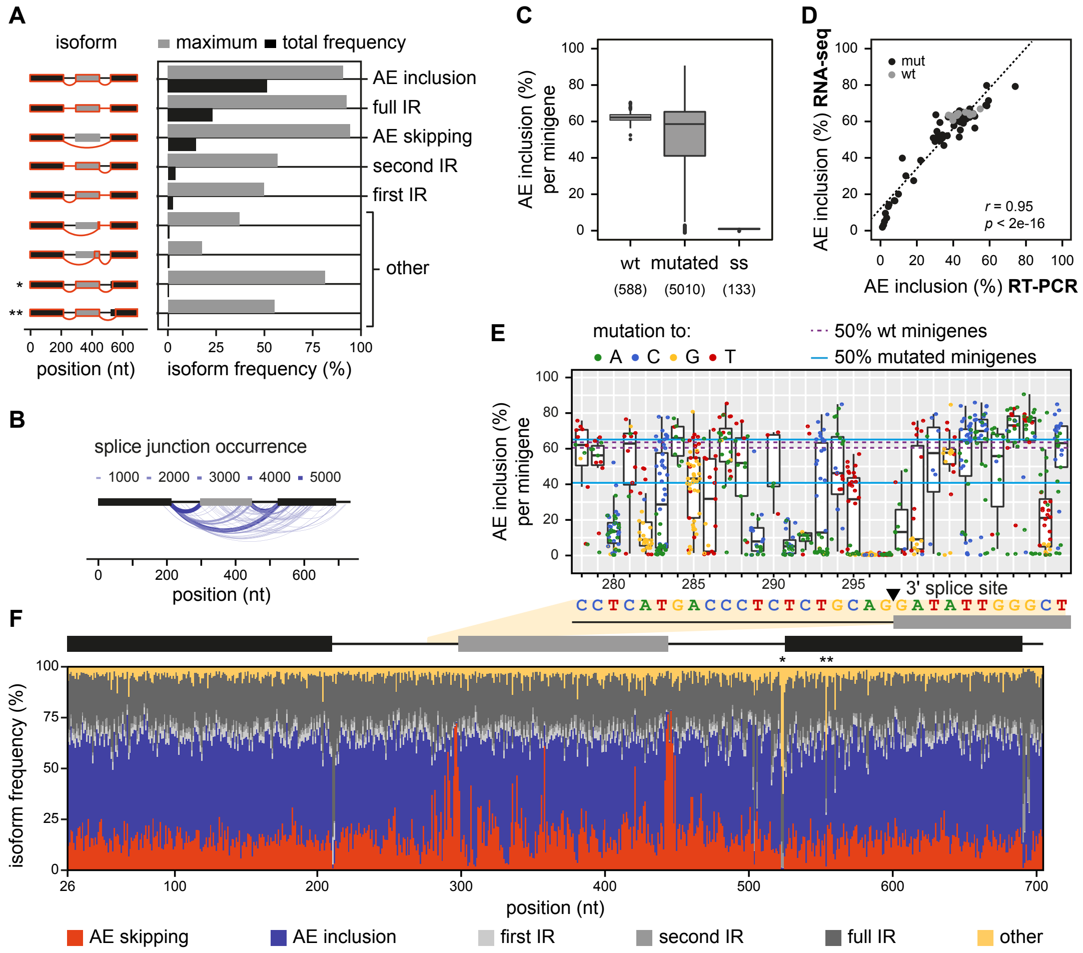


Figure 3.9: High-throughput mutagenesis screening of *RON* minigene library involves quantitative splicing outcomes. (A) Canonical isoforms compose the most frequent splicing variants in the library. Bar plot shows the nine most frequent isoforms sorted by total frequency in the RNA-seq library (black) along with the maximum frequency for an individual minigene variant (grey). (B) Splice junction frequencies. Line thickness and colour represent the number of minigene variants producing the respective splice junction. Only junctions accounting for ≥1% of all junctions for a given minigene variant were considered. (C) Widespread splicing effects of mutated minigene variants. The boxplots show the distribution of AE inclusion frequencies in % for all wild type (wt) and mutated minigenes and a subset of mutated minigene variants with mutations in the splice sites (ss) of *RON* exon 11. Minigenes in each category is given below. Whiskers represent the most extreme values within 1.5x interquartile range. (D) Validation of RNA-seq quantification by correlation with RT-PCR measurements of 59 individual clones from the library (r , Pearson correlation coefficient and associated p-value). (E) Mutation effects around the AE 3′ splice site. Boxplots show the isoform ratios displayed as AE inclusion (%) for minigenes harbouring a mutation at the respective site with colours indicating the inserted base (see legend). The purple and blue lines show the 25%- and 75%-percentiles of AE inclusion of the wt minigenes and the complete library, respectively. (F) Isoform frequencies resulting from mutations along the *RON* minigene. Stacked bar chart shows the median frequency of the six isoform categories for all minigenes with a mutation at a given position. Average of three biological replicates in HEK293T cells. Asterisks show positions that upon mutation cause increased formation of non-canonical isoforms (shown in B).

Dissecting complex mutation effects with linear regression modelling

The splicing outcome of a minigene variant is based on the underlying mutations present in the respective minigene variant. In the minigene library, mutations occurred with an average of 3.6 per minigene variant. This presented a complex situation, where only splicing effects of those variants with only one mutation or previously studied mutations could be effectively studied. In minigene variants with many more mutations, splicing effects of a mutation may depend on the effects of co-occurring mutations in the minigene variant (**Figure 3.9 E**). Such minigene variants present an overlaying of multiple mutation effects and thus, the splicing outcome of individual mutations needs to be pinpointed. In order to precisely infer the splicing effects of each mutation in the context of multiple mutations, a linear regression based modelling approach was undertaken (refer materials and methods section). Towards this, a dynamic splicing model based on splice isoform ratios instead of absolute values was formulated for calculating linear regression. The model involved isoform production versus isoform degradation in relation to the rate of the AE inclusion isoform as a reference. Using isoform ratios instead of absolute frequencies for the input of the linear regression accounted for the non-linearity of mutation effects. The non-linearity of the data was attributed to the natural boundary of isoform frequencies which were bound between 0% and 100%.

Assuming that mutation effects are additive, the linear regression model helped infer single mutation effects (log-fold changes relative to the wt) from the sum of multiple mutation effects in a minigene variant. To confirm that mutations act additively, ~600 single mutation minigenes that allow direct assessment of mutation effects were used for comparison with minigenes sharing combinations of two or three of these single mutations (**Figure 3.10**).

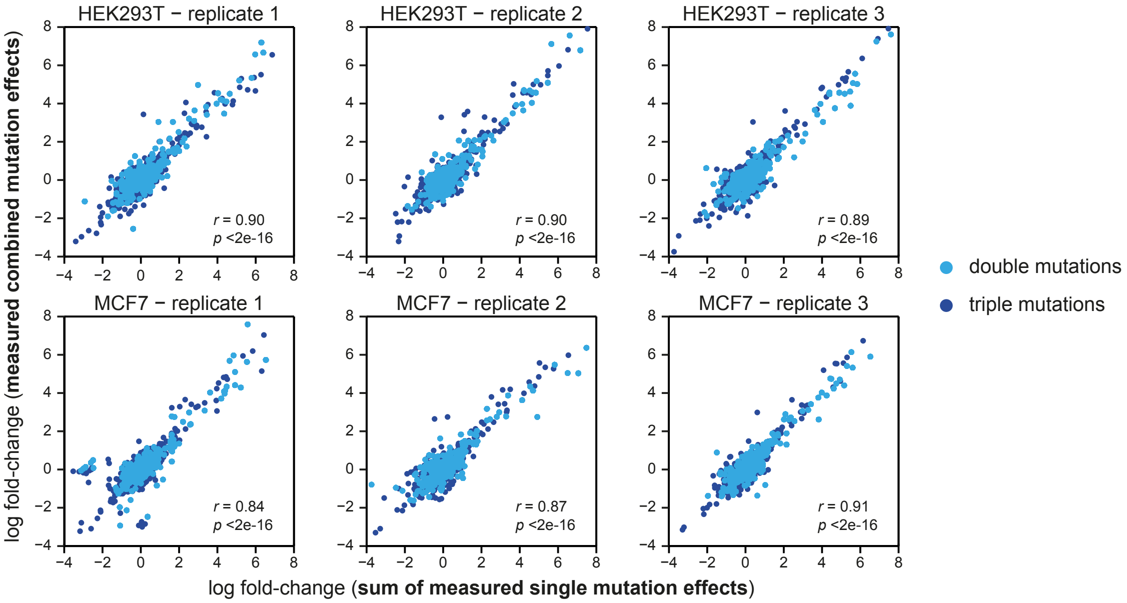


Figure 3.10: Effects of mutations show additive behaviour. Scatterplots showing the correlation of mutation effects (log-fold changes relative to the wt) for the sum of single mutation effects and measured mutation effects of minigene variants containing combinations of two (light blue) or three (dark blue) individually measured mutations. Top and bottom panels show analyses for three replicates in HEK293T and MCF7 cells, respectively. *r* - Pearson correlation coefficients, *p* - associated *p*-values.

In line with the assumption of additive behaviour, the sum of the single mutation effects highly correlated with the combined mutation effects. Hence, the regression approach allowed the accurate modelling of the experimentally measured isoform frequencies for each mutated minigene variant (**Figure** 3.11). As a validation step for the regression model, inferred single mutation effects were compared with RT-PCR derived splicing measurements of minigenes containing single mutations (**Figure** 3.12).

Median isoform frequency measurements offer a simpler approach for quantifying the mutation effects

In conjunction with the modelling approach, the median isoform frequency was calculated using isoform frequencies from all minigene variants harbouring the given mutation

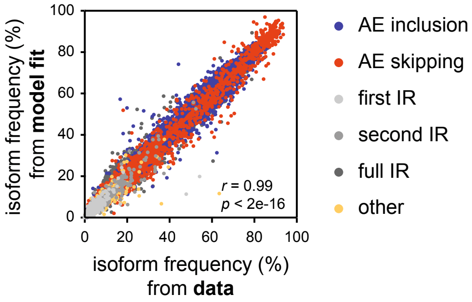


Figure 3.11: Isoform frequencies inferred from linear regression modelling correlate with measured isoform frequencies. Scatterplot showing the frequencies of splice isoforms for combined mutations calculated from the fitted model against the measured data of one biological replicate in HEK293T cells. colours indicate distinct splice isoforms (see legend). r - Pearson correlation coefficient, p - associated p-value.

(**Table** 4.5; refer materials and methods section). As a simple approach, the median was useful in estimating individual mutation effects as an approximation of the splicing outcomes of multiple minigene variants containing the same mutation (**Figure** 3.12). In fact, the median performed in line with regression approach when minigene variants carrying the same mutations occurred at least 10 times (10 plasmids carrying the same mutation). Strikingly, the mutations representing low mutation occurrences (lightly coloured points) showed poor correlation for all isoform types (**Figure** 3.12 bottom panel). This suggested that if sufficient minigene variants carrying the same mutation were present in the library, the median should average out the effect of the accompanying mutations and serve as an effective proxy for depicting the splicing outcome.

The median isoform frequency was then compared to the prediction of the regression model. However, the median-based estimation was outperformed by the linear regression derived single mutation effects (65% of all tests) and ultimately was used only for those mutations that were non-inferred by the model. Depending on the replicate and cell line, between 3-9% of the unique combinations of mutations were excluded from the calculations based on two criteria. As detailed in the methods section, we excluded certain

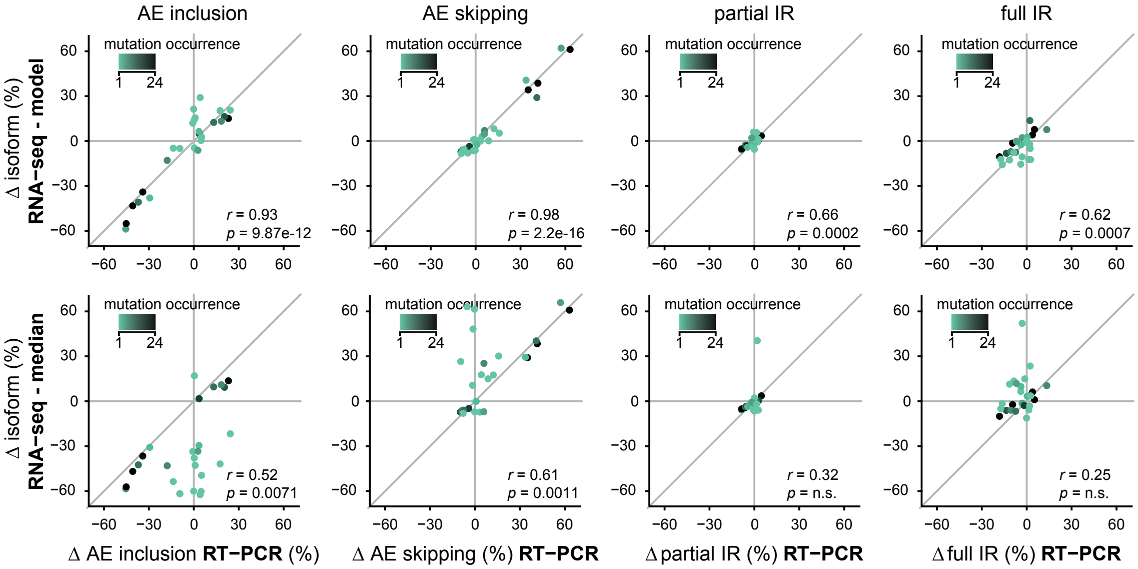


Figure 3.12: Linear regression provides an accurate estimation of single mutation effects compared to the median isoform frequency of minigenes containing the same mutations. Effects of mutations that rarely occur in the library (colour coded) correlate better with the model-inferred than the median-based estimates. Scatterplots compare the model-inferred (top row) and the median-based (bottom row) estimations of single mutation effects relative to wt (y-axes) to semi-quantitative RT-PCR measurements (x-axes) of targeted minigenes harbouring the respective single point mutations, insertions and deletions. Separate plots are shown for the different splice isoforms. First IR and second IR were summed up as *partial IR*, since these isoforms cannot be discriminated in the RT-PCR. Pearson correlation coefficients *r* and associated *p*-values are provided in each panel.

minigenes from the linear regression model that harboured either (i) combinations of two splice site mutations or (ii) that showed a strong activation of cryptic splice sites ('other'; refer materials and methods; **Table** 4.5). The effects of the rest 94-97% of the mutations present in the library could be assessed by linear regression model that covered almost the entire length of the minigene (all but 3-4 out of all 679 nucleotides in the minigene; **Figure** 4.1).

In summary, the model predictions for HEK293T cells allowed estimation of mutation effects for ~1,800 single point mutations for the five canonical isoform categories. For those mutations that fell in the excluded categories, the median estimation method was used to assess the splicing outcome. Furthermore, in cases where sufficient number of plasmids contain the given mutation, the median approach was shown to be well suited as a first-approximation approach to assess the splicing impact of a mutation in the *RON* minigene library (~3.6 mutations per minigene variant) .

The *cis*-regulatory landscape of *RON* alternative splicing

Among the ~1800 mutations, splicing was significantly altered in at least one isoform by 778 mutations (referred as splicing-effective mutations; ≥ 5% change in isoform frequency, 5% false discovery rate, FDR **Figure** 3.13). Evidently, mutations that disrupted the splice sites at the alternative exon completely prevented AE inclusion (**Figure** 3.13 **B**). In contrast, effects of mutations at the poly-pyrimidine tract upstream of the AE 3' splice site were base-specific: While pyrimidine transitions showed very little effect, transversions to purines reduced inclusion drastically e.g. position 290 (**Figure** 3.13 **B**). Similarly, degenerate positions of the branch point consensus of YUNAY sequence (positions 279 (Y), 281 (N), and 283 (Y)) were more resistant to mutation than the strictly conserved U or A at positions 280 and 282, respectively (Gao et al., 2008) (**Figure** 3.13 **B**).

We saw that mutations at the splice sites of the flanking constitutive exons showed in-

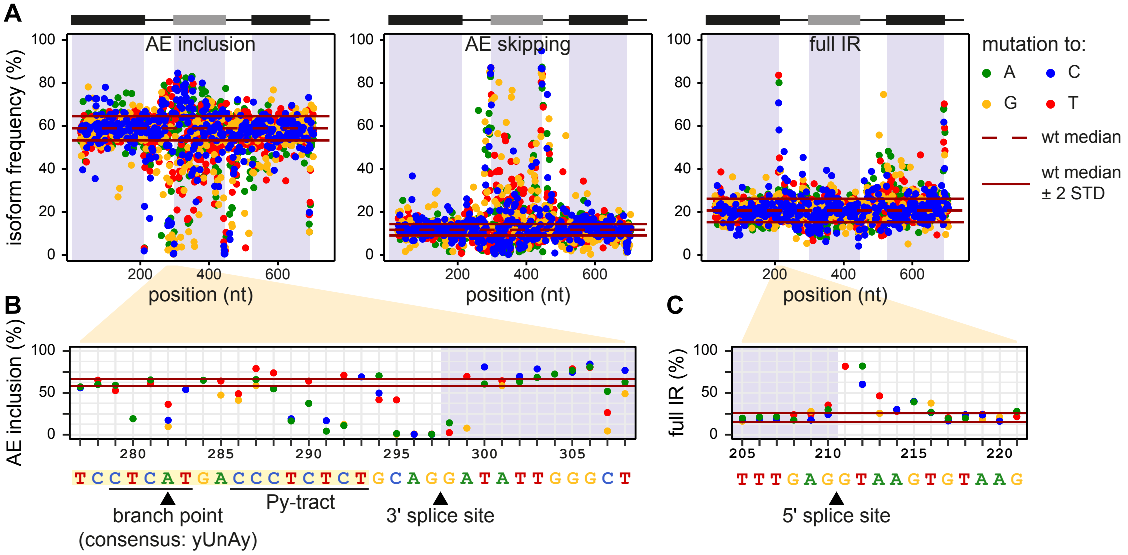


Figure 3.13: The regulatory effects of 1800 point mutations on *RON* exon 11 splicing. (A) Modelling derived quantifications of single point mutations on AE inclusion- (left), AE skipping- (middle), and full IR isoforms (right) across the *RON* minigene positions in HEK293T cells. (B) Mutation effects on AE inclusion surrounding the AE 3′ splice site. Horizontal black lines and arrowheads point to core splicing signals, including branch point, polypyrimidine tract (Py-tract) as well as the 3′ splice site. (C) Mutation effects on full IR surrounding the 5′ splice site (black arrowhead) of the upstream constitutive exon. The **Figure 3.13 B** shows the same region as in **Figure 3.9 E**.

creased full IR and lowered AE inclusion isoforms (**Figure 3.13 C**). Mutations at the 5′ splice site of the downstream constitutive exon affected AE inclusion, suggesting that exon definition at flanking exons may affect the splicing regulation of the alternative exon. The possible role of exon definition at the upstream constitutive exon to splicing regulation of the AE could not be assessed, since the 3′ splice site of the upstream constitutive exon was not present (**Figure 3.1**). This confirms previous knowledge on the possible effects of distal exons in regulating the splicing of an alternative exon (Barash et al., 2010).

To check correlation of AE inclusion levels with the splice-site strength of mutations of both 5′ and 3′ splice sites, *in-silico* predictions of splice site strength using the target in-

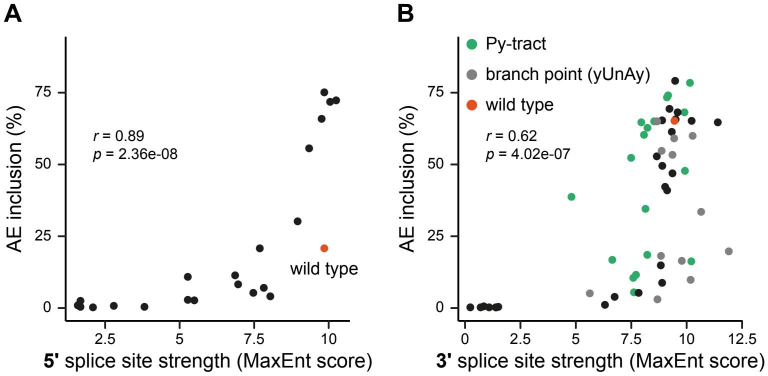


Figure 3.14: Mutation effects at both the splice sites of AE correlate with splice site strength predictions in HEK293T cells. (A, B) AE inclusion correlates with the *in-silico* predictions (MaxEntScan) of 5′ (A) or 3′ (B) splice site strength of respective sequences. Spearman correlation coefficients *r* and associated *p*-values are provided.

put sequences (refer materials and methods) were performed using MaxEntScan software (Yeo et al., 2004). *In-silico* predictions of splice site strength captured the mutation effects at the 5′ splice site of *RON* exon 11 (Spearman correlation coefficient, *r* = 0.89, *p*-value = 2.36e-08; **Figure 3.14 A**). However, predictions for 3′ splice site strength were less correlated (Spearman correlation coefficient, *r* = 0.62, *p*-value = 4.02e-07; **Figure 3.14 B**), showing that splice site strength predictions are less accurate at this site. This may be explained by the presence of a higher density of regulatory signals at the 3′ splice site compared to the 5′ splice site.

A highly dense arrangement of splicing-effective positions, i.e. positions with at least one splicing-effective mutation, was found in the alternative exon as 91% of all positions of the *RON* exon 11 (134/147 nt) were splicing-effective. In comparison, the upstream and downstream flanking introns encompassed 77% and 82% splicing-effective positions, respectively. In the constitutive exons 10 and 12, 50 and 60% of the positions were involved in the splicing regulation (**Figure 4.2**). The vast splicing-regulatory information, reflected in the overwhelming presence of splicing-effective positions within the *RON* regions is highlighted here.

In summary, the dense packaging of splicing-effective positions within the *RON* minigene region is observed not only within the alternative exon but also in atleast 50% of the positions within the flanking introns and constitutive exons. Particularly, the widespread occurrence of splicing effective positions across exons signifies the important role of splicing regulation at exonic positions.

Evolutionary conservation of splicing-effective mutations

Evolutionary selection pressures positively affect the conservation of splicing-effective positions (Xing et al., 2006). In line with this observation, the evolutionary conservation (PhyloP scores) between splicing-effective and non-effective positions in *RON* was compared (**Figure** 3.15). In introns, splicing-effective positions showed higher evolutionary conservation when compared to non-effective intronic positions. Furthermore, intronic splicing-effective positions which exhibited strong splicing changes (≥20%), such as core-splicing elements, showed conservation levels as high as exonic splicing-effective positions (**Figure** 3.15). In exons, splicing-effective positions showed the same evolutionary conservation as non-effective positions. As exonic sequences code for amino acids, which are functionally constrained by selection pressures, the splicing-function of exonic positions becomes second priority and is overridden.

Previously, synonymous mutations were suggested to alter protein sequences via splicing regulation (Mueller et al., 2015). In order to characterise the contribution of synonymous mutations in splicing regulation, absolute mutation effects were compared between non-synonymous mutations and synonymous mutations. As a result, synonymous mutations in the *RON* exons 10, 11, and 12 are observed to mediate splicing regulation with similar effect sizes and frequencies as non-synonymous mutations (**Figure** 3.16), in line with the *FAS*/*CD95* exon 6 study (Julien et al., 2016).

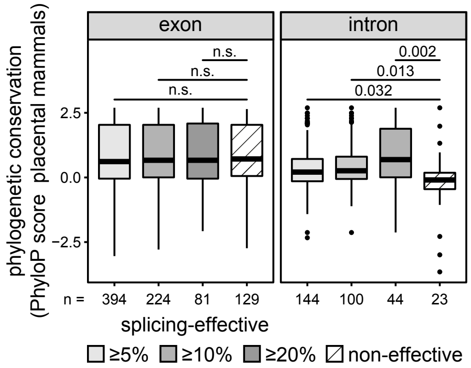


Figure 3.15: Splicing-effective positions are more conserved than non-effective positions in introns (MCF7 cells). Splicing-effective positions within introns are significantly more conserved (PhyloP score across 46 placental mammals) compared to splicing non-effective positions. This effect is not observed in Exonic positions. Splicing-effective positions were categorised according to cut-offs of ≥5%, ≥10%, or ≥20% change in isoform frequency. Number of positions counted in each box are indicated below. *p*-values correspond to two-sided Mann-Whitney-U test. n.s., not significant.

In total, 135 splicing-effective mutations were found to be both synonymous and effective in mediating splicing (**Table** 4.2). In line with previous studies, synonymous mutations contribute to disease by altering the splicing outcome and consequently, the protein function (Xing et al., 2006; Shabalina et al., 2013; Mueller et al., 2015).

In order to assess if splicing-effective mutations occur in a low frequency among human populations, allele frequency data was used from Exome Aggregation Consortium (ExAC) which consisted of exonic variations genotyped from large scale exome sequencing data. This catalogue of genetic diversity showed that the human exome contained an average of one variant every eight bases, thus implying widespread recurrence of mutations in humans (Lek et al., 2016). However, there was no significant difference between splicing-effective mutations and non-effective mutations, revalidating that the evolutionary selective pressure generated by the splicing-function of a nucleotide in an exon is overruled by protein-coding constraints (**Figure** 3.17).

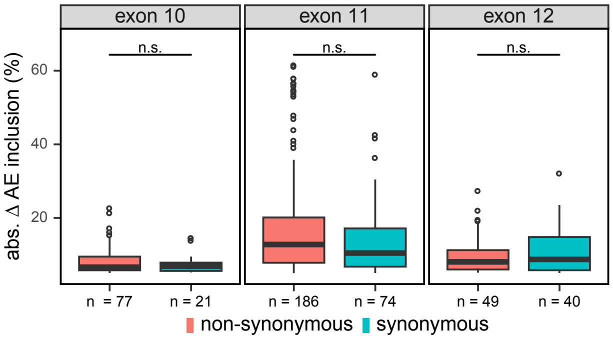


Figure 3.16: Both synonymous and non-synonymous mutations have similar effect sizes. Boxplots show the distribution of absolute changes in AE inclusion in HEK293T cells for synonymous and non-synonymous mutations across the exons of the *RON* minigene. Number of mutations in each box are indicated below. *p*-values from two-sided Mann-Whitney-Wilcoxon test. n.s., not significant.

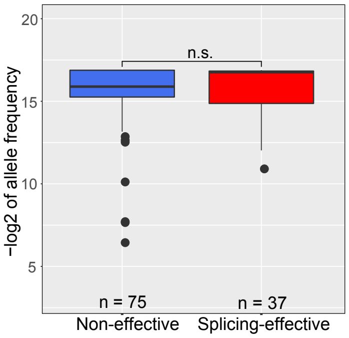


Figure 3.17: Splicing-effective and non-effective mutations show similar allelic proportions. Splicing-effective mutations when compared to non-effective mutations showed no significant changes in terms of their ExAC-derived allele frequencies as shown in the boxplots. Number of positions counted in each box are indicated below. Two-sided Mann-Whitney-U test. n.s., not significant.

Taken together, the alternative splicing of *RON* exon 11 is regulated by many intronic and exonic positions and the significance of splicing regulatory positions within introns is recapitulated by their evolutionary conservation. Moreover, stronger selection pressure is observed on protein-coding constraints of sequences in comparison to splicing constraints. Moreover, splicing regulatory mutations were also shown to be synonymous mutations. Since these mutations do not alter the sequence of the encoded protein, they might erroneously be interpreted as non-pathogenic, however, they may impair protein function through alternative splicing.

