Decoding a cancer-relevant splicing decision in the *RON* proto-oncogene using high-throughput mutagenesis

Dissertation zur Erlangung des Grades
"Doktor der Naturwissenschaften"
am Fachbereich Biologie
der Johannes Gutenberg-Universität in Mainz

Simon Braun
geb. am 24.02.1988 in Offenbach am Main

Mainz, 2018

Tag der mündlichen Prüfung: 31.01.2019

Zusammenfassung

Alternatives Spleißen stellt einen wichtigen Mechanismus der Genregulation dar, der die Kodierungskapazität des menschlichen Genoms drastisch erhöht. Aus Fehlern dieses Prozesses können verschiedene Krankheiten wie Krebs entstehen. Ermöglicht wird die Kontrolle des alternativen Spleißens durch die Rekrutierung transaktiver RNA bindender Proteine an *cis*-regulatorische Elemente. Obwohl zahlreiche dieser Interaktionen bereits im Detail beschrieben sind, ist der regulatorische Code, der der Spleiß-Entscheidung zu Grunde liegt, vielfach nicht verstanden. Hier wurde ein Hochdurchsatz-Zufallsmutagenese-Screening entwickelt, das die für einen Krebs-relevanten Spleißvorgang im Protoonkogen *RON* entscheidenden Mutation und RNA bindenden Proteine umfassend charakterisiert. Insgesamt wurden mehr als 700 Mutationen identifiziert, die die Regulation des alternativen Exons signifikant beeinflussen. Es wurde darüber hinaus gezeigt, dass die mit Hilfe des Screenings quantifizierten Mutationseffekte mit alternativem Spleißen von *RON* in Krebspatienten korrelieren. Zudem wurden zahlreiche transaktive Regulatoren enthüllt und das heterogene nukleäre Ribonucleoprotein H (HNRNPH) als extensiver Regulator des *RON* Spleißens in Zelllinien und Krebs identifiziert. Mittels iCLIP und Synergieanalyse zwischen Mutationen und HNRNPH Knock-down wurden die funktionell wichtigsten HNRNPH-Bindestellen in *RON* lokalisiert. Schließlich wurde aufgezeigt, dass kooperative HNRNPH-Bindung ein Umschalten des Spleißens von *RON* Exon 11 reguliert. Diese Ergebnisse demonstrieren, dass Spleiß-Regulation durch nahezu jedes Nukleotid innerhalb des alternativen Exons vermittelt wird. Darüber hinaus weist die große Zahl an transaktiven Regulatoren darauf hin, dass ein komplexes Spleiß-Netzwerk an der Kontrolle des *RON* Spleißens beteiligt ist.

Summary

Alternative splicing is an important regulatory mechanism of gene expression that dramatically increases the coding capacity of the human genome. Various diseases including cancer can arise from defects in this process. Control of alternative splicing is realized by *cis*-regulatory elements, which recruit *trans*-acting RNA-binding proteins. Although several of those interactions are already described in detail, the regulatory code that underlies specific splicing decisions is frequently not understood. Here, a high-throughput random mutagenesis screen was established that comprehensively characterized determinants of a cancer-relevant splicing event in the proto-oncogene *RON*. In total, more than 700 mutations were found to significantly affect the regulation of the alternative exon. It was moreover shown that mutation effects quantified from the screening correlate with *RON* alternative splicing in cancer patients. In addition, numerous *trans*-acting regulators were revealed and the heterogeneous nuclear ribonucleoprotein H (HNRNPH) was identified as an extensive regulator of *RON* splicing in cell lines and in cancer. iCLIP and analysis of synergy between mutations and HNRNPH knockdown allowed localization of the functionally most relevant HNRNPH binding sites across *RON*. Finally, cooperative HNRNPH binding was shown to mediate a splicing switch of *RON* exon 11. These results demonstrate that splicing regulation is conferred by nearly every nucleotide within the alternative exon. Furthermore, the large number of *trans*-acting regulators point to a complex splicing regulatory network involved in the control of *RON* splicing.