Comparisons to *in-silico* splicing effect and G-quadruplex predictions

In order to check if *in-silico* splicing effect predictions were in line with those quantified by the screen, SPIDEX (SPANR) splicing predictions were retrieved (Xiong et al., 2015). In the SPIDEX dataset, mutations predicted to disrupt *RON* exon 11 splicing belonged within the *RON* exon 11 region, i.e. only those mutations present within the *RON* exon 11 were predicted to strongly influence the splicing regulation of the *RON* exon 11. Among this set of mutations, only 12 mutations (10 coinciding with the screen) exhibited strong splicing changes (dPSI > 20%) and were far fewer than those identified by the screen (**Figure 3.18 B**; **Figure 4.2**). These mutation numbers are fewer in nature compared to the screening approach, because the SPIDEX dataset consisted of predictions for only the strongest splice-regulatory effect of a mutation. In contrast, from the screening approach, 47% (HEK293T) and 51% (MCF7) of the splicing-effective mutations trigger changes of greater than 10% in at least one splice isoform (363 out of 778 in HEK293T, and 521 out of 1022 in MCF7), and 1/5 of them exceed 20% (136 in HEK293T and 189 in MCF7, respectively). Importantly, intronic and other exonic (exon 10 and 12) mutations, identified by the screening approach, too show stronger splicing effects on *RON* exon 11 inclusion. Hence, a large

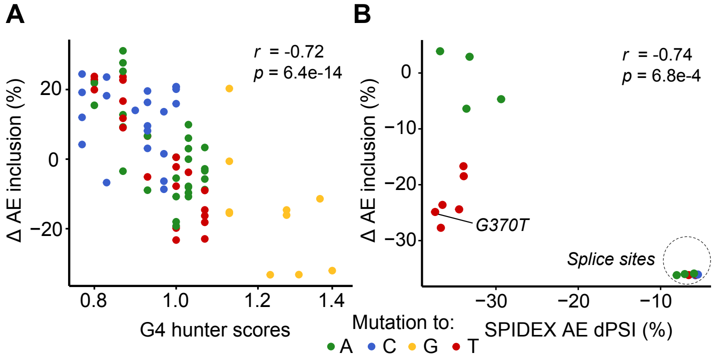


Figure 3.18: *In-silico* predictions show fewer strong mutation effects compared to the screen. (A) G4 prediction scores of the G-quadruplex region in *RON* AE correlates with splicing changes obtained from the screen. (B) SPIDEX predictions are limited to the strongest effect of a mutation and partially correlate to splicing effects obtained from the screen. To-A mutations (left top corner) show opposing splicing changes in the screening approach compared to SPIDEX predictions. Coloured points represent mutations to different bases. Spearman correlation coefficients *r* and associated *p*-values are provided.

number of potent mutations identified by the screening approach, substantially alter the splicing outcome for *RON* exon 11.

Furthermore, the splicing regulatory role of mutations like G370T is an important discovery of the screening approach (refer **Figure 3.21**), that is also confirmed by SPIDEX predictions albeit higher in magnitude by ~15%. Nevertheless, SPIDEX predicted to-A (mutation from wt base to adenosine) mutations show negligible effect in comparison to those inferred by the screening approach (**Figure 3.18 B**; top left corner), which may be attributed to the weak predictive power of the tool, especially when predicting non-splice sites (Soukarieh et al., 2016; Grodecká et al., 2017; Moles-Fernández et al., 2018). Hence, our screening approach can be useful in updating and validating *in-silico* based predictive models.

In order to assess if a possible G-quadruplex mediates *RON* splicing regulation, G-quadruplex propensity scores for the putative G-quadruplex (refer materials and methods) were retrieved from G4hunter tool (Bedrat et al., 2016). Then, the propensity scores were compared to the splicing effects at the respective sites. As expected, mutations to-guanine (right bottom corner) show high splicing effects (>20% change in IF) and G4hunter scores (> 1) where as mutations disrupting a guanine have low G4hunter scores (left top corner; r = -0.72, p = 6.4e-14). Since the *RON* alternative exon cluster has the highest density of splicing-effective mutations, the G-quadruplex might have a role in mediating splicing regulation of *RON* (**Figure** 3.18 **A**; refer **Figure** 4.2). However, further studies on G-quadruplexes are required to identify the possible role of these structural elements in modulating alternative splicing (refer discussion).

In summary, *in-silico* tools like SPIDEX predict fewer strong mutation effects when compared to the screening approach. Mutation effects quantified by the screening approach correlate with G-quadruplex propensity scores and may help assess the role of structural elements in mediating alternative splicing. Furthermore, the mutation effects from the screening approach may complement *in-silico* predictions and help in further validation studies.

Clinical relevance of *RON* mutations in cancer

In tumours, *RON* proto-oncogene is involved in progression to invasive behaviour through elevated levels of the AE skipping (RONΔ165) isoform (Collesi et al., 1996; Chakedis et al., 2016). In line with this idea, the minigene library was transfected as a pool into MCF7 cells (human breast cancer cell line) and the mutation effects were quantified as previously done. Consequently, increased AE skipping levels of the library were seen in MCF7 compared to the HEK293T cells (**Figure** 3.4 **A**). Moreover, overall mutation effects were reproducible when expressed in fold-change (in relation to mutant against wild-type) of the isoform ratios (**Figure** 3.19). The distribution of splicing-effective mutations and positions showed similarities between the two cell lines (refer **Figure** 4.2

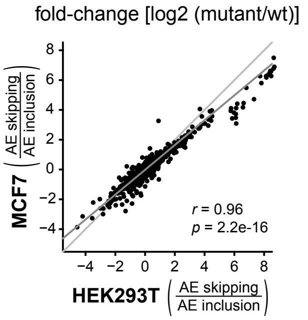


Figure 3.19: Mutation effects are similar between MCF7 and HEK293T cells. Scatterplot shows changes in splice isoform ratios predicted for mutations analysed in MCF7 against those in HEK293T cells. Light and dark grey lines correspond to diagonal and linear regression line, respectively. *r*, Pearson correlation coefficient and associated *p*-value.

and **Figure** 4.3), implying that regulation of *RON* splicing is mostly consistent between the two cell lines.

In order to identify *RON* disease-relevant mutations in cancer, the COSMIC (Catalogue of Somatic Mutations in Cancer) database was used. COSMIC compiles somatic mutations comprehensively from several publications and databases. The *RON* minigene encompassed 33 COSMIC entries, of which 20 intersected with splicing-effective mutations deduced by the screening approach (**Figure** 3.20). Strikingly, 7 synonymous mutations (with respect to the encoded RON protein) were observed in this set. This suggested that synonymous mutations may play a harmful role by perturbing alternative splicing regulation in cancer.

Next, the disease relevance of these mutations was further assessed by analysing splicing quantification from TCGA (The Cancer Genome Atlas; `https://cancergenome.nih.gov/`) cancer patient datasets. In contrast to the COSMIC database, TCGA additionally lists RNA-seq datasets (i.e. splicing quantifications) associated with the mutations, thus comparisons of *in-vivo* mutation effects with those of the screening approach were made possible. In total, 51 mutations within the *RON* minigene region were listed from 19 dif-

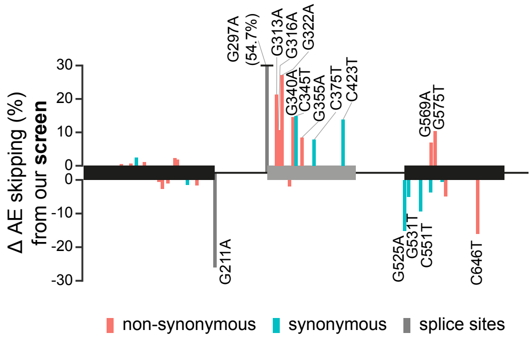


Figure 3.20: *RON* cancer somatic mutations show enrichment of splicing-effective mutations. Bar plot shows the AE skipping changes quantified from the screening in MCF7 cells for 33 mutations that were present in COSMIC database within the *RON* minigene region. Splicing-effective mutations are labelled above relevant bars. Orange, blue and grey indicate non-synonymous, synonymous and splice site mutations, respectively.

ferent cohorts (representing different cancer types in 153 patients). Only those mutations that occur specifically in the tumour but not in the corresponding normal samples (**Table 4.4**) were retrieved. The change of AE skipping levels between mutation-harbouring tumour tissue and tumour tissue without the corresponding mutations were compared with those of the respective mutations from the mutagenesis screening (**Figure 3.21**).

The strongest mutation effects were caused by G370T and G297A mutations in Head-Neck Squamous Cell Carcinoma and Thyroid Carcinoma, respectively (**Figure 3.21 B**). Both these mutations led to increased AE skipping levels, either via disrupting the AE 3′ splice site (G297A) or by alteration of a putative binding element for QKI and HNRNPL in *RON* exon 11 (G370T). The observed correlation (Pearson correlation coefficient, $r = 0.62$, p-value = 4.8e-05) indicated that data from the screening allowed the estimation of many strong *in-vivo* mutation effects in cancer (refer discussion). In summary, the mutagenesis screening is helpful not only in the assessment of the pathophysiological relevance of cancer mutations but also in deducing the splicing disrupting role of non-synonymous mutations in cancer.

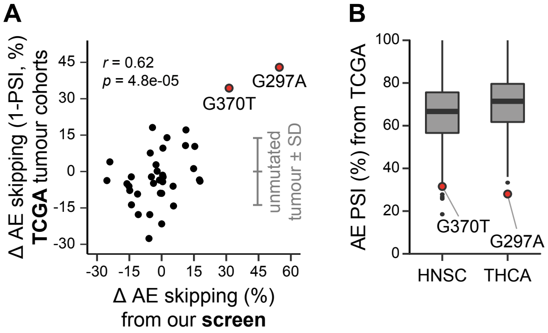


Figure 3.21: *RON* splicing in mutation-bearing cancer patients correlates with mutation effect quantifications from the screen. (A) Scatterplot comparing AE skipping changes of mutations found in TCGA database with screening derived quantifications. Highlighted mutations are detailed in (**B**). Grey lines indicate mean and standard deviation of unmutated tumor samples. Only mutations from cohorts with more than 24 supporting reads on average were used in **Figure 3.21 A** (**Table 4.4**). (B) Boxplots showing AE PSI distribution in Head-Neck Squamous Cell Carcinoma (HNSC) and Thyroid Carcinoma (THCA). The highlighted mutations strongly reduce AE PSI levels.

Identification of *trans*-acting factors mediating *RON* alternative splicing regulation

As a result of the screening approach, the set of splicing-effective mutations contained 1,022 mutations in MCF7 cells (**Figure 4.3**), implying that multiple *cis*-regulatory elements exist along the minigene. In order to find the *trans*-acting factors that regulate *RON* splicing by targeting these *cis*-regulatory elements, the following approaches were considered:

Knockdown of *trans*-acting factors and subsequent analysis of *RON* splicing via RT-PCR and screening approach:

A siRNA-mediated KD of a *RON* splicing regulator should affect *RON* isoform levels, which can quantitatively be measured by RT-PCR. In order to select candidate RBPs, a previously published large-scale RBP KD screen was of immense use (Papasaikas et al.,

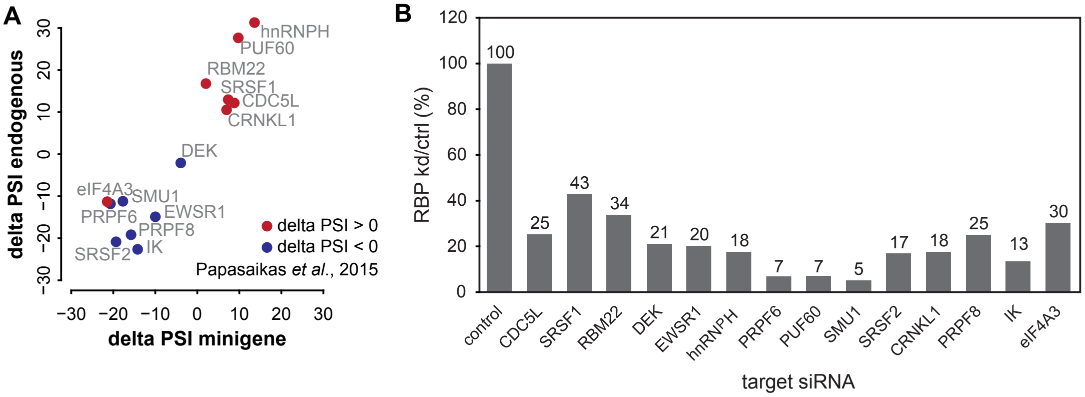


Figure 3.22: Splicing activators and repressors regulate *RON* exon 11 splicing. (A) Scatterplot comparing splicing changes induced by KD of selected RBPs between endogenous *RON* and *RON* minigene splicing. Red and blue colour codes for positive and negative splicing changes measured by (Papasaikas et al., 2015), respectively. PSI percent-spliced-in. (B) Corresponding KD efficiencies analysed by RT-qPCR.

2015). Here, the authors measured the KD effect of ~250 RBPs on several splicing events including *RON* exon 11 in HeLa cells. Putative regulators from this study were selected based on their strong and replicate-consistent effect on *RON* splicing. In total, the splicing change induced by KD of 14 selected RBPs was compared between endogenous and *RON* minigene splicing (**Figure** 3.22 **A**) and successful KDs were confirmed by RT-qPCR analysis (**Figure** 3.22 **B**). The observed splicing changes were in the same direction and consistent with the results of Papasaikas et al. (2015) study, except for the KD of *eIF4A3* that showed opposing splicing effects. Furthermore, the KD of *DEK* had almost no effect on neither endogenous *RON* or *RON* minigene splicing, while it caused strong reduction of AE inclusion in the Papasaikas et al. (2015) study. Taken together, these results indicate that *RON* splicing is extensively regulated by several splicing activators and repressors, with SRSF2 and HNRNPH as the strongest activator and repressor, respectively.

The splicing patterns of each minigene plasmid was analysed in the context of the 5 most effective RBP KDs, via the high-throughput screening approach. These RBPs comprised

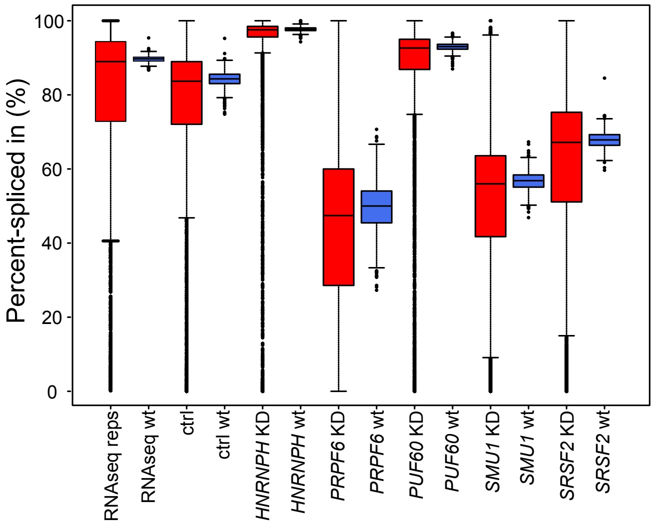


Figure 3.23: Widespread effects of mutations under RBP KDs in the context of the *RON* minigene library. Boxplots show the Percent-spliced in (PSI) of top 5 effective RBP KDs in the context of minigene library compared to wt *RON* plasmid pool and the *RON* ctrl (control)/RNAseq (RNASeq reps) minigene library.

of both core splicing factors and auxiliary RBPs namely HNRNPH, SRSF2, PRPF6, PUF60 and SMU1. Expectedly, the splicing outcome in PSI of plasmids under KDs compared to the wt plasmids and control library, shows widespread effects due to constituent mutations (**Figure** 3.23).

The underlying idea, referred to as synergy analysis (more details in the following section), is that minigene variants with mutations that change binding of a regulatory RBP will behave differently upon the KD compared to the remainder of the minigene library. For instance, minigene variants that lost an RBP binding site will no longer respond to the KD of this particular factor. In such scenarios, minigene variants with mutations in the RBP binding site would also deviate from the remainder of the library in the KD of other complex components. In order to understand this behaviour, the splicing outcome of the control minigene library (ctrl AE inclusion) was compared to the knockdown effect of the library. The knockdown effect is given by the difference in splicing outcome of the KD library (KD AE inclusion) in relation to the control library (ctrl AE inclusion), which is given by

ΔKD AE inclusion = KD AE inclusion - ctrl AE inclusion

Expectedly, in the case of KD of repressors HNRNPH and PUF60, the *RON* minigene library exhibits an overall increase in AE inclusion levels from the baseline and the rest exhibit an overall decrease in AE inclusion (**Figure** 3.24 and **Figure** 3.23). Among these 5, HNRNPH was chosen as an initial candidate for further analyses.

Screening approach and *in-silico* binding site predictions identify RBPs regulating *RON* at SRBS:

Using predictions from the ATtRACT database, RBPs and their binding sites across the *RON* minigene were identified and catalogued (Giudice et al., 2016). Furthermore, using information from the screening approach, RBP motif predictions were filtered for the presence of at least 60% splicing-effective positions which were termed as splice-regulatory binding sites, SRBS. This resulted in a list of 76 putative regulators of *RON* along with their quantifications in terms of splicing changes, suggesting that *RON* splicing is extensively controlled by multiple RBPs. The analysis also recapitulated previously published binding sites of *RON* regulators HNRNPH and SRSF1 (Ghigna et al., 2005; Lefave et al., 2011; **Figure** 3.25). In order to prioritise among the 76 RBPs, our data was overlaid with the large-scale knockdown (KD) screen which tested the KD effect of 31 RBPs from the list of 76 RBPs on *RON* exon 11 splicing (Papasaikas et al., 2015). Among the new list of 31 RBPs, 17 RBPs substantially affected *RON* splicing with SRFS2 and HNRNPH emerging as the strongest activator and repressor, respectively (**Figure** 3.25).

Correlation of RBP expression profiles and *RON* splicing levels in tumour samples:

TCGA tumour sample datasets containing *in-vivo* RBP expression profiles and *RON* splicing data were used for mining novel RBP regulators of *RON* splicing in tumours. In par-

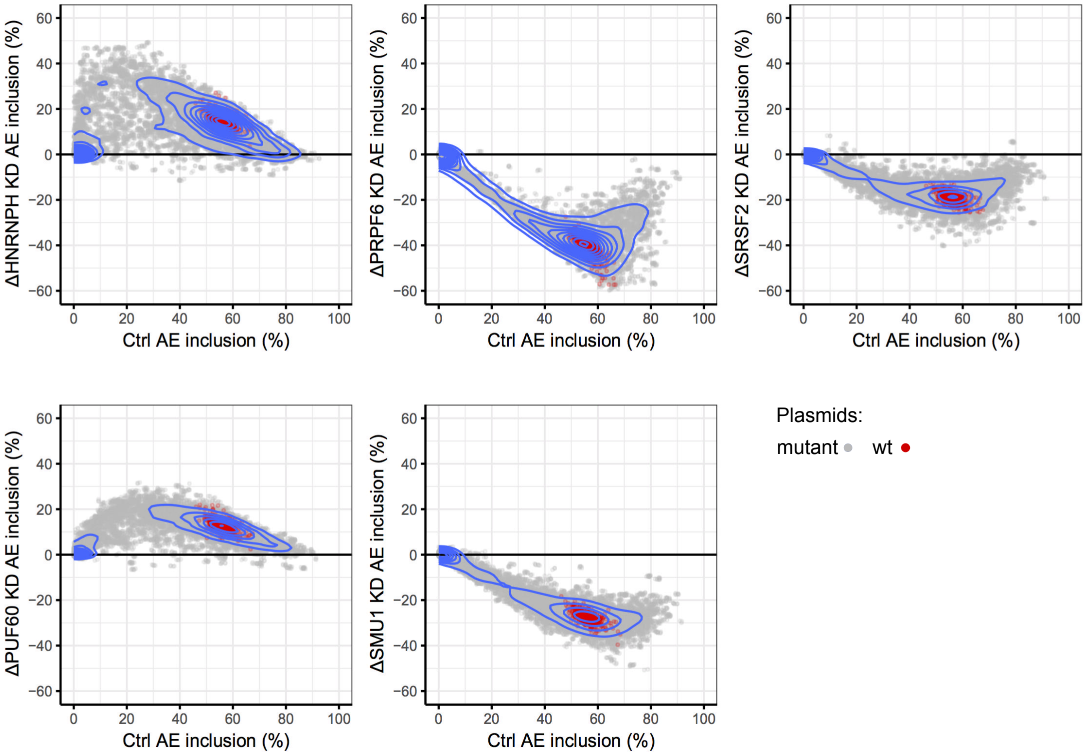


Figure 3.24: The synergistic behaviour of plasmids under RBP KDs in the context of the *RON* minigene library. The scatterplots show the effect of the top 5 RBP KDs affecting the splicing outcome of *RON* minigene library. The splicing effects of individual plasmids are represented by gray dots, exhibiting an increased AE inclusion from the baseline in the case of *HNRNPH* KD and *PUF60* KD and decreased AE inclusion in the other RBP KDs. The contours in blue show maximum concentration of plasmids and the concentric range of most plasmids. In the middle of contours, the wt plasmids are represented in red.

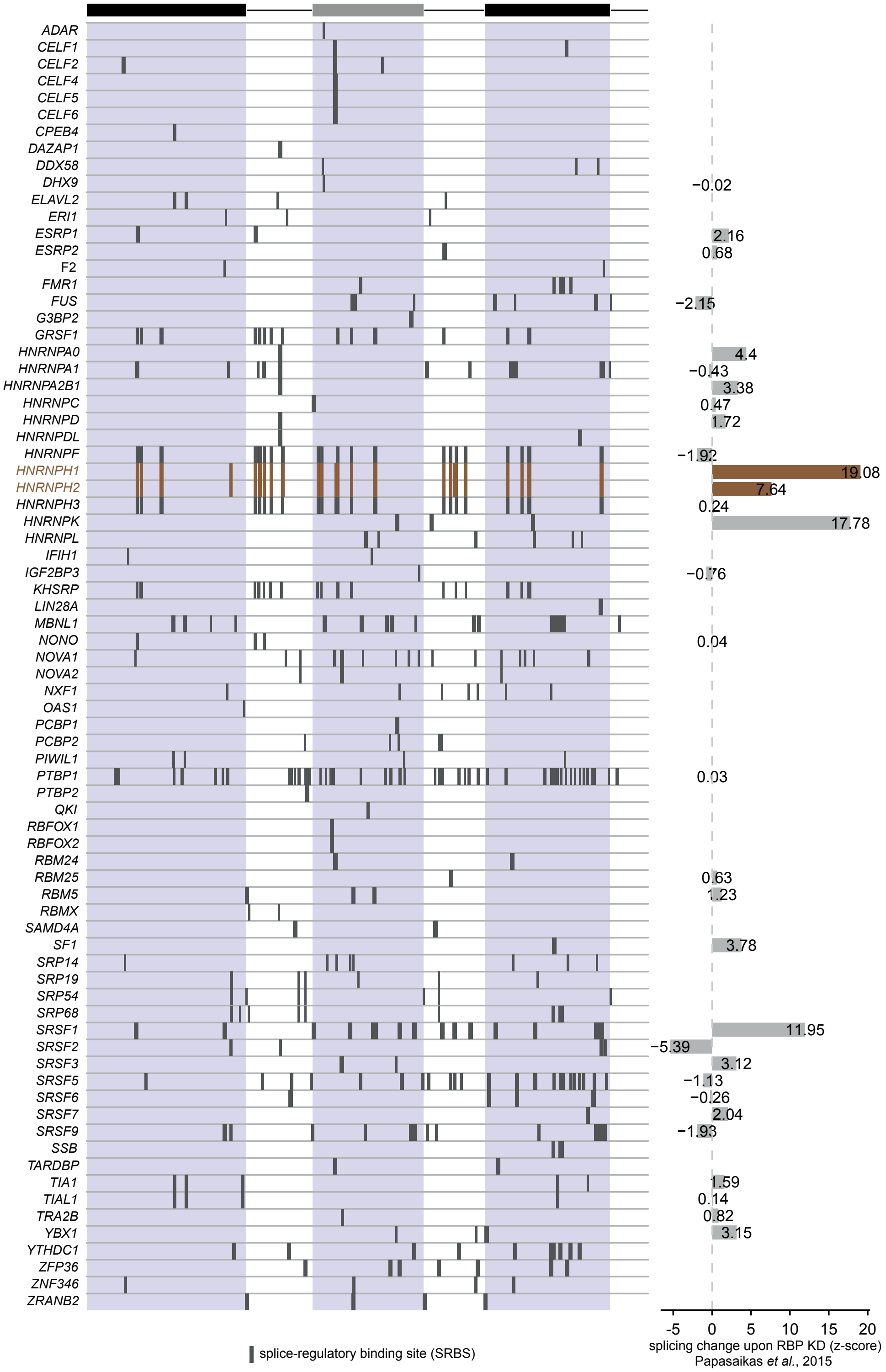


Figure 3.25: *RON* splicing regulators identified using data from *in-silico* binding site predictions and the screening approach. Boxes indicate locations of splice-regulatory binding sites (SRBS), i.e. sites of *in-silico* predicted RBP motifs that contain at least 60% splicing-effective positions (see methods). SRBS for HNRNPH1 and HNRNPH2 are highlighted in brown. Bar chart on the right provides the splicing change measured upon KD of the indicated RBPs from previously published data (in z-scores; z-scores >1 imply increased AE inclusion upon KD, z-scores <1 imply decreased AE inclusion upon KD).

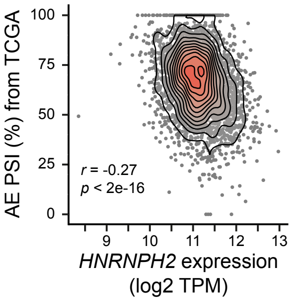


Figure 3.26: *HNRNPH2* expression correlates with *RON* AE PSI levels across TCGA tumour samples. Density scatter plot shows *HNRNPH2* expression (in transcripts per million, TPM) and *RON* exon 11 PSI across all TCGA tumour samples. *r*, Spearman correlation coefficient and associated *p*-value.

ticular, the expression of 190 RBPs and corresponding *RON* AE PSI levels were analysed in 4,514 cancer patient samples for correlation (**Table** 4.3). Strikingly, the strongest association between RBP expression levels and *RON* AE PSI was seen for *HNRNPH2* (**Figure** 3.26), which upon KD also showed stronger splicing changes in the RBP KD screen (**Table** 4.3). As in other cases of splicing regulation, *RON* splicing is also regulated *in-vivo* by multiple RBPs with varying magnitudes of association.

In summary, *RON* splicing is extensively regulated by multiple RBPs in varying and bidirectional magnitudes and HNRNPH is crucial for *RON* splicing both *in-vivo* (from tumour samples, refer **Figure** 3.26) and in cells (**Figure** 3.22 **and** 3.25). Hence, HNRNPH acts as a valid target RBP to study splicing regulation in *RON*.

Position-dependent regulation of *RON* splicing by HNRNPH

HNRNPH is predicted to bind at 22 different non-overlapping SRBS spread across the *RON* minigene (**Figure** 3.25). These SRBS arrange into five distinct clusters across intronic and exonic regions, with each comprising of 3-5 SRBS. In order to narrow down on specific binding interactions of HNRNPH with the *RON* minigene, individual-nucleotide resolu-

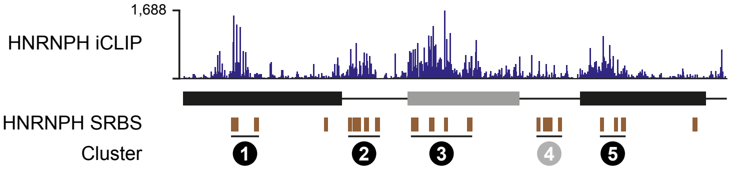


Figure 3.27: HNRNPH binds to both exons and introns across the *RON* minigene. Binding of HNRNPH to the predicted HNRNPH SRBS is validated for four out of five HNRNPH SRBS clusters. Bar diagram shows HNRNPH iCLIP crosslinking events per position across the *RON* minigene (top). Brown boxes indicate HNRNPH SRBS that were assigned to five clusters (circled numbers; bottom).

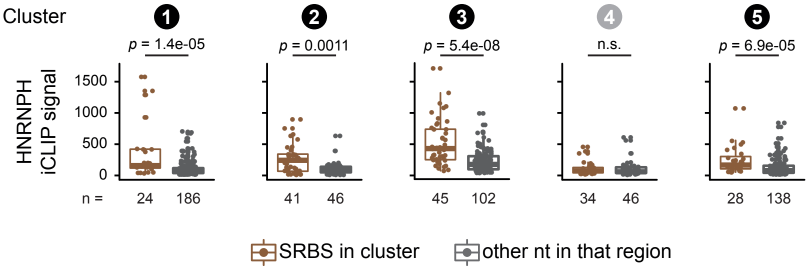


Figure 3.28: HNRNPH iCLIP signal is enriched in four out of five HNRNPH SRBS clusters. Boxplots compare the HNRNPH crosslink events within HNRNPH SRBS ±2 nt (brown) with HNRNPH crosslink events for other positions within the same intron/ exon region (grey). Number of positions per boxplot is given below. p-values correspond to two-sided Wilcoxon Rank-Sum test.

tion cross-linking and immunoprecipitation (iCLIP) was done in HEK293T cells in replicates (**Figure** 3.27). Among four out of five HNRNPH SRBS clusters, the iCLIP approach confirmed HNRNPH binding as seen by iCLIP signals (**Figure** 3.28). However, cluster 4 lacked a significant iCLIP signal, suggesting that it might possibly be bound by another RBP with similar binding profiles.

The binding motif for HNRNPH was previously shown to consist of guanine-rich sequences (Uren et al., 2016). We found that guanine-rich sequences (G-runs) were abundantly present in the defined HNRNPH SRBS. In order to understand the possible role of G-runs in altering HNRNPH binding, we dissected the clusters into their constituent

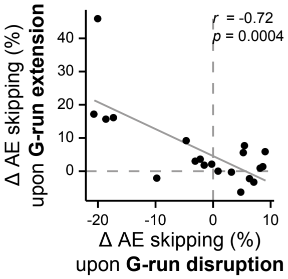


Figure 3.29: Opposing effects on splicing by extension and disruption of G-run stretches in the *RON* minigene. Scatterplot showing correlation of median AE skipping induced by mutations that extend or disrupt G-runs within HNRNPH SRBS. *r*, Spearman correlation-coefficient; *p*, associated *p*-value.

SRBS. In detail, the five SRBS at the AE cluster 3 and their splicing effects on G-run disruptions are shown (**Figure** 3.30). Strikingly, the effects are similar across all these SRBS with increased AE inclusion and simultaneous decrease in AE skipping. In summary, splicing was altered upon disruption or extension of G-runs by an additional guanine, which may lead to reduced or enhanced HNRNPH binding, respectively (**Figure** 3.29).

Importantly, the HNRNPH-mediated splicing changes showed opposing behaviour depending on where HNRNPH binding sites were positioned across the minigene. Consequently, HNRNPH binding in the alternative exon led to repression of the AE inclusion, however binding at the upstream intron activated splicing. In detail, mutations in cluster 1 and 5, located in the up- and downstream constitutive exons, respectively, reduced intron retention and in cluster 5, led to increased AE skipping (**Figure** 3.31 **A**). In line with previously described literature, context-dependent regulation of splicing regulators also extends to HNRNPH (Xiao et al., 2009; Katz et al., 2010). The strongest splicing effects were seen in mutations of HNRNPH SRSBS cluster 3 in the alternative exon (**Figure** 3.31 **A**). Moreover, *HNRNPH* KD induces a similar splicing change as mutations in cluster 3, underlining the significance of HNRNPH SRBS cluster 3 in the HNRNPH-mediated regulation of *RON* splicing (**Figure** 3.31 **B**). In conclusion, HNRNPH acts simultaneously as

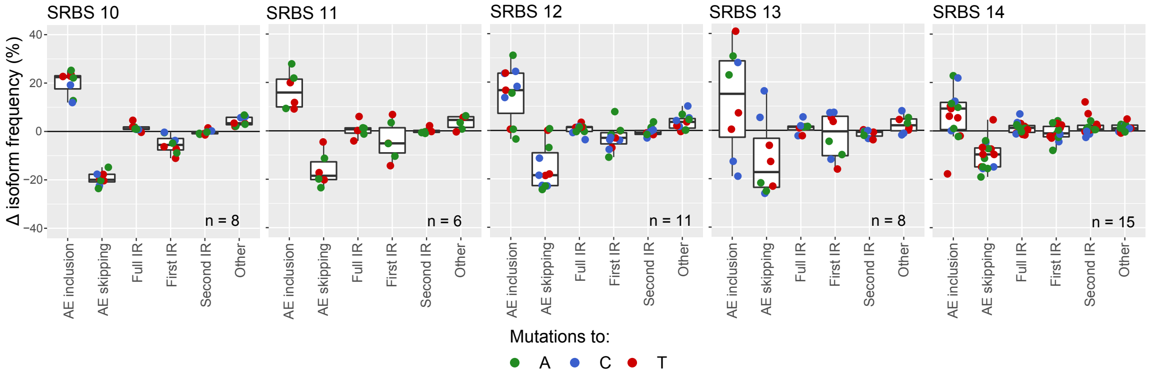


Figure 3.30: HNRNPH-mediated regulation of *RON* splicing is dependent on G-runs at the SRBS in the *RON* AE. SRBS 10-14 at the *RON* AE show similar strong splicing effects in the same direction. Boxplots show the isoform frequencies as a result of G-run disrupting mutations encompassed in HNRNPH SRBS 10-14 (above) in MCF7 cells (average of three biological replicates). Number of mutations for each SRBS is given.

an activator and repressor of *RON* splicing and mediates complex regulation of different isoforms via multiple SRBS clustering across the minigene region.

Synergy analysis shows interactions between mutations and *HNRNPH* knockdown

Although HNRNPH is shown to bind across the entire minigene (**Figure 3.27**), the mutation effects showed a varied response between different HNRNPH binding sites (**Figure 3.31**). This suggests that binding sites differentially affect *RON* splicing regulation i.e. not all binding sites contribute in the same manner towards *RON* splicing regulation. In order to identify the most relevant sites involved in HNRNPH-regulation, splicing outcome of the minigene library was analysed under *HNRNPH* KD conditions. A synergy implies that the combined response of mutation and knockdown are greater or smaller than the response expected from their independent contributions respectively (‘the whole is greater

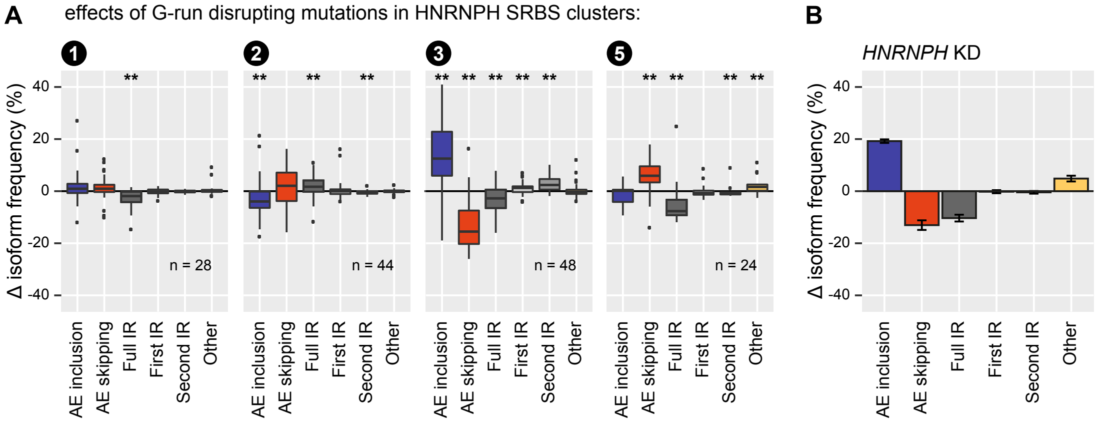


Figure 3.31: HNRNPH-mediated regulation of *RON* splicing is dependent on HNRNPH binding position across the *RON* minigene. (A) Boxplots show the isoform frequencies as a result of G-run disrupting mutations encompassed in different HNRNPH SRBS clusters (circled numbers) in MCF7 cells (average of three biological replicates). Number of mutations for each cluster is given. * p-value < 0.05, ** p-value < 0.01, one-sample Wilcoxon's test against population mean of zero. (B) Bar diagram showing the isoform frequency changes of the wt minigenes resulting from *HNRNPH* KD. Error bars indicate standard error of the mean from three biological replicates.

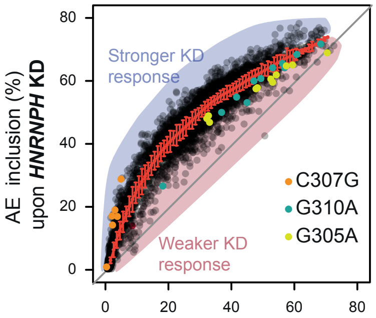


Figure 3.32: Minigene variants in the library are differentially affected by *HNRNPH* KD. Scatterplot showing splicing change of variants upon *HNRNPH* KD compared to ctrl. Several variants are strongly (blue shade) or weakly (red shade) affected by the *HNRNPH* KD compared to the majority of variants (described by the running mean ± standard deviation, red line). Minigene variants harbouring indicated point mutations C307G, G310A, and G305A are highlighted in orange, blue, and yellow, respectively.

or even lesser than the sum of its parts'). This means that the mutations that affect HNRNPH binding, were expected to show either positive or negative synergy in the context of *HNRNPH* KD. To be precise, mutations disturbing HNRNPH binding sites, would display reduced KD response in relation to wt minigenes (negative synergy), while mutations strengthening HNRNPH binding sites would display enhanced KD response in relation to wt minigenes (positive synergy).

Importantly, *HNRNPH* KD led to increase in AE inclusion on a system-wide scale (**Figure 3.32**) and as expected, subsets of minigenes displayed weaker or stronger KD responses compared to the rest of the library. For instance, minigenes that harbour mutations within HNRNPH SRBS in cluster 3, such as G305A or G310A, showed a consistently weaker KD response but higher AE inclusion in control cells, implying negative synergy. In contrast, minigenes with C307G mutations which extend a pre-existing HNRNPH binding motif by an additional guanine, showed a stronger KD response but lower control AE inclusion. This confirms positive synergy as the *HNRNPH* KD promoted a stronger-than-average increase of AE inclusion in these minigenes.

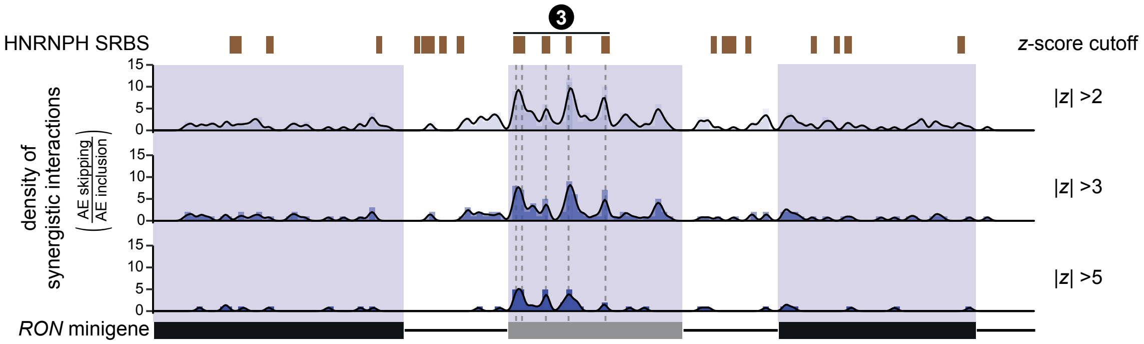


Figure 3.33: Synergistic interactions between mutations and *HNRNPH* KD overlap at HNRNPH SRBS regions in the *RON* alternative exon. Bar plots quantify significant synergistic interactions affecting AE skipping-to-inclusion isoform ratio using different z-score cutoffs in adjacent 5-nt windows. Line indicates density in 5-nt sliding window. Splice sites ± 2 nt were excluded. Predicted HNRNPH SRBS (brown) are given above.

The synergy effects for each mutation were calculated as the difference of the mutation effects between ctrl (control) and *HNRNPH* KD. Then, to account for the experimental variation of the wt minigenes, a z-score normalization approach was used to define synergy effects quantitatively (**Figure** 3.33). If resulting z-score tended to zero, then the mutations show similar effects between ctrl and *HNRNPH* KD, implying that those mutations and *HNRNPH* KD are not synergistic. However, z-scores greater or smaller than zero indicate positive or negative synergy, respectively.

In total, 358 mutations in 281 positions (41%) showed significant synergy for at least one splice isoform ($|z\text{-score}|>2$, adjusted p-value < 0.001, Stouffer's test). Importantly, 93% of positions within SRBS cluster 3 were synergistic (**Figure** 3.34 **A**). Hence, such strong synergistic interactions within the alternative exon represents the most functionally relevant region involved in HNRNPH-mediated splicing regulation in *RON*.

Furthermore, to validate the importance of AE cluster, the splicing outcome under control and *HNRNPH* KD conditions of ten different minigene variants was measured. Each variant harboured point mutations in one of the five clusters (**Figure** 3.34 **B**). In line with this idea, all the point mutations located in cluster 3 showed largely unchanged response upon *HNRNPH* KD. However, splicing outcome of minigenes harbouring point mutations in other clusters still showed changes upon *HNRNPH* KD, which agreed with the reduced synergy quantified for these clusters (other than cluster 3).

Cooperative HNRNPH binding establishes a splicing switch

An important finding of the screening approach is that individual point mutations in cluster 3 are able to completely release HNRNPH-mediated repression of AE inclusion. For instance, mutation effects of point mutations (G305A, G331C and G348C) in the alternative exon (cluster 3) resemble HNRNPH-mediated repression of *RON* AE inclusion. This suggests that HNRNPH binding occurs in an interdependent manner (**Figure 3.34 B**). Furthermore, mutations that antagonise HNRNPH binding also prevented HNRNPH binding at other SRBS. In order to test this finding, iCLIP experiments were repeated in the context of minigene variants harbouring only single point mutations in cluster 3 (**Figure 3.35**). The reduction of iCLIP signal was not limited to the site of the point mutation but extended across the entire alternative exon (cluster 3), implying a cooperative mode of binding.

The cooperative regulation of *RON* splicing by HNRNPH implied that *RON* isoform levels react to changes in HNRNPH concentrations and show a switch-like behaviour in turn resembling a steep, sigmoidal dose-response curve on cartesian co-ordinates. In order to show this behaviour, *RON* minigene and endogenous *RON* splicing levels were measured after gradually changing HNRNPH levels in both directions (*HNRNPH* KD- and *HNRNPH1* over-expression titration) (**Figure 3.36 A and 3.36 C-H**). As a result, a switch-like splicing response of *RON* AE was observed with changing HNRNPH concentrations with a high Hill's co-efficient for both the *RON* minigene and endogenous *RON*, validating that HNRNPH binds strongly in a cooperative manner. Additionally, *HNRNPH2* showed

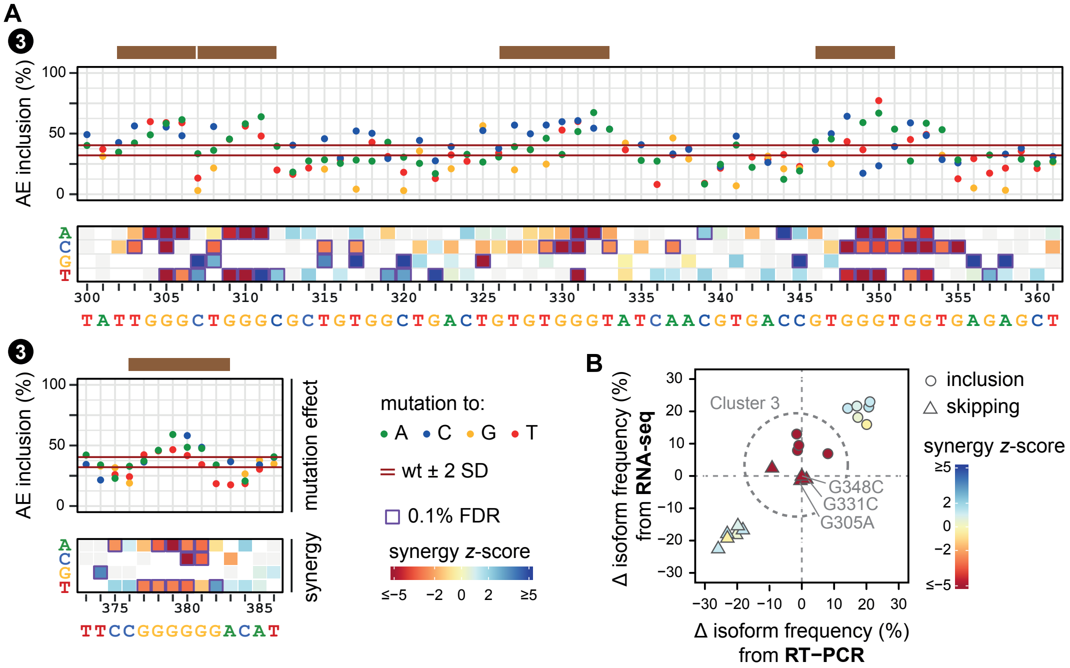


Figure 3.34: AE cluster shows high functional relevance for HNRNPH-mediated regulation of *RON* splicing. (A) Dot plots (top) display single mutation effects (inserted base, see legend) on AE inclusion (mean, n=3). Red lines indicate median isoform frequency of wt minigenes ± 2 standard deviations (SD). HNRNPH SRBS (brown) are given above. Heatmaps (bottom) show z-scores as measure of synergy (mean, n=3) per inserted base. White or grey fields indicate mutations that were not present or filtered out, respectively (see methods). Purple boxes highlight significant synergistic interactions (0.1% FDR). (B) Scatterplot compares the splicing change of point mutation harbouring minigenes between control and *HNRNPH* KD conditions (mean, n=3) as inferred by the model from RNA-seq data (y-axis) and RT-PCR analysis (x-axis). AE inclusion (circles) and AE skipping (triangle) are shown separately. In summary, KD followed by RNA-seq of the mutated minigene library followed by inference of synergy between mutations and *HNRNPH* KD enabled the ranking of the functionally relevant binding sites (SRBS).

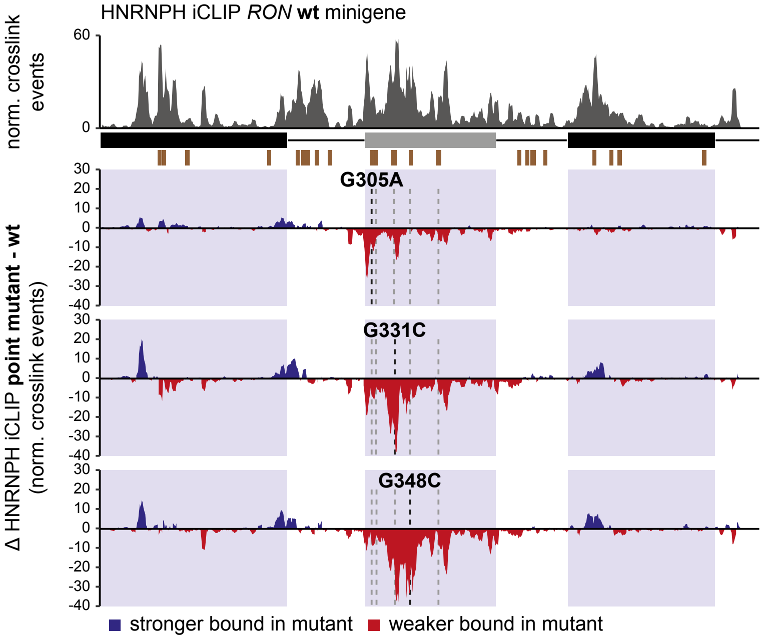


Figure 3.35: HNRNPH iCLIP signal reduction extends towards neighbouring HNRNPH binding sites in different point mutation minigene variants. Bar diagram shows the normalised cross-linking events per position across the wt *RON* minigene in a sliding 5-nt window (top; HNRNPH iCLIP). Bar diagram showing the difference of normalised cross-linking events of indicated point mutants and the wt *RON* minigene in 5-nt sliding windows (bottom). Strong or weak binding in the mutants is indicated with blue and red, respectively. HRNPH SRBS are depicted with brown boxes and dashed lines highlight their location within the bar diagrams.

the steepest regression slope among the 190 RBPs tested for correlation with TCGA cancer expression data, confirming *in-vivo* cooperative regulation (**Figure** 3.36 **B**). In summary, HNRNPH behaves as a splicing switch of *RON* exon 11 through cooperative binding, resulting in a major splicing shift due to minor changes in HNRNPH protein level.

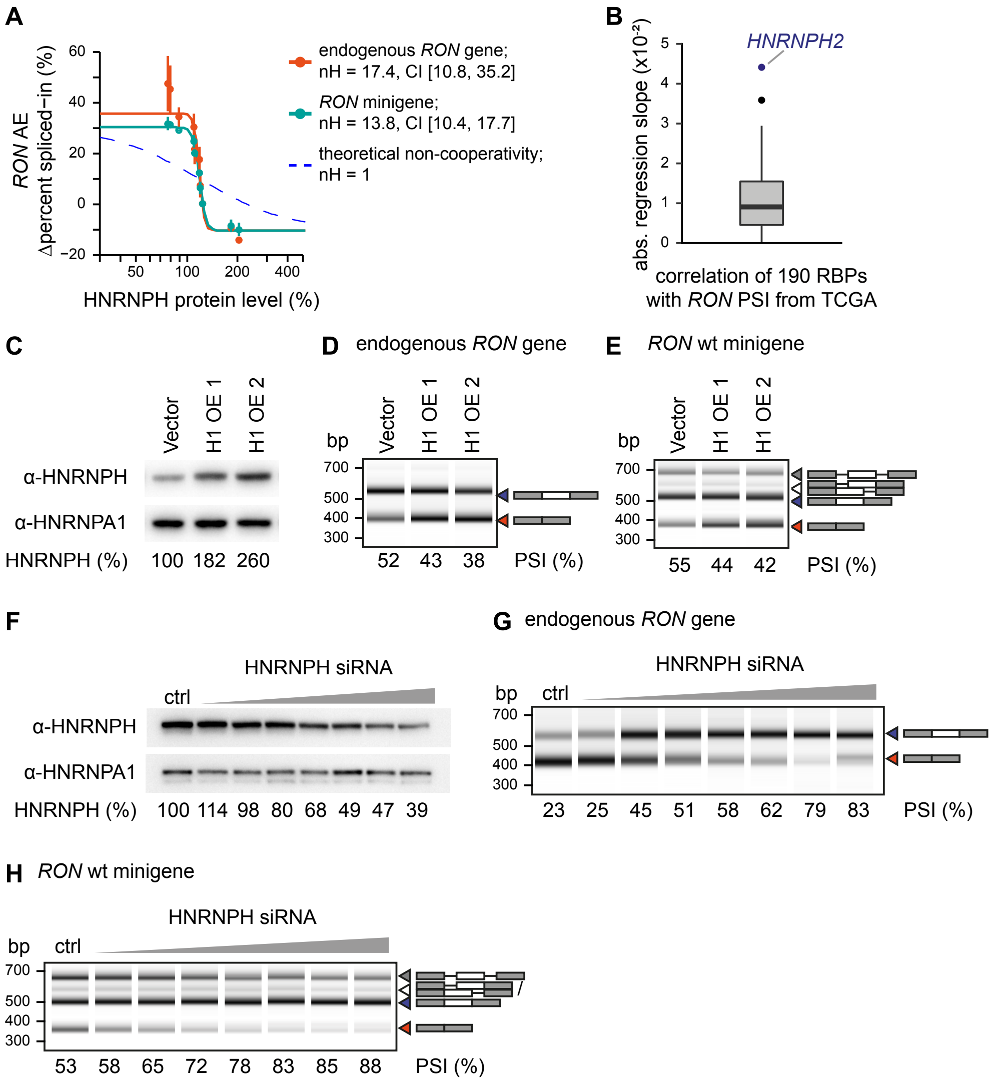


Figure 3.36: Cooperative binding of HNRNPH at *RON* AE results in a splicing switch. (A) Dose-response curve showing the change of *RON* AE percent-spliced-in for changes in HNRNPH protein levels for endogenous *RON* (orange) and the *RON* minigene (blue). Degree of cooperativity is quantified by fitting the Hill equation (solid lines) and compared to the theoretical fit for non-cooperativity (dashed line). Error bars denote standard deviation of biological triplicates. CI confidence intervals. (B) Boxplot shows the distribution of regression slopes from the correlation of 190 RBPs with *RON* exon 11 splicing in TCGA samples (refer materials and methods). *HNRNPH2* is highlighted. (C) Representative Western Blot of *HNRNPH1* overexpression using either empty vector or increasing amount of *HNRNPH1* overexpression construct for transfection (H1 OE 1 and H1 OE 2). HNRNPA1 was used for normalization. (D, E) Representative RT-PCR results corresponding to the *HNRNPH1* overexpression shown in (**C**) for the endogenous *RON* gene (**D**) or the *RON* minigene (**E**). Gel-like representation of capillary electrophoresis. (**F**) Representative Western Blot of gradual HNRNPH KD obtained through transfection of increasing siRNA amount. HNRNPA1 was used for normalization. (**G**, **H**) Representative RT-PCR results corresponding to the *HNRNPH* KD shown in (**F**) for the endogenous *RON* gene (**G**) or the *RON* minigene (**H**). Gel-like representation of capillary electrophoresis. PSI percent-spliced-in.

4

Discussion

Preface

At the outset of this PhD project, previous mutagenesis studies on determining the effect and extent of *cis*-regulatory elements on regulating AS were found to be lacking. These studies were either confined to short exonic regions or to synthetic reporter constructs. Secondly, the mutations affecting the splicing functions were predicted based on different models and may not always accurately represent *in-vivo* quantifications of mutation effects. In order to mitigate this, our novel high-throughput screening approach to decode the alternative splicing of *RON* exon 11 was devised. The screening approach forms the basis of this project. Our study involves the integration of experimental, bioinformatics and mathematical modelling aspects, which has resulted in the most complete *cis*-regulatory landscape of an alternative splicing minigene model till date. In detail, the repertoire of multiple mutations affecting *RON* exon 11 splicing was determined with precise quantitative information. In addition, linear regression based approaches with a dynamic model of splicing enabled the accurate dissection of complex mutation effects. The results enabled the recapitulation of splicing regulation in tumour patients and provided new insights into the regulatory role of sequences affecting downstream protein function. Furthermore, a key player in the regulation of *RON* alternative splicing, *HNRNPH* was knocked down in the context of the mutagenised library. From these experiments, a concentration dependent regulation of splicing at multiple binding locations of HNRNPH was observed. By systematically linking *cis*-regulatory elements to the synergistic activity of the *trans*-acting factor HNRNPH, an additional regulatory layer of alternative splicing was observed. Finally, the cooperative HNRNPH binding was shown to mediate a splicing switch of *RON* exon 11. The screening approach provided an unprecedented view of the complexity of splicing regulation of a single exon. The approach also gave rise to several insights into the molecular details of *RON* alternative splicing which are discussed in the following sections.

The widespread ever-expanding landscape of splicing-effective mutations

Experimental studies have previously estimated that 30%-60% of human exonic sequences contains putative *cis*-regulatory elements (Parmley et al., 2006) (Savisaar et al., 2017a). An early computational study predicted that alternative exons on average contain 10.2 *cis*-regulatory 8-mers per 140-nt sequence, which corresponds to 58% of exon positions involved in splicing regulation (Zhang et al., 2004), implying that the sequences that influence splicing may be very abundant in pre-mRNA. Yet, the occurrence of splicing regulation elements like enhancers and silencers are still vastly underestimated (Hurst et al., 2017).

In the current study focussed on the *RON* minigene, we showed that more than 90% of exon 11 positions affect splicing suggesting that splicing regulation is conferred by most nucleotides within the alternative exon. In this respect, *RON* exon 11 is not atypical given that the other mutagenesis studies report a high density of splicing effective mutations in *CFTR* exon 12, *SMN1* exon 7, and *FAS*/*CD95* exon 6 (87 bp, 54 bp, and 63 bp respectively) (Pagani et al., 2005; Mueller et al., 2015; Julien et al., 2016). However, these studies were done on similar-sized small exons and raise questions on the validity of these insights in the case of larger exons (Savisaar et al., 2017b). In particular, the previous minigene studies with shorter exons were considered to overestimate the density of splicing-effective mutations and computational predictions were thought to be more suitable (Savisaar et al., 2017b). Nevertheless, in line with aforementioned studies, we found that splicing information was densely spread in the *RON* alternative exon. In fact, the length of *RON* alternative exon, exon 11 (147 bp) was greater than the median length of a typical human exon (134 bp), implying that indeed, the compact and dense arrangement of splicing signals is typical across average sized exons. In summary, splicing regulation is conferred in a complex way as splicing regulation stems from many more positions in the alternative exon than previously anticipated.

Many *trans*-acting factors regulate *RON* splicing

In the project, the catalogue of *trans*-acting factors that putatively regulate *RON* exon 11 splicing were short-listed by three approaches. Firstly, individual KDs of *trans*-acting factors from a previously described high-throughput candidate screening approach (Papasaikas et al., 2015) resulted in a list of *trans*-acting factors interacting with *RON* splicing.