Preface

I declare that I have written this thesis independently and all cited resources have been listed in the references. Partial results of the presented work have been published in (Braun et al., 2018). The content of this thesis is based on a research collaboration between the laboratories of [redacted] and [redacted] (IMB, Mainz), [redacted] (BMLS, Frankfurt), and [redacted] (iMM, Lisbon). I was involved in design and evaluation of all experiments and analyses. All experimental work was performed by myself, with the following exceptions: the iCLIP experiments were carried out by [redacted] (IMB, Mainz) and parts of the validation experiments shown in Figure 37B were contributed by [redacted] (IMB, Mainz). [redacted] (BMLS, Frankfurt) performed most bioinformatics analyses. [redacted] (IMB, Mainz) carried out the RBP binding site predictions, as well as analyses of mutation effects and synergistic interactions. [redacted] (IMB, Mainz) performed iCLIP- and RNA-sequencing data processing, and splice isoform quantifications. [redacted] and [redacted] (both IMB, Mainz) designed the mathematical modelling and performed corresponding analyses. [redacted] and [redacted] (both iMM, Lisbon) analyzed the TCGA- and GTEx data. The bioinformatics analyses were supervised by [redacted] (IMB, Mainz) and [redacted] (BMLS, Frankfurt) and the experimental work was supervised by [redacted] (IMB, Mainz).

Table of Contents

List of Abbreviations

A	Adenine
AE	Alternative exon
BMLS	Buchmann Institute for Molecular Life Sciences
BSA	Bovine serum albumin
C	Cytosine
cDNA	Complementary DNA
CI	Confidence interval
DNA	Deoxyribonucleic acid
DNA-seq	DNA-sequencing
dNTP	Deoxy nucleosid triphosphate
E. coli	*Escherichia coli*
EMT	Epithelial to mesenchymal transition
FDR	False discovery rate
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
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
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
RNA-seq	RNA-sequencing
RON	Recepteur d'Origine Nantais
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

Introduction

1.1 Pre-mRNA splicing

Eukaryotic genes are transcribed as precursor-messenger RNAs (pre-mRNAs) that require several processing steps, before they can serve as a template for protein synthesis as a mature mRNA. Since eukaryotic genes are discontinuously organized, one of these steps, called splicing, involves the removal of introns from the pre-mRNA and subsequent fusion of exons to result with a continuous message.

Splicing catalysis is a stepwise process that is executed by the spliceosome, a highly complex and dynamic multi-megadalton ribonucleoprotein (RNP) particle composed of numerous small nuclear ribonucleoproteins (snRNPs) (Fica et al., 2013). The assembly of the spliceosome is initiated by the interaction of U1 snRNP with conserved sequence elements located at the 5’ end of the intron, the so-called 5’ splice site through base pairing with the 5′ end of the U1 small nuclear RNA (snRNA; **Figure 1**) (Roca et al., 2013). Next, splicing factor binding to the 3’ end of the intron (3' splice site) leads to early spliceosomal complex E formation: SF1 binds to the branch point, whereas the U2AF heterodimer binds to the polypyrimidine tract and the downstream 3’ splice site. This allows transition to the prespliceosome (complex A) by subsequent binding of U2 snRNP to the 3’ end of the intron. Interaction between U1 snRNP and U2 snRNP brings together both ends of the intron and therefore defines the intron for subsequent excision. Alternatively, interaction of an upstream U2 snRNP with a downstream U1 snRNP first marks exons within the pre-mRNA, before the introns are excised according to an exon definition model (Conti et al., 2013).