The idea behind the approach was not to identify *de novo* regulators but to list out the strong regulators of *RON* splicing. Particularly, among the top 5 RBPs affecting *RON* regulation, we saw a global shift and a widespread range of splicing effects of mutations upon the KD of the respective RBPs in the context of the minigene library (**Figure** 3.23). We observed that the mutation effects scale non-monotonically with the starting inclusion level of the exon in different RBP KDs (**Figure** 3.24). This means that the HNRNPH KD resulted in mutated minigene variants showing different splicing effects based on their starting control inclusion levels. Two recent studies that dealt with understanding the behaviour of splice-altering mutations, suggested that exons with intermediate inclusion levels have a higher tendency to be affected by splicing mutations or be alternatively spliced differently than those exons with maximum or minimum inclusion levels. This was referred to as the non-monotonic global scaling law (Baeza-Centurion et al., 2019; Jaganathan et al., 2019). As an initial approach, the KD of HNRNPH in the context of the minigene library in multiple replicates, allowed for the study and synergistic modelling of HNRNPH-mediated splicing regulation.

The second approach employed an unbiased *in-silico* prediction and validation of RBPs regulating *RON* AE using the ATtRACT database. The high number of putative predicted factors (76) from the ATtRACT database suggested that *RON* splicing regulation is more complex than previously anticipated. However, this led to us to asking whether all these factors were actively involved in *RON* splicing regulation. Additionally, multiple *trans*-acting factors may share overlapping *cis*-regulatory elements, which complicates the selection of *cis*-regulatory elements. Previously, it has been shown that alternative splicing is highly context- and tissue-specific, which has mainly been attributed to the combinatorial nature of alternative splicing regulation and cell type-specific RBP expression patterns (Fu et al., 2014). Furthermore, Jaganathan et al. (2019) suggest that weak or intermediate cryptic splice mutations play central roles in generating tissue specificity. This choice of splice site usage may depend on its environment; for instance the concentration of RBPs in the pool. Hence, the putative predicted factors might compete for binding in a context dependent manner in a restricted sequence space, leading to many *trans*-acting factors putatively sharing contiguous *cis*-regulatory elements. Accordingly, several putative regulators suggested in this work by the ATtRACT database analysis are cell type- or tissue-specifically expressed and therefore, may result in drawbacks in the predictions of the given binding site under the experimental conditions. Firstly, the given binding site prediction might not be fully replicable under specific experimental conditions. Secondly, the binding preferences of RBPs in the database were generated from *in vitro* based methods like SELEX which might be inaccurate *in-vivo* (Wu et al., 2016) and may not always sufficiently recapitulate the role played by cofactors influencing the binding of splicing factors *in-vivo*. Furthermore, splicing factors bind to a subset of their natural binding motifs present in the transcriptome and that their resultant binding is shaped by other cofactors (Sutandy et al., 2018). Hence, further validation experiments are required in order to confirm these regulators. In our screening approach, we find a large number of validated regulators recapitulating the idea of extensive regulation of *RON* splicing (**Figure 3.22**). This finding is in line with the numerous splicing factors linked to the *RON* splicing regulatory network, shown by the Papasaikas et al. (2015) study.

Finally, in order to identify novel regulators of *RON* splicing that were not considered in the RBP KD screen (Papasaikas et al., 2015), *in-vivo* RBP expression and *RON* splicing was analysed from TCGA tumour datasets. The association between RBP expression levels and *RON* AE PSI was observed for many numerous RBPs, with HNRNPH2 emerging as the strongest splicing determinant *in-vivo* (**Figure 3.26**). In summary, we find that a large number of *trans*-acting factors regulate *RON* splicing, signifying a splicing regulatory network of *RON* composed of many factors, among which HNRNPH suggesting that *RON* splicing regulation is complex and composed of a network of splicing factors instead of simpler, unit interactions between *cis*-regulatory elements and individual *trans*-acting factors. In addition, combining several approaches for the identification of putative splicing regulators results in candidate RBPs that still require downstream validation.

Cooperative regulation in *RON* splicing is mediated by HNRNPH

A previous study by Dominguez et al. (2010) showed that HNRNPH pre-mRNA targets harbour multiple G-tract binding sites. In line with this finding, we confirm that HNRNPH binds at multiple G-runs (SRBS) and extending or shortening G-runs can affect *RON* splicing (**Figure 3.29**). In addition, we see that mutating individual binding sites of HNRNPH can recapitulate the full effect of *HNRNPH* KD in terms of altering AE inclusion. iCLIP experiments showed that individual point mutations alone are potent enough to alter HNRNPH binding in the cluster 3, and small changes in HNRNPH concentration led to drastic AE inclusion changes. In tumours, *RON* exon 11 inclusion shows a particularly steep correlation with *HNRNPH2* expression levels. In summary, these results point to the extensive regulation of *RON* splicing mediated by HNRPNH. In fact, we show that HNRNPH exhibits a cooperative model of splicing regulation in *RON*. The cooperative behaviour of HNRNPH is attributed to the structural underpinnings of the process as detailed in this section.

A common feature among splicing factors is that they bind as multimers. For instance, the polypyrimidine tract-binding protein (PTB or HNRNPI) binds to a long stretch of polypyrimidine tracts in multimers (Clerte et al., 2006). This preference is attributed to the modularity of structural features of these proteins, as they possess multiple RRMs which have both protein-protein and RNA-protein binding affinities in addition to interactions with other RRMs (Cléry et al., 2008). Indeed, other HNRNP proteins too form multimeric assemblies and bind along the pre-mRNA (Casas-Finet et al., 1993; Bedard et al., 2004) at multiple locations via qRRMs (Dominguez et al., 2010). Strikingly, such an assembly is attributed to oligomerisation of the inherent GY-rich structural domains of HNRNP proteins (Akerman et al., 2009; Gueroussov et al., 2017). Moreover, Van Dusen et al. (2010)

showed that the glycine/tyrosine (GY)-rich domains are important for oligomerisation of HNRNPH and in turn, mediate cooperativity, finally resulting in nucleocytoplasmic shuttling. As a consequence, the cooperative interactions of HNRNPH affect the splicing regulation of the target pre-mRNAs (Gueroussov et al., 2017) in line with similar cooperative behaviour exhibited by other splicing related proteins, such as SF1 and U2AF65 (Berglund et al., 1998; Corioni et al., 2011).

Furthermore in this study, we find that the cooperative regulation of *RON* splicing enables a switch-like splicing response. Apart from abnormal disease states, the switch response might prove beneficial under normal physiological states that require a switch between two states. A possible scenario that explains this, would be as a switch during human embryonic development, where a switch occurs between initial rapid cell proliferation (4th week post-fertilisation) to later organ development (9th week post-fertilisation) (Yi et al., 2010). In fact, *HNRNPH2* ranks among the top 10% of the genes that are differentially regulated in this development switch (Yi et al., 2010) which signifies increased *RON* AE inclusion and lower RONΔ165 levels and in turn, reduced cell proliferation and activation of migratory pathways. In summary, *HNRNPH* gene expression changes which lead to the switch-like splicing of *RON*, may enable the precise modulation of the proliferation and migration activity of cells during embryogenesis.

Pre-mRNA structural context affects *RON* splicing regulation

Many studies have detailed on how RNA structures can affect splicing. In one such study, the splicing activity of RBP MBNL was shown to be dependent on the structural context i.e the splicing outcome was dependent on the arrangement and number of binding sites of RBP within unstructured RNA (Taylor et al., 2018). Similarly, it was shown that the structural context in the target pre-mRNA was crucial for HNRNPH-mediated splicing (Jablonski et al., 2008).

While structural studies are outside the purview of the project, the effects of G-quadruplex forming sequences could still be explored (**Figure** 3.18 **A**). Splicing-effective mutations in the G-runs disrupt or enhance the propensity of G-quadruplex formations and thus affect splicing regulation of *RON* (**Figure** 3.18 **A**; **Figure** 4.2). Although, there is uncertainty about the formation of G-quadruplexes (Guo et al., 2016), many studies implicate these structural elements in regulating splicing regulation. For instance, a G-quadruplex structure in stem-loop structure within *H-Ras* pre-mRNA was shown to regulate splicing through interactions with HNRNPH (Camats et al., 2008). Accordingly, a multimeric assembly and cooperative regulation of HNRNPH might allow remodelling of RNA structural features, including G-quadruplexes, to maintain them in single-stranded conformation. In the current study of *RON* exon 11 AS regulation, this might suggest a dual-state switching model of splicing regulation where the mode of regulation shifts between an open and a closed state. A relaxed and open conformation mediated by an HNRNPH assembly might allow the sequestering of G-runs and thus hinder the formation of the G-quadruplex. In contrast, in a released state or under low HNRNPH concentration, the G-runs may readily fold into G-quadruplexes thus explaining the observed switch-like splicing response upon altered HNRNPH levels (**Figure** 3.36).

However, it is still unclear if HNRNPH maintains G-quadruplex-forming RNA in single-stranded conformation or binds to pre-formed G-quadruplex structures. Previous studies in this regard have attributed both the aforementioned aspects of HNRNPH binding to G-quadruplexes. In detail, Conlon et al. (2016) showed that G-quadruplex structures are bound and held together by HNRNPH, where as von Hacht et al. (2014) proposed that HNRNPH binds and dissolves G-quadruplexes (Hacht et al., 2014; Conlon et al., 2016). This suggests that splicing regulation by G-quadruplexes may act in a context-dependent manner and its action may differ in individual targets. In particular, it was shown that G-quadruplex-resolving helicase DHX36 unwinds G-quadruplexes to allow access of HNRNPH/F and that DDX5 and DDX17 interact with HNRNPH/F at G-quadruplex-forming sites to mediate splicing regulation (Dardenne et al., 2014; Newman et al., 2017). Therefore, different modes of regulation seem to be target-dependent and increasing evidence highlights the importance of G-quadruplexes and associated G-quadruplex-resolving helicases in HNRNPH-mediated splicing regulation.

Taken together, cooperativity could be achieved through HNRNPH assemblies on *RON* exon 11. In fact, G-quadruplex structures might be involved in the control of *RON* splicing and their interplay with HNRNPH assemblies might cause the switch-like splicing, but further studies are required in this regard. Furthermore, our data might help extend studies on deducing the role of G-quadruplexes, for instance, in the recently developed high-throughput screens to unravel role of G-quadruplexes in gene regulation and EMT (Zhang et al., 2019).

Clinical and physiological consequences of splicing-effective mutations

Alternative splicing of the RON receptor kinase plays an active role in cancers of the breast, pancreas, lung and colon (Ghigna et al., 2005; Mayer et al., 2015; Chakedis et al., 2016). Among the different RON isoforms produced aberrantly in cancer, the isoform formed by the skipping of the *RON* alternative exon 11 ‘RONΔ165’, promotes tumour formation by aberrant RON activation (Chakedis et al., 2016). In fact, RONΔ165 is consistently upregulated in metastatic cancer, and contributes to tumour invasiveness by promoting epithelial-to-mesenchymal transition (Collesi et al., 1996; Zhou et al., 2003; Wang et al., 2004; Ghigna et al., 2005; Mayer et al., 2015). Hence, targetting the defective splicing mechanisms of *RON* signalling, involve either antibody-drug conjugates (ADCs) like Zt/gr-DM1 and H-Zt/g4-MMAE which have been shown to eradicate tumours *in-vivo* by inducing RON internalisation (Yao et al., 2006, 2019; Li et al., 2010) or involve therapeutic antibodies (mABs) that bind to RON receptors and block MSP-RON signalling pathways (antibody RON8, Narnatumab, ImClone; O'Toole et al. (2006)). Although, the process of RON-antibody binding does affect MSP-RON signalling, constitutive activation of RONΔ165 is not affected as it is ligand independent and tumours expressing this isoform can escape such therapies (Chakedis et al., 2016). Therefore, more information on the splicing impact of mutations on RONΔ165 pathogenic isoform is required to advance personalised therapy. Hence, for further directions on targetted therapy, one might require complete knowledge of the *cis*-regulatory elements. For instance, the set of splicing-effective mutations in patients. In this regard, the screen identified a set of ~1,800 mutations of which 778 and 1022 significantly affected splicing in HEK293T and MCF7 cells, respectively (refer **Figure** 4.2 and **Figure** 4.3). In comparison to previous limited targetted approaches or computational predictions, the mutagenesis screen resulted in precise measurements in terms of splicing quantifications. A quantitative measure from the screening data like AE skipping or inclusion, representing the splicing outcome of mutations can also be used to determine pathogenicity in cancers as the screening captures *in-vivo* splicing effects (**Figure** 3.21; refer **Figure** 4.2). Importantly, the mutation effects in the screening approach are reflected in cancer patients harbouring the same mutations. In detail, two splicing-effective mutations 'G297A' and 'G370T' were observed to be present in tumour patients with increased AE skipping levels of 43% and 34% on average, respectively. The same mutations showed 55% and 31% increased AE skipping levels in the screening data, respectively. It is evident that 'G297A' being a splice site mutation, shows a strong impact on AE skipping. In contrast, the mutation 'G370T' is a non-synonymous mutation and forms a premature termination codon (PTC). PTC-containing mRNAs are degraded early on through the nonsense-mediated mRNA decay (Nicholson et al., 2010). As a consequence of which, the alternative exon is not included in the mature mRNA and AE skipping isoforms are more abundant. These results from the screen direct to us to a new inverted role of the mutation where by the physiological consequences are themselves altered. Thus, instead of forming a less functional receptor, increased RONΔ165 levels are observed, leading to constitutive receptor activation as shown in previous studies (Collesi et al., 1996; Zhou et al., 2003).

Under normal physiological conditions, RON signalling regulates inflammatory immune response and wound healing in humans (Yao et al., 2013; Faham et al., 2016). Moreover, *RON* mRNA is lowly transcribed in many epithelial cells (Gaudino et al., 1994, 1995; refer introduction). Consequently, the read coverage for *RON* expression profiles in publicly available genome-wide sequencing data is low or zero (GTEX; Carithers et al., 2015; DeLuca et al., 2015) in accordance with its low expression levels (Wang et al., 1996). Therefore, alternative approaches are necessitated in order to assess the role of mutations within *RON*. In this regard, the mutagenesis screening approach provides quantitative splicing data for mutations affecting *RON* splicing. Taken together, the potential results from the screening approach will be helpful in training *in-silico* splicing prediction algorithms in the future and thereby, assist in clinical diagnosis of potential harmful mutations.

Significance of intronic and synonymous mutations in disease

Intronic mutations, including cryptic splice mutations, have been recently implicated in several cancers (Supek et al., 2014; Jung et al., 2015). In our study, intronic splicing-effective positions, even those with only a 5% change in isoform frequency, showed an increased evolutionary conservation compared to non-effective positions in MCF7 cells (**Figure 3.15**). Strong evolutionary conservation in the intronic splicing-effective positions could signify their relevance in mediating splicing deregulation in cancer. Moreover, effects of splicing-effective mutations from the screening approach recapitulates strong *in-vivo* splicing effects in cancers (**Figure 3.21**).

In addition, we saw that splicing-effective mutations distribute in proportion among the synonymous and non-synonymous mutations (**Table 4.2**), implying the equally potent deleterious role of synonymous mutations. However, synonymous mutations were previously considered as silent mutations (Greenman et al., 2006) and have been limited to fewer studies (Gotea et al., 2015). The role of evolutionary pressures on selecting synonymous variants have been observed both in normal (Savisaar et al., 2018) and cancer cells (Gartner et al., 2013; Supek et al., 2014). In line with this finding, we identify several synonymous mutations present in cancer patients with a possible deleterious splice-altering role (**Figure** 3.20).

These mutations do not alter the encoded protein content, but disrupt function through alternative splicing by inverting their functional or coding role. Such mutations might prove helpful in potential functional variant reclassification efforts (Cummings et al., 2017). Additionally, synonymous mutations have been previously implicated in altering HNRNPH2 binding sites in oncogenes, thus leading to tumour progression (Supek et al., 2014). In our study, exonic mutations at the AE cluster (cluster 3; SRBS 10-14), destroyed HNRNPH binding sites and led to decreased AE skipping (**Figure** 3.30). However, intronic mutations altering HNRNPH binding sites, particularly in SRBS cluster 2 and 5, promoted increased levels of the pathogenic AE skipping isoform, in contrast to reduced AE skipping on a system-wide scale upon HNRNPH KD (**Figure** 3.31). Hence synonymous mutations can alter HNRNPH binding, thus causing a deregulation of alternative splicing. In summary, the screening approach underscores the functional relevance of splicing-effective mutations and positions either intronic or/and synonymous that affect splicing in cancers. In particular, intronic mutations that cause cryptic splice site activations in disease could help identify unclassified variants (Findlay et al., 2018) and potentially benefit patients with severe genetic disorders (Bao et al., 2019).

Technical limitations of other studies

Many previous mutagenesis approaches were found lacking because of technology drawbacks at the time of study; such as limitation in oligonucleotide synthesis technology (Soemedi et al., 2017), which led to the study of mutation effects in short exons (<100 nt in length). Recent mutagenesis studies have examined the regulatory role of RNA sequences in a detailed manner (Mueller et al., 2015; Julien et al., 2016; Ke et al., 2018). However, these approaches were limited by only targetting mutations in a short synthetic spacer region or exons alone. Additionally, computational models that predict alternative splicing outcomes from sequences exist, for instance, percent spliced-in (PSI) predictions in SPIDEX (SPANR tool) (Xiong et al., 2015). However, predictive models may not be specific to tissue-types, where as splicing is indeed tissue-specific and requires minigene experimental data to run validation studies (Cheng et al., 2019). Hence, *in-vivo* splicing effects, derived from the screening approach, can be useful in updating and validating *in-silico* based predictive models (Baralle et al., 2017). Besides, splicing is shown to be extensively regulated by numerous *cis*-regulatory elements and both splice-enhancing and -silencing sequences frequently overlap. This presents a challenge as to how splicing regulation is organised and offers an explanation on the difficulty of predicting mutation effects (Grodecká et al., 2017). In comparison to all these studies, our study enabled a systematic quantification of mutation effects in intronic positions and showed that in addition to exons, a large number of positions within the introns of the *RON* minigene mediate splicing regulation. The average human introns are 3.3 kb long (Lander et al., 2001). In contrast, the introns flanking the AE of the *RON* minigene are much smaller in size (~80 bp). This in turn, enables a compact arrangement of splicing regulation signals in the atypically short introns. Thus, results from the screen still extend to other models albeit with short range interactions between the splicing signals.

Off-late, intronic regions are of special interest as cryptic splice mutations are involved in cancer (Supek et al., 2014; Jung et al., 2015) and the role of new cryptic splice sites in disease remain less understood. However, clinical sequencing has focused on rare coding mutations, largely disregarding variation in the non-coding genome due to the difficulty of interpretation (Jaganathan et al., 2019). Thus, *in-silico* prediction tools so far have not been rigorous enough to warrant a clinical diagnostic standard and instead compact mutagenesis studies are required to better understand splicing regulation (Baralle et al., 2017). Although newer approaches involving deep learning have led to promising discoveries, they still present potential problems. Jaganathan et al. (2019) showed that, while deep learning can automatically extract sequence features that are not well described by human experts, it still incorporates features that do not reflect the true behaviour of the spliceosome. These features have been shown to reduce the accuracy of predicting the splice-altering effects induced by genetic variation. Along with these drawbacks, complex interaction effects such as synergistic effects between the KD of an interacting RBP and mutations, are difficult to model by deep learning methods (Han et al., 2005). In contrast, our models allowed the identification of synergistic interactions of mutations with the HNRNPH KD data and enabled us to pinpoint the functionally relevant HNRNPH binding sites in *RON*. Moreover, we were able to probe motif interactions via mutations of SRBS along with their relative locations and distances, in relation to *RON* exon 11. In summary, as an extension to previous approaches where splicing regulatory elements within introns have not been systematically studied (Mueller et al., 2015; Julien et al., 2016; Ke et al., 2018), our approach enables the comprehensive study of positions that mediate splicing regulation both within the introns and exons. Furthermore, synergy analysis allows the identification of the most functionally relevant binding sites of RBPs and extends previous information on RBP regulation mechanisms.

Outlook

Other minigenes and approaches

The current study resulted in several novel insights into the sequence determinants and splicing regulation of *RON* alternative exon. As an extension of the project, currently, other minigene models are being investigated upon using the screening approach. In particular, random mutagenesis studies of minigenes derived from BRCA2, Ddx21 (mouse) and CD19 genes have been initiated. In detail, the BRCA2 minigene consisted of repeat elements leading to exonisation of intronic elements and the resulting malformed isoforms have been implicated in adenocarcinomas. The CD19 minigene model with the Exon 2 skipping pathological isoform is associated with leukemia in patients. In comparison, the Ddx21 minigene would be employed in understanding the regulatory role of sequence determinants in alternative polyadenylation. For reasons of initial prototyping during the purview of our study, it was imperative that a splicing minigene model limited to the confines of current sequencing technology be chosen. This resulted in the selection of the smaller sized *RON* minigene model and the short-read Illumina MiSeq for the sequencing of the *RON* minigene library. However, the next generation of sequencing technologies like the SMRT sequencing technology of Pacific Biosciences or Nanopore sequencing by Oxford Nanopore Technologies (Clarke et al., 2009; Eid et al., 2009) have come of age and are already in vogue (Ardui et al., 2018). Besides, Fuselli et al. (2018) previously used a combination approach in which problematic regions in long reads were resequenced with short read technology to adjust for any errors from long reads. Hence, these newly considered minigenes can be employed for library generation using these technologies. The brighter outlook is not only in terms of reducing costs of reagents but also computational resources, since it is less cumbersome and more efficient to align longer reads (Ardui et al., 2018).

In addition to SNVs, our approach also enables the quantification of the splicing effects of insertions and deletions (INDELS). In particular, a frameshift-causing indel may introduce pre-mature termination codon (PTC) into the mRNA and elicit non-sense mediated decay (NMD) (Lykke-Andersen et al., 2015) and indeed, many such isoforms may not be recovered by RNA-seq. Nevertheless, when mRNAs containing INDELS are translatable into aberrant protein products (pseudo-mRNAs), they may induce deregulation of splicing such as exon skipping by disruption of exon splicing enhancers (Tuladhar et al., 2019). INDELS were unexplored in the context of *RON* exon 11 splicing study as the study was optimised for single nucleotide variations and INDELS were not covered sufficiently by multiple plasmids. The regulatory effects of INDELS on splicing could be further explored as a next step in future screening approaches.

Another direction where more insights could be obtained, is from analysing the structural features of the minigene models and its correlation to splicing regulation. In light of the finding that only a small set of RNA binding motifs exist with low complexity (Dominguez et al., 2018), other contextual features such as flanking composition and RNA stucture play a prominent role in in deciding in which RBP finally elicits its splicing regulatory role. It was prevously shown that the structural context of pre-mRNA, in terms of the positional arrangement of its regulatory motifs in RNA hairpins, can influence the binding of RBPs and hence mediate splicing regulation (Taylor et al., 2018). Accordingly, models that predict RNA stability from sequence (Shi et al., 2015, 2018; Xu et al., 2016; Cheng et al., 2017) could be integrated to the screening approach to analyse splicing regulation in the context of RNA stability. Additionally, there needs to be more effort to learn relevant motifs not just from sequence features, but from additional features from other layers such as RBP kinetics (Sutandy et al., 2018). Moreover, chromatin modifications are known to affect splicing regulation in a significant manner (Kolasinska-Zwierz et al., 2009; Wang et al., 2015). Hence, the approach could possibly benefit from models taking chromatin states (histone accessibility scores) into account.

Recently, it has been shown that deep learning techniques may be particularly well suited to solve problems of big data biology (Ching et al., 2018). Especially in the context of alternative splicing regulation, recent studies offered new insights into the pathogenic roles of noncoding sequence variants, which were previously of unknown clinical significance (Baeza-Centurion et al., 2019; Jaganathan et al., 2019). In light of these findings, our data may help assess the disease relevance of mutations in clinical diagnostics and may additionally be used to train *in-silico* splicing prediction tools. Furthermore, where a single model may be insufficient to predict splicing-effective variants with high accuracy, Avsec et al. (2019) provides the Kipoi repository where multiple machine learning models may be employed in unison resulting in better accuracy. Taken together, our study shows that the combination of bioinformatics, mathematical modelling and experimental approaches lead to several insights into splicing regulation and thereby, such integrated approaches will provide for a deeper understanding of the splicing code.

Synergy analysis of other *trans*-acting proteins

Further studies to decipher the complete splicing regulatory network could be undertaken, by which many more *trans*-acting factors that interact strongly with *RON* splicing can be identified. Our study showed that *RON* splicing is regulated by multiple *trans*-acting factors and shows a non-monotonic synergistic behaviour when KD in the context of the minigene library (synergy analysis). Since, synergy analysis allows detection of functionally most relevant binding sites, assaying further regulators will enable to link additional trans-acting factors to their cognate cis-regulatory elements. Comprehensive analysis of similarities in the recruitment patterns along the minigene for a large number of *trans*-acting factors is expected to allow reconstruction of the full *RON* splicing regulatory network. This is attributed to the assumption that KD of factors that act in the same molecular complex should display synergies with the same set of cis-regulatory elements.

Other splicing switches mediated by HNRNPH

A more comprehensive analysis of other linear and switch-like splicing responses could be undertaken, from a deeper study of all the events that mediate HNRNPH cooperative splicing regulation. In order to investigate this, we generated transcriptome-wide RNA-seq datasets from cells with gradually changing HNRNPH levels. The RNA-seq dataset will be further used to determine the link between G-quadruplexes and cooperative HNRNPH splicing regulation in the studied events, in light of previous studies linking HNRNPH splicing regulation to G-quadruplexes (Bedrat et al., 2016; Kwok et al., 2016).

Need for further studies of *RON* splicing regulation

Although, AS plays an important role in every hallmark of cancer, it is rarely examined in either profiling of tumours or in biomarker and drug development in oncology. In fact, this is attributed to the difficulty in analysing splicing regulation data and a lack of partnership between bioinformaticians and cancer researchers (Robinson et al., 2019). In particular, the *RON* receptor tyrosine kinase has been recently shown to be a predictive marker for aggressive prostate cancer in African Americans (Bedolla et al., 2019) and drugs that internalise the receptor, like H-Zt/g4-MMAE have been in current development (Yao et al., 2019). Although there is some interest in these studies, there needs to be a renewed impetus to identify the complete role of *RON* splicing deregulation in cancers.

Abbreviations

Table 4.1: List of abbreviations

A	Adenine
ADC	Antibody-drug conjugates
AE	Alternative exon
AS	Alternative splicing
BMLS	Buchmann Institute for Molecular Life Sciences
BSA	Bovine serum albumin
C	Cytosine
cDNA	Complementary DNA
CI	Confidence interval
ctrl	Control
DNA	Deoxyribonucleic acid
DNA-seq	DNA-sequencing
dNTP	Deoxy nucleoside triphosphate
E. coli	Escherichia coli
EMT	Epithelial to mesenchymal transition
FDR	False discovery rate

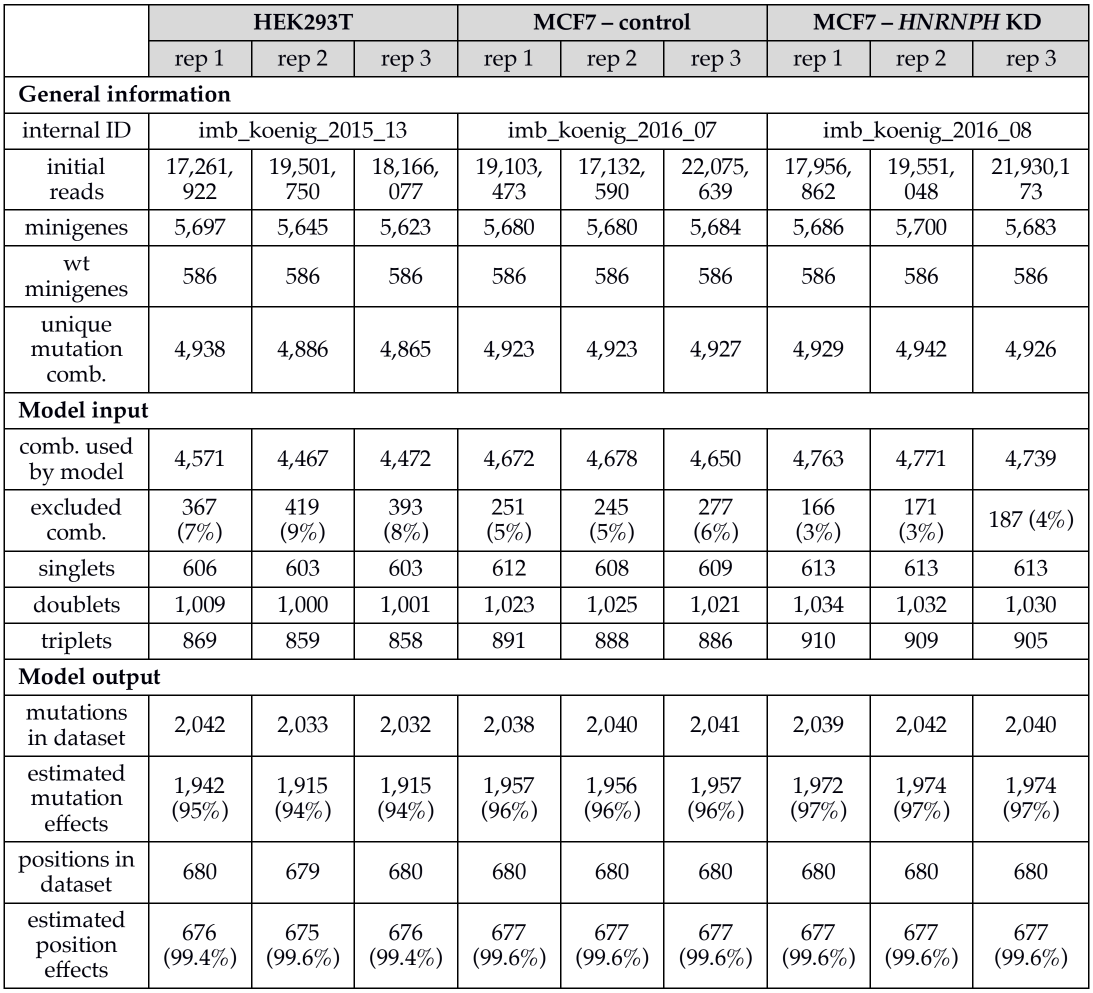
Table 4.1: List of abbreviations *(continued)*

fwd	Forward
G	Guanine
GFP	Green fluorescent protein
iCLIP	Individual-nucleotide resolution UV crosslinking and immunoprecipitation
IMB	Institute of Molecular Biology
iMM	Instituto de Medicina Molecular
INDELS	Insertions and Deletions
IR	Intron retention
KD	Knockdown
mRNA	Messenger ribonucleic acid
MSP	Macrophage-stimulating protein
MST1	Macrophage-stimulating 1
MST1R	Macrophage-stimulating protein receptor
N	Any nucleotide
NMD	Non-sense Mediated Decay
No.	Number
nt	Nucleotide
PBS	Phosphate Buffered Saline
PCR	Polymerase chain reaction
pre-mRNA	Precursor messenger ribonucleic acid
PTC	Premature termination codon
qRRM	Quasi RNA recognition motif
RBP	RNA binding protein
rev	Reverse
RNA	Ribonucleic acid

Table 4.1: List of abbreviations *(continued)*

RNA-seq	RNA-sequencing
RON	Recepteur dOrigine Nantais
RRM	RNA recognition motif
RT-PCR	Reverse transcription PCR
RT-qPCR	Real-time quantitative reverse transcription PCR
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
siRNA	small interfering RNA
SRBS	Splicing-regulatory binding site
T	Thymine
TCGA	The Cancer Genome Atlas
TPM	Transcripts Per Million
U	Uracil
wt	Wild type

Appendix of tables and supplementary materials



	HEK293T			MCF7 – control			MCF7 – *HNRNPH* KD		
	rep 1	rep 2	rep 3	rep 1	rep 2	rep 3	rep 1	rep 2	rep 3
General information									
internal ID	imb_koenig_2015_13			imb_koenig_2016_07			imb_koenig_2016_08		
initial reads	17,261, 922	19,501, 750	18,166, 077	19,103, 473	17,132, 590	22,075, 639	17,956, 862	19,551, 048	21,930,1 73
minigenes	5,697	5,645	5,623	5,680	5,680	5,684	5,686	5,700	5,683
wt minigenes	586	586	586	586	586	586	586	586	586
unique mutation comb.	4,938	4,886	4,865	4,923	4,923	4,927	4,929	4,942	4,926
Model input									
comb. used by model	4,571	4,467	4,472	4,672	4,678	4,650	4,763	4,771	4,739
excluded comb.	367 (7%)	419 (9%)	393 (8%)	251 (5%)	245 (5%)	277 (6%)	166 (3%)	171 (3%)	187 (4%)
singlets	606	603	603	612	608	609	613	613	613
doublets	1,009	1,000	1,001	1,023	1,025	1,021	1,034	1,032	1,030
triplets	869	859	858	891	888	886	910	909	905
Model output									
mutations in dataset	2,042	2,033	2,032	2,038	2,040	2,041	2,039	2,042	2,040
estimated mutation effects	1,942 (95%)	1,915 (94%)	1,915 (94%)	1,957 (96%)	1,956 (96%)	1,957 (96%)	1,972 (97%)	1,974 (97%)	1,974 (97%)
positions in dataset	680	679	680	680	680	680	680	680	680
estimated position effects	676 (99.4%)	675 (99.6%)	676 (99.4%)	677 (99.6%)	677 (99.6%)	677 (99.6%)	677 (99.6%)	677 (99.6%)	677 (99.6%)

Figure 4.1: Information on the input and output data of the mathematical model on the different RNA-seq replicates. For each RNA-seq replicate (rep), the internal library identifier is given together with information on the number of total and wt minigene variants detected in each dataset, the number of unique mutation combinations (differentiated into those used or excluded from the model analysis; refer **Table 4.5**) as well as the used single-/double-/triple-mutation combinations (singlets/doublets/triplets, respectively). Output information summarises the mutation and position effects that can be estimated by the model in relation to all mutations and mutated positions represented in each dataset.

		exon 10	intron 10	exon 11	intron 11	exon 12	intron 12	total
HEK293T	**mutations**	555	261	441	240	498	42	2037
	measured	487 (87.8%)	224 (85.8%)	381 (86.4%)	190 (79.2%)	430 (86.4%)	35 (83.3%)	1747 (85.8%)
	any isoform	117 (23.4%)	118 (51.5%)	270 (70.0%)	108 (55.1%)	144 (32.7%)	21 (60.0%)	778 (43.5%)
	AE inclusion	100 (20.5%)	111 (49.6%)	263 (69.0%)	87 (45.8%)	92 (21.4%)	19 (54.3%)	672 (38.5%)
	AE skipping	20 (4.1%)	67 (29.9%)	185 (48.6%)	53 (27.9%)	29 (6.7%)	9 (25.7%)	363 (20.8%)
	first IR	2 (0.4%)	6 (2.7%)	3 (0.8%)	0 (0.0%)	4 (0.9%)	2 (5.7%)	17 (1.0%)
	second IR	0 (0.0%)	1 (0.5%)	0 (0.0%)	3 (1.6%)	6 (1.4%)	10 (28.6%)	20 (1.1%)
	full IR	70 (14.4%)	74 (33%)	107 (28.1%)	79 (41.6%)	113 (26.3%)	16 (45.7%)	459 (26.3%)
	other	0 (0.0%)	0 (0.0%)	2 (0.5%)	4 (2.1%)	1 (0.2%)	0 (0.0%)	7 (0.4%)
	positions	185	87	147	80	166	14	679
	measured	184 (99.5%)	87 (100.0%)	147 (100.0%)	77 (96.2%)	166 (100.0%)	14 (100.0%)	675 (99.4%)
	any isoform	92 (49.7%)	67 (77.0%)	134 (91.2%)	64 (82.1%)	99 (59.6%)	13 (92.9%)	469 (69.3%)
	AE inclusion	81 (44.0%)	63 (72.4%)	134 (91.2%)	53 (68.8%)	73 (44.0%)	12 (85.7%)	416 (61.6%)
	AE skipping	20 (10.9%)	42 (48.3%)	110 (74.8%)	32 (41.6%)	28 (16.9%)	4 (28.6%)	236 (35.0%)
	first IR	2 (1.1%)	4 (4.6%)	3 (2.0%)	0 (0.0%)	4 (2.4%)	2 (14.3%)	15 (2.2%)
	second IR	0 (0.0%)	1 (1.2%)	0 (0.0%)	3 (3.9%)	6 (3.6%)	5 (35.7%)	15 (2.2%)
	full IR	59 (32.1%)	48 (55.2%)	70 (47.6%)	53 (68.8%)	82 (49.4%)	11 (78.6%)	323 (47.9%)
	others	0 (0.0%)	0 (0.0%)	2 (1.4%)	4 (5.2%)	1 (0.6%)	0 (0.0%)	7 (1.0%)

Figure 4.2: Splicing-effective mutations and positions per region in the *RON* minigene in HEK293T cells.

		exon 10	**intron 10**	**exon 11**	**intron 11**	**exon 12**	**intron 12**	**total**
MCF7	**mutations**	555	261	441	240	498	42	2037
	measured	501 (90.3%)	229 (87.7%)	386 (87.5%)	196 (81.7%)	440 (88.4%)	35 (83.3%)	1787 (87.7%)
	any isoform	150 (29.9%)	149 (65.1%)	300 (77.7%)	137 (70.0%)	264 (60.0%)	22 (62.9%)	1022 (57.2%)
	AE inclusion	81 (16.2%)	115 (50.2%)	260 (67.4%)	99 (50.5%)	91 (20.7%)	16 (45.7%)	662 (37.1%)
	AE skipping	86 (17.2%)	125 (54.6%)	271 (70.2%)	102 (52.0%)	217 (49.3%)	18 (51.4%)	819 (45.8%)
	first IR	5 (1.0%)	14 (6.1%)	5 (1.3%)	3 (1.5%)	6 (1.4%)	0 (0.0%)	33 (1.9%)
	second IR	1 (0.2%)	2 (0.9%)	12 (3.1%)	7 (3.6%)	15 (3.4%)	11 (31.4%)	48 (2.7%)
	full IR	79 (15.8%)	63 (27.5%)	62 (16.1%)	82 (41.8%)	185 (42.1%)	16 (45.7%)	487 (27.3%)
	other	3 (0.6%)	2 (0.9%)	8 (2.1%)	14 (7.1%)	13 (3.0%)	0 (0.0%)	40 (2.2%)
	positions	185	87	147	80	166	14	679
	measured	185 (100.0%)	87 (100.0%)	147 (100.0%)	78 (97.5%)	166 (100.0%)	14 (100.0%)	677 (99.7%)
	any isoform	108 (58.4%)	74 (85.1%)	139 (94.6%)	70 (89.7%)	147 (88.6%)	12 (85.7%)	550 (81.2%)
	AE inclusion	72 (39.0%)	62 (71.3%)	136 (92.5%)	57 (73.1%)	73 (44.0%)	11 (78.6%)	411 (60.7%)
	AE skipping	71 (38.4%)	65 (74.7%)	131 (89.1%)	58 (74.4%)	136 (81.9%)	9 (64.3%)	470 (69.4%)
	first IR	5 (2.7%)	11 (12.6%)	5 (3.4%)	3 (3.9%)	6 (3.6%)	0 (0.0%)	30 (4.4%)
	second IR	1 (0.5%)	2 (2.3%)	11 (7.5%)	6 (7.7%)	13 (7.8%)	6 (42.9%)	39 (5.8%)
	full IR	63 (34.1%)	44 (50.6%)	47 (32.0%)	55 (70.5%)	116 (69.9%)	10 (71.4%)	335 (49.5%)
	others	3 (1.6%)	1 (1.2%)	8 (5.4%)	12 (15.4%)	9 (5.4%)	0 (0.0%)	33 (4.9%)

Figure 4.3: Splicing-effective mutations and positions per region in the *RON* minigene in MCF7 cells.

		exon 10	**intron 10**	**exon 11**	**intron 11**	**exon 12**	**intron 12**	**total**
MCF7 – synergistic interactions with *HNRNPH* knockdown	**mutations**	555	261	441	240	498	42	2037
	measured	501 (90.3%)	229 (87.7%)	386 (87.5%)	196 (81.7%)	440 (88.4%)	35 (83.3%)	1787 (87.7%)
	any isoform	73 (14.6%)	50 (21.8%)	147 (38.1%)	63 (32.1%)	63 (14.3%)	13 (37.1%)	409 (22.9%)
	AE skipping	39 (53.4%)	30 (60%)	112 (76.2%)	24 (38.1%)	40 (63.5%)	5 (38.5%)	250 (61.1%)
	first IR	10 (13.7%)	6 (12%)	6 (4.1%)	5 (7.9%)	9 (14.3%)	2 (15.4%)	38 (9.3%)
	second IR	10 (13.7%)	12 (24%)	10 (6.8%)	6 (9.5%)	8 (12.7%)	1 (7.7%)	47 (11.5%)
	full IR	30 (41.1%)	24 (48%)	56 (38.1%)	19 (30.2%)	18 (28.6%)	5 (38.5%)	152 (37.2%)
	other	21 (28.8%)	23 (46%)	65 (44.2%)	39 (61.9%)	12 (19%)	4 (30.8%)	164 (40.1%)
	positions	185	87	147	80	166	14	679
	measured	185 (100%)	87 (100%)	147 (100%)	78 (97.5%)	166 (100%)	14 (100%)	677 (99.7%)
	any isoform	63 (34.1%)	38 (43.7%)	98 (66.7%)	46 (59%)	58 (34.9%)	9 (64.3%)	312 (46.1%)
	AE Skipping	36 (57.1%)	24 (63.2%)	74 (75.5%)	20 (43.5%)	38 (65.5%)	4 (44.4%)	196 (62.8%)
	first IR	10 (15.9%)	5 (13.2%)	6 (6.1%)	5 (10.9%)	9 (15.5%)	2 (22.2%)	37 (11.9%)
	second IR	10 (15.9%)	11 (28.9%)	9 (9.2%)	6 (13%)	8 (13.8%)	1 (11.1%)	45 (14.4%)
	full IR	29 (46%)	22 (57.9%)	47 (48%)	13 (28.3%)	18 (31%)	5 (55.6%)	134 (42.9%)
	other	21 (33.3%)	19 (50%)	47 (48%)	31 (67.4%)	11 (19%)	3 (33.3%)	132 (42.3%)

Figure 4.4: Synergistic interactions of mutations and positions per region in the *RON* minigene in MCF7 cells under HNRNPH KD.

Table 4.2: Significant splicing-regulatory effects are observed with equal frequency among synonymous and non-synonymous mutations.

Mutations	effective mutations	non-effective mutations	percentage of effective mutations
non-synonymous	312	568	35%
synonymous	135	272	33%

Table 4.3: List of top ten RBPs whose expression levels show strongest association with *RON* AE PSI levels in tumour samples from TCGA.

gene	ensembl-ID	Spearman-correlation	regression-slope	p-value	FDR
CDK10	ENSG00000185324	0.21	0.04	9.1e-47	8.7e-45
NXF1	ENSG00000162231	0.19	0.03	6.6e-38	4.2e-36
SNRNP70	ENSG00000104852	0.18	0.03	1.2e-32	4.7e-31
PRPF31	ENSG00000105618	0.17	0.03	1.3e-30	4.2e-29
PAXBP1	ENSG00000159086	0.16	0.03	7.3e-28	1.7e-26
PRPF4	ENSG00000136875	-0.15	-0.02	2.4e-25	3.9e-24
HNRNPF	ENSG00000169813	-0.15	-0.03	1.7e-25	2.9e-24
DHX8	ENSG00000067596	-0.16	-0.03	4.5e-29	1.2e-27
SLU7	ENSG00000164609	-0.18	-0.03	1.3e-33	6.2e-32
HNRNPH2	ENSG00000126945	-0.27	-0.04	1.0e-73	2.0e-71

Table 4.4: Mutation effects on *RON* exon 11 splicing in cancer patients.

Mutation	Genomic position (hg19)	Cohort	A	B	C	D	E	F
G43T	chr3:49933795	BLCA	242	1	26.50	-17.76	-17.76	-5.52
G56A	chr3:49933782	COAD	281	1	76.50	9.64	9.64	0.58
G56T	chr3:49933782	SKCM	44	1	26.00	-6.06	-6.06	5.32
G76T	chr3:49933762	CESC	247	1	181.50	-1.52	-1.52	0.45
G81A	chr3:49933757	BLCA	242	1	39.50	13.96	13.96	2.48
G103T	chr3:49933735	HNSC	423	1	19.00	NA	NA	NA
A118T	chr3:49933720	OV	175	1	14.50	NA	NA	NA
C119A	chr3:49933719	ESCA	172	1	297.50	-17.75	-17.75	-10.85
C124T	chr3:49933714	LUAD	444	1	138.00	2.78	2.78	-2.62
C133T	chr3:49933705	STAD	391	1	39.00	-3.52	-3.52	-1.02
C144A	chr3:49933694	OV	178	1	23.50	NA	NA	NA
G221T	chr3:49933617	HNSC	423	1	21.50	NA	NA	NA
G222A	chr3:49933616	BRCA	778	1	18.00	NA	NA	NA
G222T	chr3:49933616	COAD	281	1	87.50	-21.58	-21.58	0.02
G242T	chr3:49933596	THYM	49	1	16.00	NA	NA	NA
A246G	chr3:49933592	STAD	391	1	95.50	-3.45	-3.45	17.55
T277G	chr3:49933561	BLCA	242	6	46.00	5.74	1.20	14.85
T277G	chr3:49933561	PAAD	159	1	63.00	-10.56	1.20	14.85
T277G	chr3:49933561	OV	175	8	29.81	3.02	1.20	14.85
T277G	chr3:49933561	BRCA	739	8	27.44	-3.45	1.20	14.85
T277G	chr3:49933561	LUSC	268	9	34.61	-2.96	1.20	14.85
T277G	chr3:49933561	CESC	247	10	46.55	4.12	1.20	14.85
T277G	chr3:49933561	KIRC	5	1	10.00	NA	NA	NA
T277G	chr3:49933561	KIRP	45	1	11.50	NA	NA	NA

Table 4.4: Mutation effects on *RON* exon 11 splicing in cancer patients. *(continued)*

Mutation	Genomic position (hg19)	Cohort	A	B	C	D	E	F
T277G	chr3:49933561	THCA	271	12	22.62	NA	NA	NA
T277G	chr3:49933561	LUAD	444	15	51.93	4.94	1.20	14.85
T277G	chr3:49933561	STAD	391	4	127.50	4.12	1.20	14.85
T277G	chr3:49933561	HNSC	423	18	45.31	-1.72	1.20	14.85
C278G	chr3:49933560	STAD	391	1	99.00	10.70	10.70	11.08
C287T	chr3:49933551	SKCM	44	1	12.00	NA	NA	NA
G297A	chr3:49933541	THCA	271	1	25.00	42.96	42.96	54.65
C345A	chr3:49933493	SKCM	44	2	27.00	-3.98	-3.98	17.82
G348T	chr3:49933490	HNSC	423	1	21.00	NA	0.00	NA
G370T	chr3:49933468	HNSC	423	1	74.50	34.37	34.37	31.32
C375A	chr3:49933463	CESC	247	1	116.00	10.35	10.35	15.38
C376A	chr3:49933462	OV	178	1	23.50	NA	NA	NA
G381A	chr3:49933457	CESC	247	1	16.00	NA	NA	NA
C398A	chr3:49933440	HNSC	423	1	34.50	-2.20	-2.20	-4.75
C398A	chr3:49933440	SKCM	44	1	16.00	NA	NA	NA
C403T	chr3:49933435	LUSC	268	1	44.00	7.62	7.62	-5.28
C406A	chr3:49933432	SKCM	44	1	15.50	NA	NA	NA
C411A	chr3:49933427	HNSC	423	1	34.50	-2.20	-2.20	-14.08
C437A	chr3:49933401	HNSC	423	1	44.50	-4.87	-4.87	-3.78
G471A	chr3:49933367	DLBC	3	1	57.00	17.14	17.14	11.28
C478T	chr3:49933360	BRCA	739	1	30.50	-17.15	-9.00	0.02
C478T	chr3:49933360	COAD	281	1	66.00	-0.85	-9.00	0.02
G479T	chr3:49933359	HNSC	423	1	24.00	NA	NA	NA
G479T	chr3:49933359	SKCM	44	1	11.50	NA	NA	NA

Table 4.4: Mutation effects on *RON* exon 11 splicing in cancer patients. *(continued)*

Mutation	Genomic position (hg19)	Cohort	A	B	C	D	E	F
G499T	chr3:49933339	BLCA	242	1	44.50	0.14	0.14	-2.08
G500T	chr3:49933338	LUSC	268	1	55.00	-14.19	-14.19	5.52
A547T	chr3:49933291	UCEC	61	1	22.00	NA	NA	NA
A549G	chr3:49933289	CESC	247	1	43.00	-2.26	-2.26	0.68
G568A	chr3:49933270	BLCA	242	1	11.50	NA	NA	NA
G579T	chr3:49933259	BLCA	242	1	36.00	18.15	18.15	-4.25
G582A	chr3:49933256	COAD	281	1	32.00	-13.73	-13.73	-14.02
G582T	chr3:49933256	THCA	271	1	18.00	NA	NA	NA
C587A	chr3:49933251	HNSC	423	1	19.00	NA	NA	NA
G599A	chr3:49933239	HNSC	423	1	15.00	NA	NA	NA
C602T	chr3:49933236	COAD	281	1	47.00	-8.94	-8.94	-0.58
C615A	chr3:49933223	BRCA	739	1	19.50	NA	NA	NA
G636T	chr3:49933202	BLCA	242	1	23.00	NA	NA	NA
C640G	chr3:49933198	THCA	271	1	24.50	-29.04	-7.81	-14.82
C640G	chr3:49933198	HNSC	423	1	40.00	13.42	-7.81	-14.82
C646A	chr3:49933192	BRCA	739	1	56.00	-27.60	-27.60	-5.88
C646A	chr3:49933192	LIHC	24	1	21.00	NA	NA	NA
C646T	chr3:49933192	CESC	247	1	87.50	-5.02	-5.02	-16.12
G656T	chr3:49933182	BRCA	739	1	39.00	-4.21	-4.21	0.48
G676T	chr3:49933162	SKCM	44	1	25.00	-9.29	-9.29	-11.18
G690T	chr3:49933148	SKCM	44	1	26.00	-6.06	-6.06	-15.18
T692C	chr3:49933146	COAD	281	1	60.00	-3.73	-3.73	-25.42
A694G	chr3:49933144	COAD	281	1	133.50	3.97	3.97	-24.05

Information on 51 mutations that are present in tumours but not matched normal samples from 153 patients in The Cancer Genome Atlas(TCGA), including the mutation, its genomic coordinate (human genome version hg38), the tumour cohort of the patient with the total number of patients and of mutation-bearing patients therein, the number of RNA-seq reads supporting the PSI in the TCGA samples (average across mutation-bearing samples from the cohort), as well as changes in alternative exon (AE) skipping from TCGA (in 1-PSI) and the screen (in % isoform frequency). Abbreviations of cancer types: BLCA, Bladder Urothelial Carcinoma; BRCA, Breast Invasive Carcinoma; CESC, Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma; COAD, Colon Adenocarcinoma; DLBC, Lymphoid Neoplasm Diffuse Large B-cell Lymphoma; ESCA, Esophageal Carcinoma; HNSC, Head-Neck Squamous Cell Carcinoma; KIRC, Kidney Renal Clear Cell Carcinoma; KIRP, Kidney Renal Papillary Cell Carcinoma; LIHC, Liver Hepatocellular Carcinoma; LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Cell Carcinoma; OV, Ovarian Serous Cystadenocarcinoma; PAAD, Pancreatic Adenocarcinoma; SKCM, Skin Cutaneous Melanoma; STAD, Stomach Adenocarcinoma; THCA, Thyroid Carcinoma; THYM, Thymoma; UCEC, Uterine Corpus Endometrial Carcinoma. Only mutations from cohorts with more than 24 supporting reads on average were used in (**Figure** 3.21 **A**). Columns A, B, C, D, E and F stand for 'Total patients', 'Patients with mutation', 'Average supporting reads', 'ΔAE skipping (1-PSI; TCGA)', 'ΔAE skipping weighted average (1-PSI; TCGA)', 'ΔAE skipping (our screen)' respectively. NA implies no value available.

Table 4.5: HEK293T modelling-excluded mutations that were annotated with median isoform frequencies.

mutation	type	position	occurrence	A	B	C	D	E	F
A219C	SNV	chr3:49933619	1	-38.32	43.59	-1.69	-2.30	-0.90	-0.18
A224C	SNV	chr3:49933614	1	-54.57	64.48	-5.66	-3.50	-1.58	1.09
A299C	SNV	chr3:49933539	1	-61.30	67.03	-13.64	-4.63	-1.83	14.58
A371C	SNV	chr3:49933467	1	-61.58	73.65	-4.39	-4.77	-1.63	-1.01
A428T	SNV	chr3:49933410	10	-53.06	9.40	4.35	-4.67	3.27	30.67
A475C	SNV	chr3:49933363	1	-61.39	59.61	10.02	-4.71	-1.69	-1.69
A506C	SNV	chr3:49933332	1	-60.67	17.99	32.67	-4.70	8.14	6.49
A506G	SNV	chr3:49933332	20	-60.73	-2.88	32.99	-4.71	19.19	13.14
A513C	SNV	chr3:49933325	4	-7.78	-1.31	-3.30	-1.72	1.31	5.56
A523G	SNV	chr3:49933315	33	-60.81	-7.89	24.92	-4.71	13.45	37.40
A523T	SNV	chr3:49933315	21	-61.26	-7.94	2.57	-4.72	4.75	66.95
A535C	SNV	chr3:49933303	1	-60.13	-7.90	27.50	-4.77	23.40	21.09
A542T	SNV	chr3:49933296	10	-7.56	-0.98	2.34	-1.71	0.51	4.97
A549T	SNV	chr3:49933289	4	-4.55	1.28	-0.10	-0.69	0.68	4.20
A560G	SNV	chr3:49933278	18	-41.83	-7.65	15.38	-4.18	7.62	29.98
AG296A	DEL	chr3:49933542	1	-60.75	11.44	5.07	0.31	-1.72	45.20
AG523A	DEL	chr3:49933315	1	-60.78	-7.75	12.97	-4.74	4.80	55.15
C135CT	INS	chr3:49933703	1	1.72	-6.44	5.13	-0.31	0.18	-0.06
C139G	SNV	chr3:49933699	1	-59.70	-8.09	31.83	-4.77	30.74	9.46
C151A	SNV	chr3:49933687	1	1.14	-3.07	2.32	-0.50	0.27	-0.07
C168A	SNV	chr3:49933670	2	-60.44	-4.50	25.77	-4.23	1.50	41.44
C175A	SNV	chr3:49933663	5	-35.72	1.05	-0.61	-1.48	-0.39	3.31
C198G	SNV	chr3:49933640	1	-2.69	-2.39	5.37	0.11	-0.52	0.00
C231G	SNV	chr3:49933607	1	-38.32	43.59	-1.69	-2.30	-0.90	-0.18
C237A	SNV	chr3:49933601	1	-61.60	79.13	-9.68	-4.66	-1.79	-1.12
C281G	SNV	chr3:49933557	1	10.55	-6.34	-7.63	3.07	0.30	0.12

Table 4.5: HEK293T modelling-excluded mutations that were annotated with median isoform frequencies. *(continued)*

mutation	type	position	occurrence	A	B	C	D	E	F
C324G	SNV	chr3:49933514	2	-23.04	27.09	-0.62	-2.14	-0.75	-0.39
C344A	SNV	chr3:49933494	1	-11.11	2.97	6.06	1.59	-0.09	0.68
C396A	SNV	chr3:49933442	5	-20.04	-1.86	-0.32	-1.18	-0.26	21.04
C404G	SNV	chr3:49933434	2	-38.66	34.00	3.41	-2.07	-0.88	4.13
C410G	SNV	chr3:49933428	2	-61.30	33.62	4.18	-4.72	-1.13	29.46
C415G	SNV	chr3:49933423	3	-30.78	0.86	26.64	-2.52	3.91	1.69
C418G	SNV	chr3:49933420	4	-21.17	21.37	-7.93	-3.62	-1.05	9.24
C450G	SNV	chr3:49933388	1	-9.10	5.57	0.65	-2.52	5.15	-0.36
C455A	SNV	chr3:49933383	2	-0.09	-2.54	4.06	-2.71	1.24	0.09
C455G	SNV	chr3:49933383	1	-18.55	2.84	8.98	0.20	-0.37	7.00
C489G	SNV	chr3:49933349	1	-45.83	9.40	4.35	-4.64	3.27	32.96
C502G	SNV	chr3:49933336	1	-61.45	-8.07	-12.99	-4.75	3.47	83.65
C503G	SNV	chr3:49933335	2	-47.09	-1.36	37.75	-3.78	5.31	9.15
C512G	SNV	chr3:49933326	3	-11.41	-5.08	4.55	-2.40	4.23	9.17
C514G	SNV	chr3:49933324	2	-53.01	-6.96	35.96	-4.21	13.52	14.87
C518A	SNV	chr3:49933320	4	-26.90	-6.92	22.47	-3.17	8.45	6.84
C518G	SNV	chr3:49933320	2	-25.68	-7.75	5.05	-3.62	24.43	7.48
C520G	SNV	chr3:49933318	1	3.72	0.76	-11.00	-2.54	-0.14	9.35
C521CA	INS	chr3:49933317	1	8.51	-6.03	-10.63	-1.50	-0.24	9.45
C522A	SNV	chr3:49933316	2	-28.68	-5.73	21.04	-3.94	8.50	8.79
C522G	SNV	chr3:49933316	3	-58.41	-7.80	19.51	-4.69	9.48	38.69
C530G	SNV	chr3:49933308	9	-29.10	-6.81	19.62	-3.65	8.86	6.68
C60G	SNV	chr3:49933778	1	2.74	-3.35	-0.56	-0.41	0.32	1.42
C63A	SNV	chr3:49933775	2	-9.37	-4.69	1.84	-2.56	5.09	9.65
C64G	SNV	chr3:49933774	1	2.38	12.29	-10.75	-2.79	-0.23	-0.87