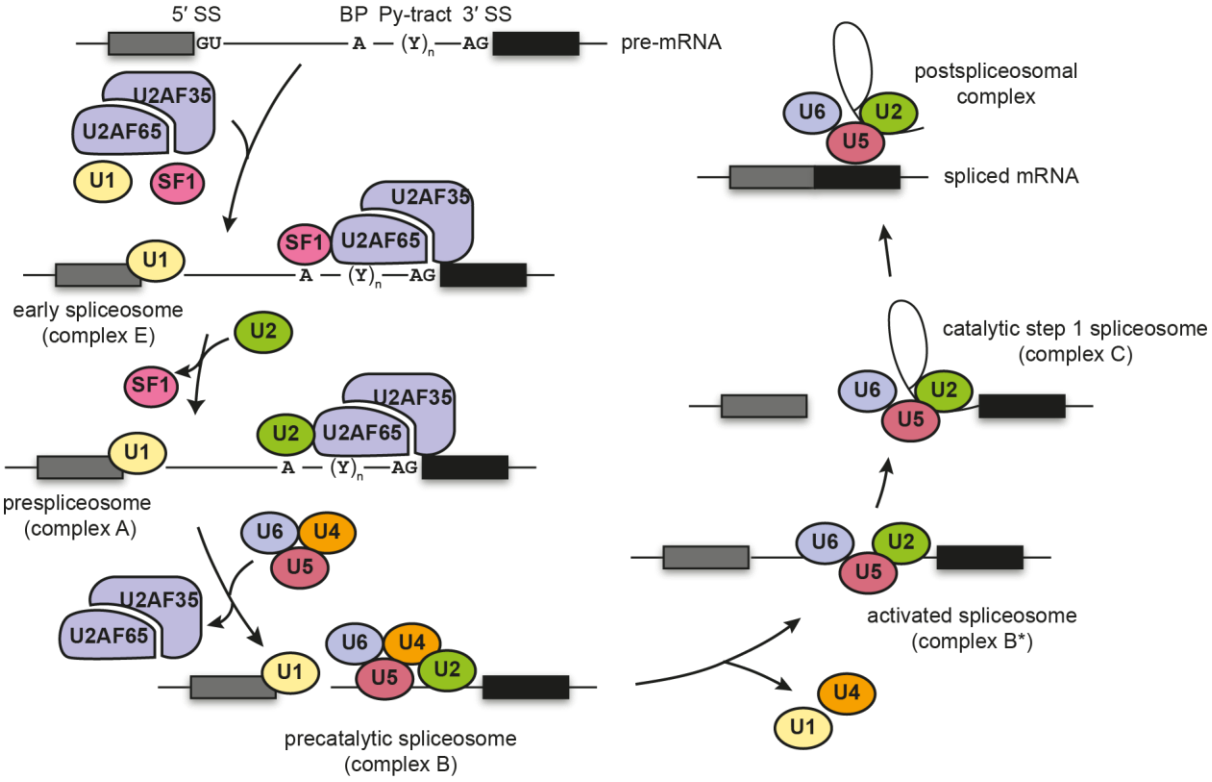


Figure 1: Pre-mRNA splicing by the spliceosome is a multi-step process. Core-splicing signals within introns of pre-mRNAs comprise the 5' splice site (5' SS), the branch point (BP), the polypyrimidine-tract (Py-tract), and the 3' splice site (3' SS). Early splicing factors U1 snRNP, SF1 and the U2AF 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 (not shown). Boxes represent exons while the connecting introns are depicted with lines. Adapted from (Wahl et al., 2009).

Splicing catalysis proceeds with precatalytic spliceosome formation (complex B) by recruitment of U4/U6.U5 tri-snRNP, which is followed by the release of U1 and U4 snRNPs to result with the activated spliceosome (complex B*). Now two consecutive transesterification steps take place: First, the 2'-OH of the conserved branch point adenine attacks the phosphate of the conserved guanine at the 5' splice site, cleaving between the exon and the guanine in the intron and thereby resulting in a free 3'-OH at the upstream exon and an intron lariat attached to the downstream exon. In the catalytic step 1 spliceosome (complex C) the second step of

catalysis is then executed by a nucleophilic attack of the 3'-OH of the upstream exon on the phosphate at the 5' end of the downstream exon to covalently join both exons and release the intron lariat (Wahl et al., 2009; Shi, 2017).

Together, five RNAs and more than 170 proteins associate with the metazoan spliceosome during the splicing reaction to catalyze a comparably simple chemical reaction (Will & Lührmann, 2011). The compositional complexity of the spliceosome can be attributed to the relevance of the splicing process. Furthermore, a multi-step splicing reaction involving many proteins allows control and modulation via a process known as alternative splicing (Johnson & Vilardell, 2012).