Table 4.5: HEK293T modelling-excluded mutations that were annotated with median isoform frequencies. *(continued)*

mutation	type	position	occurrence	A	B	C	D	E	F
C668G	SNV	chr3:49933170	3	-29.49	26.50	3.55	-4.32	-0.13	3.21
C85G	SNV	chr3:49933753	1	-61.53	-8.04	2.18	-4.77	2.09	69.35
CA384C	DEL	chr3:49933454	1	-60.84	-7.66	33.59	-4.70	38.74	0.28
CA442C	DEL	chr3:49933396	3	-61.36	47.67	-1.39	-4.75	-0.39	18.78
CG339C	DEL	chr3:49933499	1	13.86	-6.58	-7.95	0.21	1.56	-0.90
CG478C	DEL	chr3:49933360	1	-20.32	0.77	0.62	-1.86	0.41	20.55
CT515C	DEL	chr3:49933323	1	-58.76	65.16	-2.15	-4.33	-1.39	1.27
CT687C	DEL	chr3:49933151	1	-61.75	84.88	-15.87	-4.77	-1.86	-0.32
G104T	SNV	chr3:49933734	3	-58.93	-7.60	-0.10	-4.34	4.71	9.34
G211A	SNV	chr3:49933627	16	-59.88	-7.53	57.19	8.50	-1.44	3.12
G211C	SNV	chr3:49933627	3	-59.91	-7.26	51.55	7.01	-1.60	9.91
G215T	SNV	chr3:49933623	1	-57.52	34.17	29.85	-4.04	-1.36	-1.41
G221C	SNV	chr3:49933617	1	-57.61	-6.74	43.08	-4.52	13.08	12.14
G228C	SNV	chr3:49933610	1	-5.02	1.51	4.91	-0.51	-0.27	-0.47
G235C	SNV	chr3:49933603	1	-5.02	1.51	4.91	-0.51	-0.27	-0.47
G245GT	INS	chr3:49933593	1	22.75	-7.94	-17.78	-2.87	3.56	0.21
G260T	SNV	chr3:49933578	1	10.55	-6.34	-7.63	3.07	0.30	0.12
G284T	SNV	chr3:49933554	1	10.26	-3.48	-6.26	-1.10	0.90	-0.20
G304C	SNV	chr3:49933534	1	-9.10	5.57	0.65	-2.52	5.15	-0.36
G304T	SNV	chr3:49933534	1	-50.22	2.24	11.52	0.50	-1.36	37.06
G310C	SNV	chr3:49933528	1	13.86	-6.58	-7.95	0.21	1.56	-0.90
G342C	SNV	chr3:49933496	1	-30.17	21.60	10.22	-2.19	-0.24	0.82
G348A	SNV	chr3:49933490	9	-16.94	-7.06	2.18	-0.31	0.29	19.92
G350GT	INS	chr3:49933488	1	-7.83	-6.60	13.32	-1.78	2.27	0.26
G350T	SNV	chr3:49933488	1	-59.65	-7.11	67.69	0.51	-1.07	-0.51

Table 4.5: HEK293T modelling-excluded mutations that were annotated with median isoform frequencies. *(continued)*

mutation	type	position	occurrence	A	B	C	D	E	F
G362T	SNV	chr3:49933476	2	-31.61	0.19	31.45	-2.27	2.38	0.02
G381T	SNV	chr3:49933457	16	-8.46	-5.97	-0.11	-1.81	0.35	12.10
G422A	SNV	chr3:49933416	19	-4.95	-4.34	-1.19	-0.89	-0.06	9.13
G426C	SNV	chr3:49933412	5	-35.95	20.57	-3.15	-3.53	-0.24	8.34
G426T	SNV	chr3:49933412	26	-16.25	4.85	1.32	-1.27	-0.38	10.30
G445T	SNV	chr3:49933393	1	-61.58	73.65	-4.39	-4.77	-1.63	-1.01
G452C	SNV	chr3:49933386	2	-47.43	51.01	-0.86	-2.71	-1.39	1.45
G480C	SNV	chr3:49933358	2	-25.29	0.58	2.40	-1.61	0.71	22.88
G481A	SNV	chr3:49933357	15	-19.20	8.43	3.10	-2.14	0.15	8.29
G52GT	INS	chr3:49933786	1	-60.74	-7.62	53.47	-4.49	11.71	7.99
G524A	SNV	chr3:49933314	13	-60.79	-7.81	29.12	-4.68	8.31	30.31
G524C	SNV	chr3:49933314	1	-60.78	-7.93	10.49	-4.65	1.55	61.50
G524T	SNV	chr3:49933314	19	-61.30	-8.01	-6.25	-4.71	3.02	76.07
G525C	SNV	chr3:49933313	1	-15.48	-5.22	12.15	-1.10	2.93	6.81
G525T	SNV	chr3:49933313	1	-11.80	-2.19	10.32	-3.10	1.88	4.86
G554A	SNV	chr3:49933284	13	-45.47	-7.64	1.13	-4.18	6.10	45.60
G554T	SNV	chr3:49933284	3	-9.83	-6.69	4.13	-0.87	2.19	18.92
G555T	SNV	chr3:49933283	17	-9.25	2.35	-2.41	-1.20	-0.02	7.86
G561T	SNV	chr3:49933277	1	-11.05	-2.78	6.93	0.87	0.86	5.05
G651C	SNV	chr3:49933187	2	10.26	-3.48	-6.26	-1.10	0.90	-0.20
G679C	SNV	chr3:49933159	1	-11.34	-4.45	2.35	-3.58	14.59	2.24
G691C	SNV	chr3:49933147	1	-60.34	-7.70	30.15	-4.66	43.90	-1.43
G71T	SNV	chr3:49933767	3	2.38	-3.07	-10.73	-1.48	0.27	-0.07
GA100G	DEL	chr3:49933738	1	-61.19	-7.90	5.50	-4.68	3.95	64.28
GA245G	DEL	chr3:49933593	1	-36.61	25.86	14.44	-2.17	-1.35	0.01

Table 4.5: HEK293T modelling-excluded mutations that were annotated with median isoform frequencies. *(continued)*

mutation	type	position	occurrence	A	B	C	D	E	F
GA540G	DEL	chr3:49933298	1	-0.09	-2.54	4.06	-2.71	1.24	0.09
GA677G	DEL	chr3:49933161	1	3.72	0.76	-11.00	-2.54	-0.14	9.35
GT204G	DEL	chr3:49933634	1	-60.77	-7.54	2.53	3.32	-1.60	63.55
GT350G	DEL	chr3:49933488	1	4.92	-5.72	-2.50	0.06	1.21	1.99
GT507G	DEL	chr3:49933331	1	-7.66	-3.70	6.12	-1.84	2.38	4.73
T116G	SNV	chr3:49933722	2	-13.12	-0.54	11.17	-1.49	0.65	3.38
T117G	SNV	chr3:49933721	1	2.35	5.00	-8.92	-3.93	0.85	3.23
T293A	SNV	chr3:49933545	16	-59.81	33.87	11.13	-3.51	-1.69	7.27
T293G	SNV	chr3:49933545	2	-59.36	36.93	13.28	-2.61	-1.60	13.29
T351G	SNV	chr3:49933487	1	-47.43	51.01	-0.86	-2.71	-1.39	1.45
T420G	SNV	chr3:49933418	1	-36.57	9.70	29.52	-3.82	1.48	-0.44
T439G	SNV	chr3:49933399	1	-23.04	27.09	-0.62	-2.14	-0.75	-0.39
T446G	SNV	chr3:49933392	1	-61.39	59.61	10.02	-4.71	-1.69	-1.69
T504A	SNV	chr3:49933334	21	-52.59	-6.54	37.59	-4.43	11.61	8.53
T504C	SNV	chr3:49933334	32	-59.20	-7.52	48.23	-4.63	12.55	9.45
T504G	SNV	chr3:49933334	3	-59.64	-7.39	29.05	-4.61	12.71	4.39
T509G	SNV	chr3:49933329	2	-49.31	-4.64	36.34	-3.86	5.94	15.39
T516A	SNV	chr3:49933322	16	-60.34	-7.66	36.49	-4.66	11.83	17.73
T526A	SNV	chr3:49933312	3	-8.98	-4.34	4.19	-1.70	1.45	9.31
T536G	SNV	chr3:49933302	1	-0.09	-2.54	4.06	-2.71	1.24	0.09
T616A	SNV	chr3:49933222	18	-14.70	0.42	1.99	-1.62	0.53	5.16
T616G	SNV	chr3:49933222	1	1.72	-6.44	5.13	-0.31	0.18	-0.06
T620G	SNV	chr3:49933218	2	-39.94	9.05	18.95	-2.40	1.37	13.02
T634G	SNV	chr3:49933204	1	6.50	-4.24	-7.34	-2.60	0.81	7.03
T638G	SNV	chr3:49933200	1	-61.37	62.81	6.51	-4.68	-1.46	-1.64

Table 4.5: HEK293T modelling-excluded mutations that were annotated with median isoform frequencies. *(continued)*

mutation	type	position	occurrence	A	B	C	D	E	F
TG126T	DEL	chr3:49933712	1	-54.57	64.48	-5.66	-3.50	-1.58	1.09
TG317T	DEL	chr3:49933521	1	-11.80	-2.19	10.32	-3.10	1.88	4.86
TG498T	DEL	chr3:49933340	12	-41.52	-3.60	18.17	-3.58	5.04	12.21
TG581T	DEL	chr3:49933257	11	-9.88	5.29	-5.54	-1.37	-0.45	0.50
TG622T	DEL	chr3:49933216	1	6.50	-4.24	-7.34	-2.60	0.81	7.03

List of HEK293T mutations excluded from the modelling approach and subsequently annotated with median isoform frequencies from all minigene variants containing the given mutation as the splicing outcome. Columns A, B, C, D, E and F stand for 'Median ΔAE inclusion (%)', 'Median ΔAE skipping (%)', 'Median ΔAE full IR (%)', 'Median Δfirst IR (%)', 'Median Δsecond IR (%)' and 'Median Δother (%)' respectively.

Plasmids

Table 4.6: List of plasmids used in this study.

Plasmid ID	Name	Description
pS_020	pCDNA_RON_TAT_to_GAT+50	RON minigene with a T298G single nucleotide exchange and 50 bp extension downstream of exon 12
pS_021	pcDNA_RON_TAT_to_GAT_Lefave	RON minigene with point mutations according to (Lefave et al., 2011)
pS_022	pcDNA_RON_TAT_to_GAT_Bonomi	RON minigene with point mutations according to (Bonomi et al., 2013)
pS_056	pcDNA_HNRNPH1	HNRNPH1 overexpression construct

All plasmids are in the pcDNA 3.1 (+) vector backbone (Invitrogen).

Table 4.7: List of plasmids containing point mutations or small insertions or deletions.

Plasmid ID	Mutation
1	G272A
2	G305A
3	C307G
4	C339T
5	A356G

Table 4.7: List of plasmids containing point mutations or small insertions or deletions. *(continued)*

Plasmid ID	Mutation
6	A371G
7	C415T
8	G460A
9	A483T
10	T333C
11	G229T
12	C397A
13	GT172G
14	G433T
15	G517T
16	T389G
17	G531C
18	G380C
19	TG325T
20	G71C
21	CT612C
22	TA117T
23	T581G
24	C142A
25	G348C
26	A483C
27	G228A
28	G331C

Table 4.7: List of plasmids containing point mutations or small insertions or deletions. *(continued)*

Plasmid ID	Mutation
29	G471C
30	G480T
31	G500A
32	G555C

siRNAs

All siRNAs were purchased from Sigma Aldrich.

Table 4.8: List of siRNAs used in this study.

No.	Target	Sequence
18	No target	UGGUUUACAUGUCGACUAA[dT][dT]
22	PRPF6	GAGAAGAUUGGGCAGCUUA[dT][dT]
24	HNRNPH	GGAGCUGGCUUUGAGAGGA[dT][dT]
26	SRSF2	AAUCCAGGUCGCGAUCGAA[dT][dT]
29	SMU1	GCACGAGAAGGAUGUGAUU[dT][dT]
30	PUF60	GCAGAUGAACUCGGUGAUG[dT][dT]
52	HNRNPF	UGAGAAGGCUCUAGGGAAA[dT][dT]

Antibodies

Table 4.9: List of antibodies used in this study.

Antigen	Dilution	Origin	Product number	Supplier
GFP	1:500 (0.39)	mouse	sc-9996	Santa Cruz
GFP trap	20 μl slurry / sample	lama	gtma-100	Chromotek
HNRNPA1	1:20,000	mouse	R4528	Sigma Aldrich
HNRNPH	1:10,000	rabbit	AB10374	Abcam
HNRNPF	1:500 (0.39)	mouse	3H4	Santa Cruz
mouse IgG	1:5,000	horse	7076	Cell Signalling
rabbit IgG	1:5,000	goat	7074	Cell Signalling

Oligonucleotides (next page)

Oligonucleotides were obtained either from Sigma-Aldrich or Integrated DNA Technologies

No.	Name	Sequence (5′-3′)	Purpose
oJ303	minigene_cloning_fwd	CCCAAGCTTTGTGAGAGGCAGCTTCCAGA	Cloning of wt *RON* minigene
oS111	minigene_cloning_rev	cagTCTAGANNNNNNNNNNNNNNNGGATCCgccattggttgggggtaggggctgattaaaggtagg	Cloning of wt *RON* minigene
oS428	BamHI_HNRNPH1_fwd	catGGATCCaccatgatgttgggcacggaagg	Cloning of HNRNPH1 overexpression construct
oS429	XbaI_HNRNPH1_rev	cattctagactatgcaatgtttgattgaaaatc	Cloning of HNRNPH1 overexpression construct
oS66	RT-PCR_minigene_fwd	TGCCAACCTAGTTCCACTGA	RT-PCR for *RON* minigene
oS67	RT-PCR_minigene_rev	GCAACTAGAAGGCACAGTCG	RT-PCR for *RON* minigene
oS44	RT-PCR_endo_fwd	CCTGAATATGTGGTCCGAGACCCCCAG	RT-PCR for endogenous *RON* gene
oS45	RT-PCR_endo_rev	CTAGCTGCTTCCTCCGCCACCAGTA	RT-PCR for endogenous *RON* gene
oS237	RT-PCR-endo_fwd2	GGGCAGTGGAAAGCAGGTGTGAG	RT-PCR for endogenous *RON* gene in minigene transfected cells
oS118	RON A	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCTATAGGGAGACCCAAGCTT	Illumina fwd sequencing primer for DNA-seq
oS119	RON B	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGTTCCACTGAAGCCTGAG	Illumina fwd sequencing primer for DNA-seq and RNA-seq
oS120	RON C	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNAGCTGCCAGCACGAGTTC	Illumina fwd sequencing primer for DNA-seq
oS138	RON D	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGAATCTGAGTGCCCGAGG	Illumina fwd sequencing primer for DNA-seq

No.	Name	Sequence (5′-3′)	Purpose
oS105	RON E	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNctactggctggtcctcatga	Illumina fwd sequencing primer for DNA-seq
oS106	P5 SOLEXA RON	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNATAGAATAGGGCCCTCTAGA	Illumina rev sequencing primer for DNA-seq and RNA-seq
oS153	HNRNPH1_qPCR_f	GTACACATGCGGGGATTACC	RT-qPCR validation of *HNRNPH1* KD
oS154	HNRNPH1_qPCR_r	CGAACTCGACATCTGCTTCA	RT-qPCR validation of *HNRNPH1* KD
oS155	PRPF6_qPCR_f	CCGGAGAGAACCATACCTCA	RT-qPCR validation of *PRPF6* KD
oS156	PRPF6_qPCR_r	CTGTGCCAGGTGTCATCAGT	RT-qPCR validation of *PRPF6* KD
oS157	PUF60_qPCR_f	GCAAGATCAAGTCCTGCACA	RT-qPCR validation of *PUF60* KD
oS158	PUF60_qPCR_r	GGTTCATGGAAGACACAGCA	RT-qPCR validation of *PUF60* KD
oS159	SMU1_qPCR_f	TCCTCTGCATGTGTTTCAGC	RT-qPCR validation of *SMU1* KD
oS160	SMU1_qPCR_r	TGTGCCCTCTCAAATCTCCT	RT-qPCR validation of *SMU1* KD
oS161	SRSF2_qPCR_f	GCACTAGGCGCAGTTGTGTA	RT-qPCR validation of *SRSF2* KD
oS162	SRSF2_qPCR_r	CAATCGGGAGAAAACAGGAA	RT-qPCR validation of *SRSF2* KD
oS253	HNRNPH2_qPCR_fwd	GTGGTGGTTATGGAGGTGGT	RT-qPCR validation of *HNRNPH2* KD
oS254	HNRNPH2_qPCR_rev	GTGCTCCTTCTCTACCTAAGCA	RT-qPCR validation of *HNRNPH2* KD
oS289	HNRNPF_qPCR_f	GCCTGGTAGCAACAGAAACC	RT-qPCR validation of *HNRNPF* KD
oS290	HNRNPF_qPCR_r	GTGATCTTGGGTGTGGCTTT	RT-qPCR validation of *HNRNPF* KD

No.	Name	Sequence (5′-3′)	Purpose
oS428	BamHI_HNRNPH1_fwd	catGGATCCaccatgatgttgggcacggaagg	Cloning of HNRNPH1 overexpression construct
oS429	XbaI_HNRNPH1_rev	cattctagactatgcaatgtttgattgaaaatc	Cloning of HNRNPH1 overexpression construct

Other supplementary tables

The following tables list the contents of the solutions and buffers used in the study.

Supplementary Table 1: Lysis buffer composition.

Component	Final concentration
Tris-HCl, pH 7.5	50 mM
NaCl	150 mM
EDTA	1 mM
NP-40	1% (v/v)
sodium deoxycholate	0.1% (w/v)
NEM (N-ethylmaleimide)	10 mM
cOmplete™ Protease Inhibitor Cocktail	1 tablet/ 10 ml
β-glycerophosphate	5 mM

Supplementary Table 2: Ponceau red staining solution.

Component	Final concentration
Ponceau S	0.1% (w/v)
Acetic acid	5% (v/v)

Supplementary Table 3: Buffer A+ composition.

Component	Final concentration
Hepes KOH pH 7.6	10 mM
$MgCl_2$	1.5 mM
KCl	10 mM
Igepal CA630	0.1% (v/v)
cOmplete™ Protease Inhibitor Cocktail	1 tablet/ 10 ml
DTT	0.5 mM

Supplementary Table 4: Buffer C+ composition.

Component	Final concentration
Hepes KOH pH 7.6	20 mM
$MgCl_2$	2 mM
NaCl	420 mM
Glycerol	20% (v/v)
Igepal CA630	0.1% (v/v)
cOmplete™ Protease Inhibitor Cocktail	1 tablet/ 10 ml
DTT	0.5 mM

Supplementary Table 5: Binding buffer composition.

Component	Final concentration
Tris-HCl, pH 7.5	20 mM
LiCl	1 M
EDTA	2 mM

Supplementary Table 6: Washing buffer B composition.

Component	Final concentration
Tris-HCl, pH 7.5	10 mM
LiCl	0.15 M
EDTA	1 mM

Supplementary Table 7: RNA binding buffer composition.

Component	Final concentration
NaCl	100 mM
Hepes/ HCl, pH 7.6	50 mM
Igepal CA630	0.5% (v/v)
$MgCl_2$	10 mM

Supplementary Table 8: RNA wash buffer composition.

Component	Final concentration
NaCl	250 mM
Hepes/ HCl, pH 7.6	50 mM
Igepal CA630	0.5% (v/v)
$MgCl_2$	10 mM

Supplementary Table 9: Oil-surfactant mixture for emulsion PCR.

Component	Amount	Final concentration
Span 80	2.25 ml	4.5% (vol/vol)
Tween 80	200 μl	0.4% (vol/vol)

Component	Amount	Final concentration
Triton X-100	25 μl	0.05% (vol/vol)
Mineral oil	to 50 ml	

Citations

- Figures in the Introduction section were created using BioRender.com
- The references were added and managed using Papers 3.

References

Adamson, S. I.; Zhan, L.; Graveley, B. R., 2018: Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome biology.*, **19**, 71.

Ahsan, K. B.; Masuda, A.; Rahman, M. A.; Takeda, J. i.; Nazim, M.; Ohkawara, B.; Ito, M.; Ohno, K., 2017: SRSF1 suppresses selection of intron-distal 5′ splice site of DOK7 intron 4 to generate functional full-length Dok-7 protein. *Scientific Reports.*, **7**, 10446.

Akerman, M.; David-Eden, H.; Pinter, R. Y.; Mandel-Gutfreund, Y., 2009: A computational approach for genome-wide mapping of splicing factor binding sites. *Genome biology.*, **10**, 1–14.

Amin, E. M.; Oltean, S.; Hua, J.; Gammons, M. V. R.; Hamdollah-Zadeh, M.; Welsh, G. I.; Cheung, M. K.; Ni, L.; Kase, S.; Rennel, E. S.; Symonds, K. E.; Nowak, D. G.; Royer-Pokora, B.; Saleem, M. A.; Hagiwara, M.; Schumacher, V. A.; Harper, S. J.; Hinton, D. R.; Bates, D. O. et al., 2011: WT1 mutants reveal SRPK1 to be a downstream angiogenesis target by altering VEGF splicing. *Cancer cell.*, **20**, 768–780.

Anantharaman, V.; Koonin, E. V.; Aravind, L., 2002: Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic acids research.*, **30**, 1427–1464.

Aoki, S.; Takahashi, T.; Matsumoto, K.; Kiyohara, T.; Nakamura, T., 1996: Cloning and expression of Xenopus HGF-like protein (HLP) and Ron/HLP receptor implicate their involvement in early neural development. *Biochemical and biophysical research communications.*, **224**, 564–573.

Ardui, S.; Ameur, A.; Vermeesch, J. R.; Hestand, M. S., 2018: Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic acids research.*, **46**, 2159–2168.

Auweter, S. D.; Fasan, R.; Reymond, L.; Underwood, J. G.; Black, D. L.; Pitsch, S.; Allain, F. H. T., 2006: Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. *The EMBO journal.*, **25**, 163–173.

Avsec, žiga; Kreuzhuber, R.; Israeli, J.; Xu, N.; Cheng, J.; Shrikumar, A.; Banerjee, A.; Kim, D. S.; Beier, T.; Urban, L.; Kundaje, A.; Stegle, O.; Gagneur, J., 2019: The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nature biotechnology.*, **37**, 592–600.

Baeza-Centurion, P.; Miñana, B.; Schmiedel, J. M.; Valcárcel, J.; Lehner, B., 2019: Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell.*, **176**, 549–563.e23.

Bamford, S.; Dawson, E.; Forbes, S.; Clements, J.; Pettett, R.; Dogan, A.; Flanagan, A.; Teague, J.; Futreal, P. A.; Stratton, M. R.; Wooster, R., 2004: The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British journal of cancer.*, **91**, 355–358.

Bao, S.; Moakley, D. F.; Zhang, C., 2019: The Splicing Code Goes Deep. *Cell.*, **176**, 414–416.

Baralle, F. E.; Giudice, J., 2017: Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology.*, **18**, 437–451.

Barash, Y.; Calarco, J. A.; Gao, W.; Pan, Q.; Wang, X.; Shai, O.; Blencowe, B. J.; Frey, B. J., 2010: Deciphering the splicing code. *Nature.*, **465**, 53–59.

Barbosa-Morais, N. L.; Irimia, M.; Pan, Q.; Xiong, H. Y.; Gueroussov, S.; Lee, L. J.; Slobodeniuc, V.; Kutter, C.; Watt, S.; Colak, R.; Kim, T.; Misquitta-Ali, C. M.; Wilson, M. D.;

Kim, P. M.; Odom, D. T.; Frey, B. J.; Blencowe, B. J., 2012: The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science (New York, N.Y.)*., **338**, 1587–1593.

Bartys, N.; Kierzek, R.; Lisowiec-Wachnicka, J., 2019: The regulation properties of RNA secondary structure in alternative splicing. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*., 194401.

Bassett, D. I., 2003: Identification and developmental expression of a macrophage stimulating 1/ hepatocyte growth factor-like 1 orthologue in the zebrafish. *Development genes and evolution*., **213**, 360–362.

Bedard, K. M.; Walter, B. L.; Semler, B. L., 2004: Multimerization of poly(rC) binding protein 2 is required for translation initiation mediated by a viral IRES. *RNA*., **10**, 1266–1276.

Bedolla, R. G.; Shah, D. P.; Huang, S. B.; Reddick, R. L.; Ghosh, R.; Kumar, A. P., 2019: Receptor tyrosine kinase recepteur d'origine nantais as predictive marker for aggressive prostate cancer in African Americans. *Molecular Carcinogenesis*., **58**, 854–861.

Bedrat, A.; Lacroix, L.; Mergny, J. L., 2016: Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic acids research*., **44**, 1746–1759.

Benvenuti, S.; Comoglio, P. M., 2007: The MET receptor tyrosine kinase in invasion and metastasis. *Journal of cellular physiology*., **213**, 316–325.

Ben-Yishay, R.; Shav-Tal, Y., 2019: The dynamic lifecycle of mRNA in the nucleus. *Current Opinion in Cell Biology*., **58**, 69–75.

Berglund, J. A.; Abovich, N.; Rosbash, M., 1998: A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. *Genes & development*., **12**, 858–867.

Biffi, G.; Tannahill, D.; Miller, J.; Howat, W. J.; Balasubramanian, S., 2014: Elevated levels of G-quadruplex formation in human stomach and liver cancer tissues. *PloS one*., **9**, e102711.

Black, D. L., 2003: Mechanisms of alternative pre-messenger RNA splicing. *Annual review of biochemistry.*, **72**, 291–336.

Blanchette, M.; Chabot, B., 1999: Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization. *The EMBO journal.*, **18**, 1939–1952.

Bolger, A. M.; Lohse, M.; Usadel, B., 2014: Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England).*, **30**, 2114–2120.

Bonomi, S.; Matteo, A. di; Buratti, E.; Cabianca, D. S.; Baralle, F. E.; Ghigna, C.; Biamonti, G., 2013: HnRNP A1 controls a splicing regulatory circuit promoting mesenchymal-to-epithelial transition. *Nucleic acids research.*, **41**, 8665–8679.

Boukakis, G.; Patrinou-Georgoula, M.; Lekarakou, M.; Valavanis, C.; Guialis, A., 2010: Deregulated expression of hnRNP A/B proteins in human non-small cell lung cancer: parallel assessment of protein and mRNA levels in paired tumour/non-tumour tissues. *BMC cancer.*, **10**, 434–413.

Braun, S.; Enculescu, M.; Setty, S. T.; Cortés-López, M.; Almeida, B. P. de; Sutandy, F. X. R.; Schulz, L.; Busch, A.; Seiler, M.; Ebersberger, S.; Barbosa-Morais, N. L.; Legewie, S.; König, J.; Zarnack, K., 2018: Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. *Nature Communications.*, **9**, 3315.

Braunschweig, U.; Barbosa-Morais, N. L.; Pan, Q.; Nachman, E. N.; Alipanahi, B.; Gonatopoulos-Pournatzis, T.; Frey, B.; Irimia, M.; Blencowe, B. J., 2014: Widespread intron retention in mammals functionally tunes transcriptomes. *Genome research.*, **24**, 1774–1786.

Buratti, E.; Stuani, C.; De Prato, G.; Baralle, F. E., 2007: SR protein-mediated inhibition of CFTR exon 9 inclusion: molecular characterization of the intronic splicing silencer. *Nucleic acids research.*, **35**, 4359–4368.

Camats, M.; Guil, S.; Kokolo, M.; Bach-Elias, M., 2008: P68 RNA Helicase (DDX5) Alters

Activity of Cis- and Trans-Acting Factors of the Alternative Splicing of H-Ras. *PloS one.*, **3**.

Cammas, A.; Millevoi, S., 2017: RNA G-quadruplexes: emerging mechanisms in disease. *Nucleic acids research.*, **45**, 1584–1595.

Caputi, M.; Zahler, A. M., 2002: SR proteins and hnRNP H regulate the splicing of the HIV-1 tev-specific exon 6D. *The EMBO journal.*, **21**, 845–855.

Carithers, L. J.; Ardlie, K.; Barcus, M.; Branton, P. A.; Britton, A.; Buia, S. A.; Compton, C. C.; DeLuca, D. S.; Peter-Demchok, J.; Gelfand, E. T.; Guan, P.; Korzeniewski, G. E.; Lockhart, N. C.; Rabiner, C. A.; Rao, A. K.; Robinson, K. L.; Roche, N. V.; Sawyer, S. J.; Segre, A. V. et al., 2015: A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking.*, **13**, 311–319.

Cartegni, L.; Krainer, A. R., 2002: Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nature genetics.*, **30**, 377–384.

Casas-Finet, J. R.; Smith Jr, J. D.; Kumar, A.; Kim, J. G.; Wilson, S. H.; Karpel, R. L., 1993: Mammalian Heterogeneous Ribonucleoprotein A1 and its Constituent Domains: Nucleic Acid Interaction, Structural Stability and Self-association. *Journal of molecular biology.*, **229**, 873–889.

Castello, A.; Fischer, B.; Hentze, M. W.; Preiss, T., 2013: RNA-binding proteins in Mendelian disease. *Trends in genetics : TIG.*, **29**, 318–327.

Chakedis, J.; French, R.; Babicky, M.; Jaquish, D.; Howard, H.; Mose, E.; Lam, R.; Holman, P.; Miyamoto, J.; Walterscheid, Z.; Lowy, A. M., 2016: A novel protein isoform of the RON tyrosine kinase receptor transforms human pancreatic duct epithelial cells. *Oncogene.*, **35**, 3249–3259.

Chaudhury, A.; Chander, P.; Howe, P. H., 2010: Heterogeneous nuclear ribonucleoproteins (hnRNPs) in cellular processes: Focus on hnRNP E1's multifunctional regulatory roles. *RNA.*, **16**, 1449–1462.

Chen, L.; Bush, S. J.; Tovar-Corona, J. M.; Castillo-Morales, A.; Urrutia, A. O., 2014: Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. *Molecular biology and evolution.*, **31**, 1402–1413.

Cheng, J.; Maier, K. C.; Avsec, žiga; Rus, P.; Gagneur, J., 2017: Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast. *RNA.*, **23**, 1648–1659.

Cheng, J.; Nguyen, T. Y. D.; Cygan, K. J.; Çelik, M. H.; Fairbrother, W. G.; Avsec, žiga; Gagneur, J., 2019: MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome biology.*, **20**, 48.

Cheung, R.; Insigne, K. D.; Yao, D.; Burghard, C. P.; Wang, J.; Hsiao, Y. H. E.; Jones, E. M.; Goodman, D. B.; Xiao, X.; Kosuri, S., 2019: A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. *Molecular cell.*, **73**, 183–194.e8.

Ching, T.; Himmelstein, D. S.; Beaulieu-Jones, B. K.; Kalinin, A. A.; Do, B. T.; Way, G. P.; Ferrero, E.; Agapow, P. M.; Zietz, M.; Hoffman, M. M.; Xie, W.; Rosen, G. L.; Lengerich, B. J.; Israeli, J.; Lanchantin, J.; Woloszynek, S.; Carpenter, A. E.; Shrikumar, A.; Xu, J. et al., 2018: Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society, Interface.*, **15**.

Clarke, J.; Wu, H. C.; Jayasinghe, L.; Patel, A.; Reid, S.; Bayley, H., 2009: Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology.*, **4**, 265–270.

Clerte, C.; Hall, K. B., 2006: Characterization of multimeric complexes formed by the human PTB1 protein on RNA. *RNA.*, **12**, 457–475.

Cléry, A.; Blatter, M.; Allain, F. H. T., 2008: RNA recognition motifs: boring? Not quite. *Current opinion in structural biology.*, **18**, 290–298.

Collesi, C.; Santoro, M. M.; Gaudino, G.; Comoglio, P. M., 1996: A splicing variant of the

RON transcript induces constitutive tyrosine kinase activity and an invasive phenotype. *Molecular and Cellular Biology.*, **16**, 5518–5526.

Coltri, P. P.; Santos, M. G. P. dos; Silva, G. H. G. da, 2019: Splicing and cancer: Challenges and opportunities. *Wiley Interdisciplinary Reviews: RNA.*, **10**, e1527.

Conlon, E. G.; Lu, L.; Sharma, A.; Yamazaki, T.; Tang, T.; Shneider, N. A.; Manley, J. L., 2016: The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains. *eLife.*, **5**, 345.

Cooper, T. A., 2005: Use of minigene systems to dissect alternative splicing elements. *Methods (San Diego, Calif.).*, **37**, 331–340.

Cordin, O.; Beggs, J. D., 2013: RNA helicases in splicing. *RNA biology.*, **10**, 83–95.

Corioni, M.; Antih, N.; Tanackovic, G.; Zavolan, M.; Krämer, A., 2011: Analysis of in situ pre-mRNA targets of human splicing factor SF1 reveals a function in alternative splicing. *Nucleic acids research.*, **39**, 1868–1879.

Correa, B. R.; Araujo, P. R. de; Qiao, M.; Burns, S. C.; Chen, C.; Schlegel, R.; Agarwal, S.; Galante, P. A. F.; Penalva, L. O. F., 2016: Functional genomics analyses of RNA-binding proteins reveal the splicing regulator SNRPB as an oncogenic candidate in glioblastoma. *Genome biology.*, **17**, 1–16.

Cummings, B. B.; Marshall, J. L.; Tukiainen, T.; Lek, M.; Donkervoort, S.; Foley, A. R.; Bolduc, V.; Waddell, L. B.; Sandaradura, S. A.; O'Grady, G. L.; Estrella, E.; Reddy, H. M.; Zhao, F.; Weisburd, B.; Karczewski, K. J.; O'Donnell-Luria, A. H.; Birnbaum, D.; Sarkozy, A.; Hu, Y. et al., 2017: Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science translational medicine.*, **9**, eaal5209.

Daguenet, E.; Dujardin, G.; Valcárcel, J., 2015: The pathogenicity of splicing defects: mechanistic insights into pre-mRNA processing inform novel therapeutic approaches. *EMBO reports.*, **16**, 1640–1655.

Danckwardt, S.; Neu-Yilik, G.; Thermann, R.; Frede, U.; Hentze, M. W.; Kulozik, A. E., 2002: Abnormally spliced beta-globin mRNAs: a single point mutation generates transcripts sensitive and insensitive to nonsense-mediated mRNA decay. *Blood.*, **99**, 1811–1816.

Dardenne, E.; Polay Espinoza, M.; Fattet, L.; Germann, S.; Lambert, M. P.; Neil, H.; Zonta, E.; Mortada, H.; Gratadou, L.; Deygas, M.; Chakrama, F. Z.; Samaan, S.; Desmet, F. O.; Tranchevent, L. C.; Dutertre, M.; Rimokh, R.; Bourgeois, C. F.; Auboeuf, D., 2014: RNA helicases DDX5 and DDX17 dynamically orchestrate transcription, miRNA, and splicing programs in cell differentiation. *Cell reports.*, **7**, 1900–1913.

Daubner, G. M.; Cléry, A.; Allain, F. H. T., 2013: RRMRNA recognition: NMR or crystallography…and new findings. *Current opinion in structural biology.*, **23**, 100–108.

De Conti, L.; Baralle, M.; Buratti, E., 2013: Exon and intron definition in pre-mRNA splicing. *Wiley Interdisciplinary Reviews: RNA.*, **4**, 49–60.

Decorsière, A.; Cayrel, A.; Vagner, S.; Millevoi, S., 2011: Essential role for the interaction between hnRNP H/F and a G quadruplex in maintaining p53 pre-mRNA 3′-end processing and function during DNA damage. *Genes & development.*, **25**, 220–225.

DeLuca, D. S.; Segre, A. V.; Sullivan, T. J.; Young, T. R.; Gelfand, E. T.; Trowbridge, C. A.; Maller, J. B.; Tukiainen, T.; Lek, M.; Ward, L. D.; Kheradpour, P.; Iriarte, B.; Meng, Y.; Palmer, C. D.; Esko, T.; Winckler, W.; Hirschhorn, J. N.; MacArthur, D. G.; Getz, G. et al., 2015: The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (New York, N.Y.).*, **348**, 648–660.

Ding, J.; Hayashi, M. K.; Zhang, Y.; Manche, L.; Krainer, A. R.; Xu, R. M., 1999: Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. *Genes & development.*, **13**, 1102–1115.

Djebali, S.; Davis, C. A.; Merkel, A.; Dobin, A.; Lassmann, T.; Mortazavi, A.; Tanzer, A.; Lagarde, J.; Lin, W.; Schlesinger, F.; Xue, C.; Marinov, G. K.; Khatun, J.; Williams, B. A.; Zaleski, C.; Rozowsky, J.; Röder, M.; Kokocinski, F.; Abdelhamid, R. F. et al., 2012: Landscape of transcription in human cells. *Nature.*, **489**, 101–108.

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 (Oxford, England).*, **29**, 15–21.

Dohm, J. C.; Lottaz, C.; Borodina, T.; Himmelbauer, H., 2008: Substantial biases in ultrashort read data sets from high-throughput DNA sequencing. *Nucleic acids research.*, **36**, e105–e105.

Dominguez, C.; Fisette, J. F.; Chabot, B.; Allain, F. H. T., 2010: Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. *Nature Structural & Molecular Biology.*, **17**, 853–861.

Dominguez, D.; Freese, P.; Alexis, M. S.; Su, A.; Hochman, M.; Palden, T.; Bazile, C.; Lambert, N. J.; Van Nostrand, E. L.; Pratt, G. A.; Yeo, G. W.; Graveley, B. R.; Burge, C. B., 2018: Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular cell.*, **70**, 854–+.

Dorak, M. T., 2006: In MT Dorak (Ed.), Real-time PCR.

Dorak, M. T., 2007: *Real-time PCR*. Garland Science.

Dvinge, H.; Kim, E.; Abdel-Wahab, O.; Bradley, R. K., 2016: RNA splicing factors as oncoproteins and tumour suppressors. *Nature reviews. Cancer.*, **16**, 413–430.

Eid, J.; Fehr, A.; Gray, J.; Luong, K.; Lyle, J.; Otto, G.; Peluso, P.; Rank, D.; Baybayan, P.; Bettman, B.; Bibillo, A.; Bjornson, K.; Chaudhuri, B.; Christians, F.; Cicero, R.; Clark, S.; Dalal, R.; deWinter, A.; Dixon, J. et al., 2009: Real-Time DNA Sequencing from Single Polymerase Molecules. *Science (New York, N.Y.).*, **323**, 133–138.

Erkelenz, S.; Mueller, W. F.; Evans, M. S.; Busch, A.; Schöneweis, K.; Hertel, K. J.; Schaal, H., 2013: Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA.*, **19**, 96–102.

Faham, N.; Welm, A. L., 2016: RON Signaling Is a Key Mediator of Tumor Progression in Many Human Cancers. *Cold Spring Harbor Symposia on Quantitative Biology.*, **81**, 177–188.

Fica, S. M.; Tuttle, N.; Novak, T.; Li, N. S.; Lu, J.; Koodathingal, P.; Dai, Q.; Staley, J. P.; Piccirilli, J. A., 2013: RNA catalyses nuclear pre-mRNA splicing. *Nature.*, **503**, 229–234.

Fica, S. M.; Nagai, K., 2017: Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. *Nature Structural & Molecular Biology.*, **24**, 791–799.

Findlay, G. M.; Daza, R. M.; Martin, B.; Zhang, M. D.; Leith, A. P.; Gasperini, M.; Janizek, J. D.; Huang, X.; Starita, L. M.; Shendure, J., 2018: Accurate classification of BRCA1 variants with saturation genome editing. *Nature.*, **562**, 217–222.

Fisette, J. F.; Montagna, D. R.; Mihailescu, M. R.; Wolfe, M. S., 2012: A G-rich element forms a G-quadruplex and regulates BACE1 mRNA alternative splicing. *Journal of neurochemistry.*, **121**, 763–773.

Fu, X. D.; Ares, M., 2014: Context-dependent control of alternative splicing by RNA-binding proteins. *Nature reviews. Genetics.*, **15**, 689–701.

Fuselli, S.; Baptista, R. P.; Panziera, A.; Magi, A.; Guglielmi, S.; Tonin, R.; Benazzo, A.; Bauzer, L. G.; Mazzoni, C. J.; Bertorelle, G., 2018: A new hybrid approach for MHC genotyping: high-throughput NGS and long read MinION nanopore sequencing, with application to the non-model vertebrate Alpine chamois (Rupicapra rupicapra). *Heredity.*, **121**, 293–303.

Gao, K.; Masuda, A.; Matsuura, T.; Ohno, K., 2008: Human branch point consensus sequence is yUnAy. *Nucleic acids research.*, **36**, 2257–2267.

Garg, K.; Green, P., 2007: Differing patterns of selection in alternative and constitutive splice sites. *Genome research.*, **17**, 1015–1022.

Gartner, J. J.; Parker, S. C. J.; Prickett, T. D.; Dutton-Regester, K.; Stitzel, M. L.; Lin, J. C.; Davis, S.; Simhadri, V. L.; Jha, S.; Katagiri, N.; Gotea, V.; Teer, J. K.; Wei, X.; Morken,

M. A.; Bhanot, U. K.; Program, N. C. S.; Chen, G.; Elnitski, L. L.; Davies, M. A. et al., 2013: Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proceedings of the National Academy of Sciences of the United States of America.*, **110**, 13481–13486.

Gaudino, G.; Follenzi, A.; Naldini, L.; Collesi, C.; Santoro, M.; Gallo, K. A.; Godowski, P. J.; Comoglio, P. M., 1994: RON is a heterodimeric tyrosine kinase receptor activated by the HGF homologue MSP. *The EMBO journal.*, **13**, 3524–3532.

Gaudino, G.; Avantaggiato, V.; Follenzi, A.; Acampora, D.; Simeone, A.; Comoglio, P. M., 1995: The proto-oncogene RON is involved in development of epithelial, bone and neuro-endocrine tissues. *Oncogene.*, **11**, 2627–2637.

Gerstberger, S.; Hafner, M.; Ascano, M.; Tuschl, T., 2014: Evolutionary Conservation and Expression of Human RNA-Binding Proteins and Their Role in Human Genetic Disease. In: *Systems biology of rna binding proteins*. Springer, New York, NY, New York, NY, pp. 1–55.

Ghigna, C.; Giordano, S.; Shen, H.; Benvenuto, F.; Castiglioni, F.; Comoglio, P. M.; Green, M. R.; Riva, S.; Biamonti, G., 2005: Cell Motility Is Controlled by SF2/ASF through Alternative Splicing of the Ron Protooncogene. *Molecular cell.*, **20**, 881–890.

Giudice, G.; Sánchez-Cabo, F.; Torroja, C.; Lara-Pezzi, E., 2016: ATtRACT-a database of RNA-binding proteins and associated motifs. *Database : the journal of biological databases and curation.*, **2016**, baw035.

Glisovic, T.; Bachorik, J. L.; Yong, J.; Dreyfuss, G., 2008: RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters.*, **582**, 1977–1986.

Goren, A.; Ram, O.; Amit, M.; Keren, H.; Lev-Maor, G.; Vig, I.; Pupko, T.; Ast, G., 2006: Comparative Analysis Identifies Exonic Splicing Regulatory SequencesThe Complex Definition of Enhancers and Silencers. *Molecular cell.*, **22**, 769–781.

Gotea, V.; Gartner, J. J.; Qutob, N.; Elnitski, L.; Samuels, Y., 2015: The functional relevance of somatic synonymous mutations in melanoma and other cancers. *Pigment cell & melanoma research.*, **28**, 673–684.

Graveley, B. R., 2000: Sorting out the complexity of SR protein functions. *RNA.*, **6**, 1197–1211.

Greenman, C.; Wooster, R.; Futreal, P. A.; Stratton, M. R.; Easton, D. F., 2006: Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics.*, **173**, 2187–2198.

Grodecká, L.; Buratti, E.; Freiberger, T., 2017: Mutations of Pre-mRNA Splicing Regulatory Elements: Are Predictions Moving Forward to Clinical Diagnostics? *International journal of molecular sciences.*, **18**, 1668.

Gueroussov, S.; Weatheritt, R. J.; O'Hanlon, D.; Lin, Z. Y.; Narula, A.; Gingras, A. C.; Blencowe, B. J., 2017: Regulatory Expansion in Mammals of Multivalent hnRNP Assemblies that Globally Control Alternative Splicing. *Cell.*, **170**, 324–339.e23.

Guo, J. U.; Bartel, D. P., 2016: RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science (New York, N.Y.).*, **353**, aaf5371.

Hacht, A. von; Seifert, O.; Menger, M.; Schütze, T.; Arora, A.; Konthur, Z.; Neubauer, P.; Wagner, A.; Weise, C.; Kurreck, J., 2014: Identification and characterization of RNA guanine-quadruplex binding proteins. *Nucleic acids research.*, **42**, 6630–6644.

Han, K.; Yeo, G.; An, P.; Burge, C. B.; Grabowski, P. J., 2005: A Combinatorial Code for Splicing Silencing: UAGG and GGGG Motifs. *PLoS biology.*, **3**, e158.

Havens, M. A.; Duelli, D. M.; Hastings, M. L., 2013: Targeting RNA splicing for disease therapy. *Wiley Interdisciplinary Reviews: RNA.*, **4**, 247–266.

Heiner, M.; Hui, J.; Schreiner, S.; Hung, L. H.; Bindereif, A., 2010: HnRNP L-mediated regulation of mammalian alternative splicing by interference with splice site recognition. *RNA biology.*, **7**, 56–64.

Howard, J. M.; Sanford, J. R., 2015: The RNAissance family: SR proteins as multifaceted regulators of gene expression. *Wiley Interdisciplinary Reviews: RNA.*, **6**, 93–110.

Hsiao, Y. H. E.; Bahn, J. H.; Lin, X.; Chan, T. M.; Wang, R.; Xiao, X., 2016: Alternative splicing modulated by genetic variants demonstrates accelerated evolution regulated by highly conserved proteins. *Genome research.*, **26**, 440–450.

Huang, H.; Zhang, J.; Harvey, S. E.; Hu, X.; Cheng, C., 2017: RNA G-quadruplex secondary structure promotes alternative splicing via the RNA-binding protein hnRNPF. *Genes & development.*, **31**, 2296–2309.

Huelga, S. C.; Vu, A. Q.; Arnold, J. D.; Liang, T. Y.; Liu, P. P.; Yan, B. Y.; Donohue, J. P.; Shiue, L.; Hoon, S.; Brenner, S.; Ares, M.; Yeo, G. W., 2012: Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell reports.*, **1**, 167–178.

Hug, N.; Longman, D.; Cáceres, J. F., 2016: Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic acids research.*, **44**, 1483–1495.

Hui, J.; Stangl, K.; Lane, W. S.; Bindereif, A., 2003: HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nature Structural & Molecular Biology.*, **10**, 33–37.

Hurst, L. D.; Batada, N. N., 2017: Depletion of somatic mutations in splicing-associated sequences in cancer genomes. *Genome biology.*, **18**, 213.

Ibrahim, E. C.; Schaal, T. D.; Hertel, K. J.; Reed, R.; Maniatis, T., 2005: Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. *Proceedings of the National Academy of Sciences of the United States of America.*, **102**, 5002–5007.

Jablonski, J. A.; Buratti, E.; Stuani, C.; Caputi, M., 2008: The secondary structure of the human immunodeficiency virus type 1 transcript modulates viral splicing and infectivity. *Journal of virology.*, **82**, 8038–8050.

Jacob, A. G.; Smith, C. W. J., 2017: Intron retention as a component of regulated gene expression programs. *Human Genetics.*, **136**, 1043–1057.

Jaganathan, K.; Kyriazopoulou Panagiotopoulou, S.; McRae, J. F.; Darbandi, S. F.; Knowles, D.; Li, Y. I.; Kosmicki, J. A.; Arbelaez, J.; Cui, W.; Schwartz, G. B.; Chow, E. D.; Kanterakis, E.; Gao, H.; Kia, A.; Batzoglou, S.; Sanders, S. J.; Farh, K. K. H., 2019: Predicting Splicing from Primary Sequence with Deep Learning. *Cell.*, **176**, 535–548.e24.

Johnson, T. L.; Vilardell, J., 2012: Regulated pre-mRNA splicing: the ghostwriter of the eukaryotic genome. *Biochimica et biophysica acta.*, **1819**, 538–545.

Julien, P.; Miñana, B.; Baeza-Centurion, P.; Valcárcel, J.; Lehner, B., 2016: The complete local genotypephenotype landscape for the alternative splicing of a human exon. *Nature Communications.*, **7**, 11558.

Jung, H.; Lee, D.; Lee, J.; Park, D.; Kim, Y. J.; Park, W. Y.; Hong, D.; Park, P. J.; Lee, E., 2015: Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nature genetics.*, **47**, 1242–1248.

Kahles, A.; Lehmann, K. V.; Toussaint, N. C.; Hüser, M.; Stark, S. G.; Sachsenberg, T.; Stegle, O.; Kohlbacher, O.; Sander, C.; Cancer Genome Atlas Research Network; Rätsch, G., 2018: Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer cell.*, **34**, 211–224.e6.

Karni, R.; Stanchina, E. de; Lowe, S. W.; Sinha, R.; Mu, D.; Krainer, A. R., 2007: The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nature Structural &#38; Molecular Biology.*, **14**, 185–193.

Katz, Y.; Wang, E. T.; Airoldi, E. M.; Burge, C. B., 2010: Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods.*, **7**, 1009–1015.

Ke, S.; Anquetil, V.; Zamalloa, J. R.; Maity, A.; Yang, A.; Arias, M. A.; Kalachikov, S.; Russo, J. J.; Ju, J.; Chasin, L. A., 2018: Saturation mutagenesis reveals manifold determinants of exon definition. *Genome research.*, **28**, 11–24.

Kolasinska-Zwierz, P.; Down, T.; Latorre, I.; Liu, T.; Liu, X. S.; Ahringer, J., 2009: Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature genetics.*,

41, 376–381.

Kornblihtt, A. R.; Schor, I. E.; Alló, M.; Dujardin, G.; Petrillo, E.; Muñoz, M. J., 2013: Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature Reviews Molecular Cell Biology.*, **14**, 153–165.

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**, 909–915.

Kwok, C. K.; Marsico, G.; Sahakyan, A. B.; Chambers, V. S.; Balasubramanian, S., 2016: rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nature Methods.*, **13**, 841–844.

Laguerre, A.; Hukezalie, K.; Winckler, P.; Katranji, F.; Chanteloup, G.; Pirrotta, M.; Perrier-Cornet, J. M.; Wong, J. M. Y.; Monchaud, D., 2015: Visualization of RNA-Quadruplexes in Live Cells. *Journal of the American Chemical Society.*, **137**, 8521–8525.

Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P. et al., 2001: Initial sequencing and analysis of the human genome. *Nature.*, **409**, 860–921.

Lee, Y.; Rio, D. C., 2015: Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual review of biochemistry.*, **84**, 291–323.

Lefave, C. V.; Squatrito, M.; Vorlova, S.; Rocco, G. L.; Brennan, C. W.; Holland, E. C.; Pan, Y. X.; Cartegni, L., 2011: Splicing factor hnRNPH drives an oncogenic splicing switch in gliomas. *The EMBO journal.*, **30**, 4084–4097.

Lek, M.; Karczewski, K. J.; Minikel, E. V.; Samocha, K. E.; Banks, E.; Fennell, T.; O'Donnell-Luria, A. H.; Ware, J. S.; Hill, A. J.; Cummings, B. B.; Tukiainen, T.; Birnbaum, D. P.; Kosmicki, J. A.; Duncan, L. E.; Estrada, K.; Zhao, F.; Zou, J.; Pierce-Hoffman, E.; Berghout, J. et al., 2016: Analysis of protein-coding genetic variation in 60,706 humans. *Nature.*, **536**, 285–291.

Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup, 2009: The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*., **25**, 2078–2079.

Li, Z.; Yao, H.; Guin, S.; Padhye, S. S.; Zhou, Y. Q.; Wang, M. H., 2010: Monoclonal antibody (mAb)-induced down-regulation of RON receptor tyrosine kinase diminishes tumorigenic activities of colon cancer cells. *International Journal of Oncology*., **37**, 473–482.

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**, 464–469.

Liu, X.; Jiang, Y.; Russell, J. E., 2010: A potential regulatory role for mRNA secondary structures within the prothrombin 3′UTR. *Thrombosis research*., **126**, 130–136.

Liu, X.; Ishizuka, T.; Bao, H. L.; Wada, K.; Takeda, Y.; Iida, K.; Nagasawa, K.; Yang, D.; Xu, Y., 2017: Structure-Dependent Binding of hnRNPA1 to Telomere RNA. *Journal of the American Chemical Society*., **139**, 7533–7539.

Lukong, K. E.; Chang, K. w.; Khandjian, E. W.; Richard, S., 2008: RNA-binding proteins in human genetic disease. *Trends in genetics : TIG*., **24**, 416–425.

Lykke-Andersen, S.; Jensen, T. H., 2015: Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nature Reviews Molecular Cell Biology*., **16**, 665–677.

Manning, K. S.; Cooper, T. A., 2017: The roles of RNA processing in translating genotype to phenotype. *Nature Reviews Molecular Cell Biology*., **18**, 102–114.

Marcel, V.; Tran, P. L. T.; Sagne, C.; Martel-Planche, G.; Vaslin, L.; Teulade-Fichou, M. P.; Hall, J.; Mergny, J. L.; Hainaut, P.; Van Dyck, E., 2011: G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms. *Carcinogenesis*., **32**, 271–278.

Marcucci, R.; Baralle, F. E.; Romano, M., 2007: Complex splicing control of the human Thrombopoietin gene by intronic G runs. *Nucleic acids research.*, **35**, 132–142.

Maris, C.; Dominguez, C.; Allain, F. H. T., 2005: The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *The FEBS Journal.*, **272**, 2118–2131.

Martinez-Contreras, R.; Fisette, J. F.; Nasim, F. u. H.; Madden, R.; Cordeau, M.; Chabot, B., 2006: Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS biology.*, **4**, e21.

Martinez-Contreras, R.; Cloutier, P.; Shkreta, L.; Fisette, J. F.; Revil, T.; Chabot, B., 2007: hnRNP proteins and splicing control. *Advances in experimental medicine and biology.*, **623**, 123–147.

Masuda, A.; Shen, X. M.; Ito, M.; Matsuura, T.; Engel, A. G.; Ohno, K., 2008: hnRNP H enhances skipping of a nonfunctional exon P3A in CHRNA1 and a mutation disrupting its binding causes congenital myasthenic syndrome. *Human molecular genetics.*, **17**, 4022–4035.

Mauger, D. M.; Lin, C.; Garcia-Blanco, M. A., 2008: hnRNP H and hnRNP F Complex with Fox2 To Silence Fibroblast Growth Factor Receptor 2 Exon IIIc. *Molecular and Cellular Biology.*, **28**, 5403–5419.

Mayer, S.; Hirschfeld, M.; Jaeger, M.; Pies, S.; Iborra, S.; Erbes, T.; Stickeler, E., 2015: RON alternative splicing regulation in primary ovarian cancer. *Oncology reports.*, **34**, 423–430.

Meyts, P. de; Roth, J.; Neville, D. M.; Gavin, J. R.; Lesniak, M. A., 1973: Insulin interactions with its receptors: experimental evidence for negative cooperativity. *Biochemical and biophysical research communications.*, **55**, 154–161.

Mittal, N.; Roy, N.; Babu, M. M.; Janga, S. C., 2009: Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America.*, **106**, 20300–20305.

Mo, Z.; Du, P.; Wang, G.; Wang, Y., 2017: The Multi-Purpose Tool of Tumor Immunotherapy: Gene-Engineered T Cells. *Journal of Cancer*., **8**, 1690–1703.

Moles-Fernández, A.; Duran-Lozano, L.; Montalban, G.; Bonache, S.; López-Perolio, I.; Menéndez, M.; Santamariña, M.; Behar, R.; Blanco, A.; Carrasco, E.; López-Fernández, A.; Stjepanovic, N.; Balmaña, J.; Capellá, G.; Pineda, M.; Vega, A.; Lázaro, C.; Hoya, M. de la; Diez, O. et al., 2018: Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations? *Frontiers in genetics*., **9**, 366.

Mootha, V. K.; Lindgren, C. M.; Eriksson, K. F.; Subramanian, A.; Sihag, S.; Lehar, J.; Puigserver, P.; Carlsson, E.; Ridderstråle, M.; Laurila, E.; Houstis, N.; Daly, M. J.; Patterson, N.; Mesirov, J. P.; Golub, T. R.; Tamayo, P.; Spiegelman, B.; Lander, E. S.; Hirschhorn, J. N. et al., 2003: PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*., **34**, 267–273.

Mount, S. M., 1982: A catalogue of splice junction sequences. *Nucleic acids research*., **10**, 459–472.

Mueller, W. F.; Larsen, L. S. Z.; Garibaldi, A.; Hatfield, G. W.; Hertel, K. J., 2015: The Silent Sway of Splicing by Synonymous Substitutions. *The Journal of biological chemistry*., **290**, 27700–27711.

Muraoka, R. S.; Sun, W. Y.; Colbert, M. C.; Waltz, S. E.; Witte, D. P.; Degen, J. L.; Friezner Degen, S. J., 1999: The Ron/STK receptor tyrosine kinase is essential for peri-implantation development in the mouse. *The Journal of clinical investigation*., **103**, 1277–1285.

Musunuru, K., 2003: Cell-specific RNA-binding proteins in human disease. *Trends in cardiovascular medicine*., **13**, 188–195.

Nasrin, F.; Rahman, M. A.; Masuda, A.; Ohe, K.; Takeda, J. i.; Ohno, K., 2014: HnRNP C, YB-1 and hnRNP L coordinately enhance skipping of human <i>MUSK</i> exon 10 to generate a Wnt-insensitive MuSK isoform. *Scientific Reports*., **4**, 6841.

Nazim, M.; Masuda, A.; Rahman, M. A.; Nasrin, F.; Takeda, J. i.; Ohe, K.; Ohkawara, B.; Ito, M.; Ohno, K., 2017: Competitive regulation of alternative splicing and alternative polyadenylation by hnRNP H and CstF64 determines acetylcholinesterase isoforms. *Nucleic acids research.*, **45**, 1455–1468.

Neelamraju, Y.; Gonzalez-Perez, A.; Bhat-Nakshatri, P.; Nakshatri, H.; Janga, S. C., 2018: Mutational landscape of RNA-binding proteins in human cancers. *RNA biology.*, **15**, 115–129.

Newman, M.; Sfaxi, R.; Saha, A.; Monchaud, D.; Teulade-Fichou, M. P.; Vagner, S., 2017: The G-Quadruplex-Specific RNA Helicase DHX36 Regulates p53 Pre-mRNA 3'-End Processing Following UV-Induced DNA Damage. *Journal of molecular biology.*, **429**, 3121–3131.

Nicholson, P.; Mühlemann, O., 2010: Cutting the nonsense: the degradation of PTC-containing mRNAs. *Biochemical Society transactions.*, **38**, 1615–1620.

Oberstrass, F. C.; Auweter, S. D.; Erat, M.; Hargous, Y.; Henning, A.; Wenter, P.; Reymond, L.; Amir-Ahmady, B.; Pitsch, S.; Black, D. L.; Allain, F. H. T., 2005: Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science (New York, N.Y.).*, **309**, 2054–2057.

Okunola, H. L.; Krainer, A. R., 2009: Cooperative-binding and splicing-repressive properties of hnRNP A1. *Molecular and Cellular Biology.*, **29**, 5620–5631.

Oltean, S.; Bates, D. O., 2014: Hallmarks of alternative splicing in cancer. *Oncogene.*, **33**, 5311–5318.

O'Toole, J. M.; Rabenau, K. E.; Burns, K.; Lu, D.; Mangalampalli, V.; Balderes, P.; Covino, N.; Bassi, R.; Prewett, M.; Gottfredsen, K. J.; Thobe, M. N.; Cheng, Y.; Li, Y.; Hicklin, D. J.; Zhu, Z.; Waltz, S. E.; Hayman, M. J.; Ludwig, D. L.; Pereira, D. S., 2006: Therapeutic Implications of a Human Neutralizing Antibody to the Macrophage-Stimulating Protein Receptor Tyrosine Kinase (RON), a c-MET Family Member. *Cancer Research.*, **66**, 9162–9170.

Pagani, F.; Raponi, M.; Baralle, F. E., 2005: Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proceedings of the National Academy of Sciences of the United States of America.*, **102**, 6368–6372.

Pan, Q.; Shai, O.; Lee, L. J.; Frey, B. J.; Blencowe, B. J., 2008: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics.*, **40**, 1413–1415.

Papasaikas, P.; Tejedor, J. R.; Vigevani, L.; Valcárcel, J., 2015: Functional splicing network reveals extensive regulatory potential of the core spliceosomal machinery. *Molecular cell.*, **57**, 7–22.

Paradis, C.; Cloutier, P.; Shkreta, L.; Toutant, J.; Klarskov, K.; Chabot, B., 2007: hnRNP I/PTB can antagonize the splicing repressor activity of SRp30c. *RNA.*, **13**, 1287–1300.

Parmley, J. L.; Chamary, J. V.; Hurst, L. D., 2006: Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Molecular biology and evolution.*, **23**, 301–309.

Pascual, M.; Vicente, M.; Monferrer, L.; Artero, R., 2006: The Muscleblind family of proteins: an emerging class of regulators of developmentally programmed alternative splicing. *Differentiation; research in biological diversity.*, **74**, 65–80.

Plaschka, C.; Newman, A. J.; Nagai, K., 2019: Structural Basis of Nuclear pre-mRNA Splicing: Lessons from Yeast. *Cold Spring Harbor Perspectives in Biology.*, **11**, a032391.

Pollard, K. S.; Hubisz, M. J.; Rosenbloom, K. R.; Siepel, A., 2010: Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research.*, **20**, 110–121.

Qian, W.; Liu, F., 2014: Regulation of alternative splicing of tau exon 10. *Neuroscience bulletin.*, **30**, 367–377.

Quantin, B.; Schuhbaur, B.; Gesnel, M. C.; Doll'e, P.; Breathnach, R., 1995: Restricted expression of the ron gene encoding the macrophage stimulating protein receptor during mouse development. *Developmental dynamics : an official publication of the American Association of Anatomists.*, **204**, 383–390.

Rahman, M. A.; Azuma, Y.; Nasrin, F.; Takeda, J. i.; Nazim, M.; Bin Ahsan, K.; Masuda, A.; Engel, A. G.; Ohno, K., 2015: SRSF1 and hnRNP H antagonistically regulate splicing of COLQ exon 16 in a congenital myasthenic syndrome. *Scientific Reports.*, **5**, 13208.

Rauch, J.; O'Neill, E.; Mack, B.; Matthias, C.; Munz, M.; Kolch, W.; Gires, O., 2010: Heterogeneous nuclear ribonucleoprotein H blocks MST2-mediated apoptosis in cancer cells by regulating A-Raf transcription. *Cancer Research.*, **70**, 1679–1688.

Robinson, T. J.; Freedman, J. A.; Al Abo, M.; Deveaux, A. E.; LaCroix, B.; Patierno, B. M.; George, D. J.; Patierno, S. R., 2019: Alternative RNA Splicing as a Potential Major Source of Untapped Molecular Targets in Precision Oncology and Cancer Disparities. *Clinical Cancer Research.*, **25**, 2963–2968.

Romano, M.; Marcucci, R.; Buratti, E.; Ayala, Y. M.; Sebastio, G.; Baralle, F. E., 2002: Regulation of 3′ splice site selection in the 844ins68 polymorphism of the cystathionine Beta -synthase gene. *The Journal of biological chemistry.*, **277**, 43821–43829.

Ronsin, C.; Muscatelli, F.; Mattei, M. G.; Breathnach, R., 1993: A novel putative receptor protein tyrosine kinase of the met family. *Oncogene.*, **8**, 1195–1202.

Roque, R. S.; Caldwell, R. B.; Behzadian, M. A., 1992: Cultured Müller cells have high levels of epidermal growth factor receptors. *Investigative ophthalmology & visual science.*, **33**, 2587–2595.

Rosenberg, A. B.; Patwardhan, R. P.; Shendure, J.; Seelig, G., 2015: Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell.*, **163**, 698–711.

Rothrock, C. R.; House, A. E.; Lynch, K. W., 2005: HnRNP L represses exon splicing via a regulated exonic splicing silencer. *The EMBO journal.*, **24**, 2792–2802.

Savisaar, R.; Hurst, L. D., 2017a: Estimating the prevalence of functional exonic splice regulatory information. *Human Genetics.*, **136**, 1059–1078.

Savisaar, R.; Hurst, L. D., 2017b: Both Maintenance and Avoidance of RNA-Binding Protein Interactions Constrain Coding Sequence Evolution. *Molecular biology and evolution.*, **34**, 1110–1126.

Savisaar, R.; Hurst, L. D., 2018: Exonic splice regulation imposes strong selection at synonymous sites. *Genome research.*, **28**, 1442–1454.

Schirmer, M.; Ijaz, U. Z.; D'Amore, R.; Hall, N.; Sloan, W. T.; Quince, C., 2015: Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic acids research.*, **43**, e37–e37.

Schlessinger, J., 2000: Cell signaling by receptor tyrosine kinases. *Cell.*, **103**, 211–225.

Schlessinger, J., 2014: Receptor tyrosine kinases: legacy of the first two decades. *Cold Spring Harbor Perspectives in Biology.*, **6**, a008912–a008912.

Sebestyén, E.; Singh, B.; Miñana, B.; Pagès, A.; Mateo, F.; Pujana, M. A.; Valcárcel, J.; Eyras, E., 2016: Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome research.*, **26**, 732–744.

Sedlazeck, F. J.; Rescheneder, P.; Haeseler, A. von, 2013: NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics (Oxford, England).*, **29**, 2790–2791.

Seiler, M.; Peng, S.; Agrawal, A. A.; Palacino, J.; Teng, T.; Zhu, P.; Smith, P. G.; Buonamici, S.; Yu, L.; Caesar-Johnson, S. J.; Demchok, J. A.; Felau, I.; Kasapi, M.; Ferguson, M. L.; Hutter, C. M.; Sofia, H. J.; Tarnuzzer, R.; Wang, Z.; Yang, L. et al., 2018: Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. *Cell reports.*, **23**, 282–296.e4.

Shabalina, S. A.; Spiridonov, N. A.; Kashina, A., 2013: Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic acids research.*, **41**, 2073–2094.

Shen, H.; Kan, J. L. C.; Green, M. R., 2004: Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Molecular cell.*, **13**, 367–376.

Shen, M.; Mattox, W., 2012: Activation and repression functions of an SR splicing regulator depend on exonic versus intronic-binding position. *Nucleic acids research.*, **40**, 428–437.

Shi, Y., 2017: Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nature Reviews Molecular Cell Biology.*, **18**, 655–670.

Shi, Y. Z.; Jin, L.; Wang, F. H.; Zhu, X. L.; Tan, Z. J., 2015: Predicting 3D Structure, Flexibility, and Stability of RNA Hairpins in Monovalent and Divalent Ion Solutions. *Biophysical Journal.*, **109**, 2654–2665.

Shi, Y. Z.; Jin, L.; Feng, C. J.; Tan, Y. L.; Tan, Z. J., 2018: Predicting 3D structure and stability of RNA pseudoknots in monovalent and divalent ion solutions. *PLoS computational biology.*, **14**, e1006222.

Soemedi, R.; Cygan, K. J.; Rhine, C. L.; Wang, J.; Bulacan, C.; Yang, J.; Bayrak-Toydemir, P.; McDonald, J.; Fairbrother, W. G., 2017: Pathogenic variants that alter protein code often disrupt splicing. *Nature genetics.*, **49**, 848–855.

Soukarieh, O.; Gaildrat, P.; Hamieh, M.; Drouet, A.; Baert-Desurmont, S.; Frébourg, T.; Tosi, M.; Martins, A., 2016: Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico Tools. *PLoS genetics.*, **12**, e1005756.

Stark, M.; Bram, E. E.; Akerman, M.; Mandel-Gutfreund, Y.; Assaraf, Y. G., 2011: Heterogeneous nuclear ribonucleoprotein H1/H2-dependent unsplicing of thymidine phosphorylase results in anticancer drug resistance. *The Journal of biological chemistry.*, **286**, 3741–3754.

Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P., 2005: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America.*, **102**, 15545–15550.

Sun, S.; Zhang, Z.; Fregoso, O.; Krainer, A. R., 2012: Mechanisms of activation and repression by the alternative splicing factors RBFOX1/2. *RNA.*, **18**, 274–283.

Supek, F.; Miñana, B.; Valcárcel, J.; Gabaldón, T.; Lehner, B., 2014: Synonymous mutations frequently act as driver mutations in human cancers. *Cell.*, **156**, 1324–1335.

Sutandy, F. R.; Hildebrandt, A.; König, J., 2016: Profiling the Binding Sites of RNA-Binding Proteins with Nucleotide Resolution Using iCLIP. In: *Post-transcriptional gene regulation*. Institute of Molecular Biology gGmbH, Ackermannweg 4, 55128, Mainz, Germany. Humana Press, New York, NY, New York, NY, pp. 175–195.

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**, 699–713.

Sveen, A.; Kilpinen, S.; Ruusulehto, A.; Lothe, R. A.; Skotheim, R. I., 2016: Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene.*, **35**, 2413–2427.

Taylor, K.; Sznajder, L. J.; Cywoniuk, P.; Thomas, J. D.; Swanson, M. S.; Sobczak, K., 2018: MBNL splicing activity depends on RNA binding site structural context. *Nucleic acids research.*, **46**, 9119–9133.

Tejedor, J. R.; Papasaikas, P.; Valcárcel, J., 2015: Genome-wide identification of Fas/CD95 alternative splicing regulators reveals links with iron homeostasis. *Molecular cell.*, **57**, 23–38.

Tomczak, K.; Czerwińska, P.; Wiznerowicz, M., 2015: The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology (Poznan, Poland).*, **19**, A68–

77.

Treutlein, B.; Gokce, O.; Quake, S. R.; Südhof, T. C., 2014: Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America.*, **111**, E1291–9.

Tuladhar, R.; Yeu, Y.; Piazza, J. T.; Tan, Z.; Clemenceau, J. R.; Wu, X.; Barrett, Q.; Herbert, J.; Mathews, D. H.; Kim, J.; Hwang, T. H.; Lum, L., 2019: CRISPR/Cas9-based mutagenesis frequently provokes on-target mRNA misregulation. *bioRxiv.*, **136**, 583138.

Uren, P. J.; Bahrami-Samani, E.; Araujo, P. R. de; Vogel, C.; Qiao, M.; Burns, S. C.; Smith, A. D.; Penalva, L. O. F., 2016: High-throughput analyses of hnRNP H1 dissects its multifunctional aspect. *RNA biology.*, **13**, 400–411.

Valverde, R.; Edwards, L.; Regan, L., 2008: Structure and function of KH domains. *The FEBS Journal.*, **275**, 2712–2726.

Van der Auwera, G. A.; Carneiro, M. O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; Banks, E.; Garimella, K. V.; Altshuler, D.; Gabriel, S.; DePristo, M. A., 2013: From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics.*, **43**, 11.10.1–33.

Van Dusen, C. M.; Yee, L.; McNally, L. M.; McNally, M. T., 2010: A Glycine-Rich Domain of hnRNP H/F Promotes Nucleocytoplasmic Shuttling and Nuclear Import through an Interaction with Transportin 1. *Molecular and Cellular Biology.*, **30**, 2552–2562.

Vivian, J.; Rao, A. A.; Nothaft, F. A.; Ketchum, C.; Armstrong, J.; Novak, A.; Pfeil, J.; Narkizian, J.; Deran, A. D.; Musselman-Brown, A.; Schmidt, H.; Amstutz, P.; Craft, B.; Goldman, M.; Rosenbloom, K.; Cline, M.; O'Connor, B.; Hanna, M.; Birger, C. et al., 2017: Toil enables reproducible, open source, big biomedical data analyses. *Nature biotechnology.*, **35**, 314–316.

Wachter, A.; Rühl, C.; Stauffer, E., 2012: The Role of Polypyrimidine Tract-Binding Proteins and Other hnRNP Proteins in Plant Splicing Regulation. *Frontiers in plant science.*, **3**, 81.

Wahl, M. C.; Will, C. L.; Lührmann, R., 2009: The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell.*, **136**, 701–718.

Wang, D.; Shen, Q.; Chen, Y. Q.; Wang, M. H., 2004: Collaborative activities of macrophage-stimulating protein and transforming growth factor-beta1 in induction of epithelial to mesenchymal transition: roles of the RON receptor tyrosine kinase. *Oncogene.*, **23**, 1668–1680.

Wang, M. H.; Yoshimura, T.; Skeel, A.; Leonard, E. J., 1994: Proteolytic conversion of single chain precursor macrophage-stimulating protein to a biologically active heterodimer by contact enzymes of the coagulation cascade. *The Journal of biological chemistry.*, **269**, 3436–3440.

Wang, M. H.; Dlugosz, A. A.; Sun, Y.; Suda, T.; Skeel, A.; Leonard, E. J., 1996: Macrophage-stimulating protein induces proliferation and migration of murine keratinocytes. *Experimental cell research.*, **226**, 39–46.

Wang, Y.; Liu, J.; Huang, B. O.; Xu, Y. M.; Li, J.; Huang, L. F.; Lin, J.; Zhang, J.; Min, Q. H.; Yang, W. M.; Wang, X. Z., 2015: Mechanism of alternative splicing and its regulation. *Biomedical reports.*, **3**, 152–158.

Wang, Z.; Burge, C. B., 2008: Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA.*, **14**, 802–813.

Wang, Z. L.; Li, B.; Luo, Y. X.; Lin, Q.; Liu, S. R.; Zhang, X. Q.; Zhou, H.; Yang, J. H.; Qu, L. H., 2018: Comprehensive Genomic Characterization of RNA-Binding Proteins across Human Cancers. *Cell reports.*, **22**, 286–298.

Warf, M. B.; Berglund, J. A., 2010: Role of RNA structure in regulating pre-mRNA splicing. *Trends in biochemical sciences.*, **35**, 169–178.

Weldon, C.; Dacanay, J. G.; Gokhale, V.; Boddupally, P. V. L.; Behm-Ansmant, I.; Burley, G. A.; Branlant, C.; Hurley, L. H.; Dominguez, C.; Eperon, I. C., 2018: Specific G-quadruplex ligands modulate the alternative splicing of Bcl-X. *Nucleic acids research.*, **46**, 886–896.

Whitlock, M. C., 2005: Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of evolutionary biology.*, **18**, 1368–1373.

Will, C. L.; Lührmann, R., 2011: Spliceosome Structure and Function. *Cold Spring Harbor Perspectives in Biology.*, **3**, a003707–a003707.

Williams, R.; Peisajovich, S. G.; Miller, O. J.; Magdassi, S.; Tawfik, D. S.; Griffiths, A. D., 2006: Amplification of complex gene libraries by emulsion PCR. *Nature Methods.*, **3**, 545–550.

Wu, J. Y.; Maniatis, T., 1993: Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell.*, **75**, 1061–1070.

Wu, Y. X.; Kwon, Y. J., 2016: Aptamers: The "evolution" of SELEX. *Methods (San Diego, Calif.).*, **106**, 21–28.

Xiao, X.; Wang, Z.; Jang, M.; Nutiu, R.; Wang, E. T.; Burge, C. B., 2009: Splice site strength-dependent activity and genetic buffering by poly-G runs. *Nature Structural & Molecular Biology.*, **16**, 1094–1100.

Xing, Y.; Lee, C. J., 2005: Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS genetics.*, **1**, e34.

Xing, Y.; Lee, C., 2006: Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. *Nature reviews. Genetics.*, **7**, 499–509.

Xiong, H. Y.; Alipanahi, B.; Lee, L. J.; Bretschneider, H.; Merico, D.; Yuen, R. K. C.; Hua, Y.; Gueroussov, S.; Najafabadi, H. S.; Hughes, T. R.; Morris, Q.; Barash, Y.; Krainer, A. R.; Jojic, N.; Scherer, S. W.; Blencowe, B. J.; Frey, B. J., 2015: The human splicing code reveals new insights into the genetic determinants of disease. *Science (New York, N.Y.).*, **347**, 1254806–1254806.

Xu, X.; Chen, S. J., 2016: A Method to Predict the Structure and Stability of RNA/RNA Complexes. *Methods in molecular biology (Clifton, N.J.).*, **1490**, 63–72.

Yang, X.; Coulombe-Huntington, J.; Kang, S.; Sheynkman, G. M.; Hao, T.; Richardson, A.; Sun, S.; Yang, F.; Shen, Y. A.; Murray, R. R.; Spirohn, K.; Begg, B. E.; Duran-Frigola, M.; MacWilliams, A.; Pevzner, S. J.; Zhong, Q.; Trigg, S. A.; Tam, S.; Ghamsari, L. et al., 2016: Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell.*, **164**, 805–817.

Yao, H. P.; Luo, Y. L.; Lei, F.; Cheng, L. F.; Lu, Y.; Li, W.; Wang, M. H., 2006: Agonistic monoclonal antibodies potentiate tumorigenic and invasive activities of splicing variant of the RON receptor tyrosine kinase. *Cancer Biology & Therapy.*, **5**, 1179–1186.

Yao, H. P.; Zhou, Y. Q.; Zhang, R.; Wang, M. H., 2013: MSP-RON signalling in cancer: pathogenesis and therapeutic potential. *Nature reviews. Cancer.*, **13**, 466–481.

Yao, H. P.; Feng, L.; Suthe, S. R.; Chen, L. H.; Weng, T. H.; Hu, C. Y.; Jun, E. S.; Wu, Z. G.; Wang, W. L.; Kim, S. C.; Tong, X. M.; Wang, M. H., 2019: Therapeutic efficacy, pharmacokinetic profiles, and toxicological activities of humanized antibody-drug conjugate Zt/g4-MMAE targeting RON receptor tyrosine kinase for cancer therapy. *Journal for ImmunoTherapy of Cancer.*, **7**, 1–16.

Yeo, G.; Burge, C. B., 2004: Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *home.liebertpub.com.*, **11**, 377–394.

Yi, H.; Xue, L.; Guo, M. X.; Ma, J.; Zeng, Y.; Wang, W.; Cai, J. Y.; Hu, H. M.; Shu, H. B.; Shi, Y. B.; Li, W. X., 2010: Gene expression atlas for human embryogenesis. *Faseb Journal.*, **24**, 3341–3350.

Zarnack, K.; König, J.; Tajnik, M.; Martincorena, I.; Eustermann, S.; Stévant, I.; Reyes, A.; Anders, S.; Luscombe, N. M.; Ule, J., 2013: Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements. *Cell.*, **152**, 453–466.

Zhang, J.; Harvey, S. E.; Cheng, C., 2019: A high-throughput screen identifies small molecule modulators of alternative splicing by targeting RNA G-quadruplexes. *Nucleic acids research.*, **47**, 3667–3679.

Zhang, W.; Flemington, E. K.; Zhang, K., 2018: Driver gene mutations based clustering of tumors: methods and applications. *Bioinformatics (Oxford, England).*, **34**, i404–i411.

Zhang, X.; Chasin, L. A., 2004: Computational definition of sequence motifs governing constitutive exon splicing. *Genes & development.*, **18**, 1241–1250.

Zhou, Y. Q.; He, C.; Chen, Y. Q.; Wang, D.; Wang, M. H., 2003: Altered expression of the RON receptor tyrosine kinase in primary human colorectal adenocarcinomas: generation of different splicing RON variants and their oncogenic potential. *Oncogene.*, **22**, 186–197.

Zhu, J.; Mayeda, A.; Krainer, A. R., 2001: Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Molecular cell.*, **8**, 1351–1361.

Addendum

Preface (extended with figure contributions)

The *RON* mutagenesis screening project constituted of three sections; experimental, bioinformatics and mathematical modelling and was supervised by Dr. Julian König (IMB, Mainz), Dr. Kathi Zarnack (BMLS, Frankfurt), Dr. Stefanie Ebersberger (IMB, Mainz) and Dr. Stefan Legewie (IMB, Mainz), respectively. The results of the project and part of the thesis have been published in the following article -

Braun, S.*; Enculescu, M.*; Setty, S. T.*; Cortés-López, M.; Almeida, B. P. de; Sutandy, F. X. R.; Schulz, L.; Busch, A.; Seiler, M.; Ebersberger, S.; Barbosa-Morais, N. L.; Legewie, S.; König, J.; Zarnack, K., 2018: Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. *Nature Communications.*, 9, 3315.

*These authors contributed equally: Simon Braun, Mihaela Enculescu, Samarth Thonta Setty. The author and figure contributions are as listed in the following paragraphs.

In my PhD study, I performed most of the bioinformatics analyses of the project. In this thesis, I have presented the topics that were exclusively analysed by myself (Figures 2.1, 3.6 C, 3.7, 3.8, 3.9 C, D & E, 3.13 (Modelling results from Mihaela) , 3.14, 3.15, 3.16, 3.17, 3.18 A (ΔAE inclusion (%) values) and B, 3.20, 3.23, 3.24, 3.27, 3.29, 3.30, 3.31, 3.32, 3.33, 3.34 dot plots, 3.34 B (RNA-seq values)), along with analyses performed by the other collaborators to present a complete and cohesive picture of the project for better comprehension. These collaborators and their corresponding work in brackets are as follows –

Simon Braun (method schematics - Figures 3.1 and 3.2 (with my design inputs); Experimental section - Figures 1.6, 1.7, 3.3, 3.4, 3.5, 3.6 A and B, 3.22 A and B, 3.34 B (RT-PCR values), 3.35, 3.36), F. X. Reymond Sutandy (iCLIP experiments - Figures 3.27 (figure by myself), 3.35), Mariela Cortés-López (RBP site annotation and analyses - Figures 3.25, 3.28, 3.34 heatmaps, 4.4), Laura Schulz (RT-PCR validation experiments - Figure 3.12 (RT-PCR values)), Dr. Markus Seiler (Annotation of RNA G-quadruplex sequence (G4 hunter scores calculation only) - Figure 3.18 A; Figure done by myself), Dr. Anke Busch (iCLIP and RNA-seq data processing and splice isoform quantification - Figure 3.9 A and B by Anke Busch; F - IF % quantification by Anke Busch; figure by myself), Bernardo P. de Almeida and Dr. Nuno L. Barbosa-Morais (TCGA and GTEx analyses - Figures 3.21, 3.26; iMM, Lisbon). The project was conceived by Dr. Julian König and experimental analyses were performed under his supervision. The bioinformatics analyses were performed under the supervision of Dr. Katharina Zarnack with additional supervision by Dr. Stefanie Ebersberger. Dr. Mihaela Enculescu (Figures 3.10, 3.11, 3.12 (RNA-seq - model results), 3.19) and Dr. Stefan Legewie designed and performed the mathematical modelling approach.

Erklärung

Ich erkläre hiermit, dass ich mich bisher keiner Doktorprüfung im mathematisch - naturwissenschaftlichem Bereich unterzogen habe.

Frankfurt am Main, den 14-07-2020

Samarth

Samarth Thonta Setty

Versicherung

Ich versichere hiermit, dass die vorgelegte Doktorarbeit über "Computational approaches to decipher splicing regulatory network of the *RON* proto-oncogene“ selbständig und ohne unzulässige fremde Hilfe verfasst, andere als die in ihr angegebene Literatur nicht benutzt und, dass ich alle ganz oder annährend übernommenen Textstellen, sowie verwendete Grafiken, Tabellen und Auswertungsprogramme gekennzeichnet habe. Außerdem versichere ich, dass die vorgelegte elektronische mit der schriftlichen Version der Doktorarbeit übereinstimmt.

Frankfurt am Main, den 14-07-2020

Samarth

Samarth Thonta Setty