1.2 Alternative splicing

Combinatorial assembly of pre-mRNA into mature mRNA transcripts through a process called alternative splicing enables the production of protein isoforms with different functionalities. Thus, alternative splicing is a source of phenotypic diversity from a limited number of genes (Yang et al., 2016). Moreover, alternative splicing is tissue- and development-stage specific and hence important for cell and organ differentiation as well as tissue homeostasis (Baralle & Giudice, 2017). In fact, more than 90% of human genes are alternatively spliced (Yang et al., 2016) and extensive alternative splicing can generate more than 1,000 different isoforms from a single human gene under physiological conditions (Treutlein et al., 2014). With increasing evolutionary distance from human, the frequency of alternative splicing decreases (Barbosa-Morais et al., 2012). Consequently, alternative splicing is strongly associated with organism complexity (Chen et al., 2014).

Depending on the splice site choice, five different modes of alternative splicing can be distinguished (**Figure 2**): So-called cassette exons can be either included or excluded from the mature mRNA (exon inclusion/skipping). Alternatively, exactly one of two or more possible exons is included in the mRNA (mutually exclusive exons). Unspliced introns (intron retention; IR) and alternative splice site usage (alternative 3' and alternative 5' splice site) result in exons

of varying length (Kornblihtt et al., 2013). 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 & Smith, 2017).

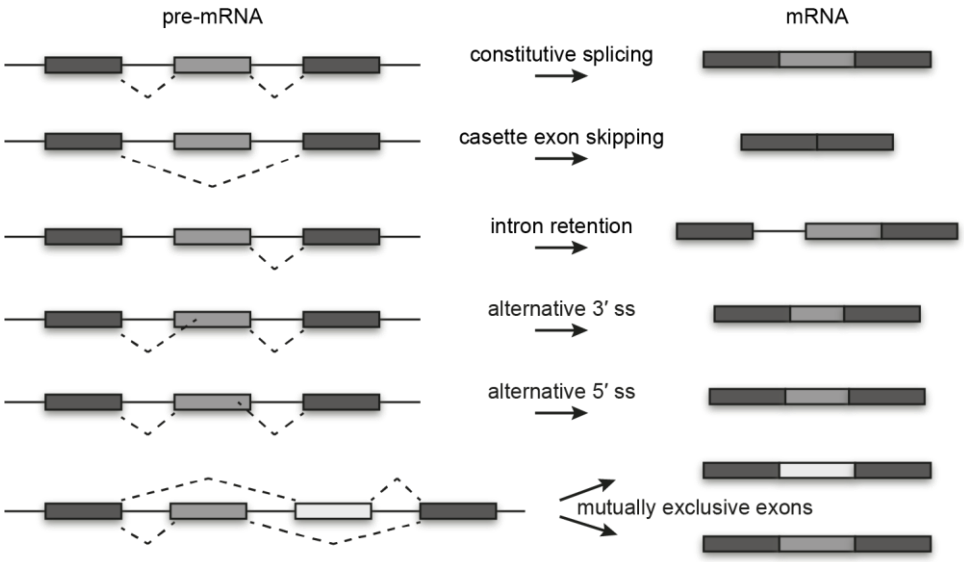


Figure 2: Different types of alternative splicing. Constitutive splicing results in exon inclusion, while various types of alternative splicing result in different mRNA isoforms. Introns are depicted with solid lines while exons are shown as boxes. Dashed lines indicate splice junctions between exons.

1.3 Regulation of alternative splicing

The first layer of alternative splicing regulation is constituted by the splice site strength, i.e. the more the sequences at the splice sites match a consensus sequence that is well recognized by the early spliceosome (Garg & Green, 2007). While strong consensus splice sites generally result in constitutive splicing, the splicing machinery is challenged by the presence of numerous additional alternative splice sites of weak or intermediate strength, that must be selected in a context-dependent manner (Baralle & Giudice, 2017). Moreover, metazoan consensus splice sites are rather poorly conserved in comparison to the splice sites of yeast,

which is thought to allow the extensive alternative splicing observed in higher organisms (Lee & Rio, 2015).

To ensure correct and context-dependent alternative splicing, the splicing machinery requires additional RNA binding proteins (RBPs), so called *trans*-acting factors. These RBPs target RNA sequence elements, so called *cis*-regulatory elements, embedded in the pre-mRNA to guide the spliceosome activity (Fu & Ares, 2014). The *cis*-regulatory elements are classified as exonic or intronic splicing enhancers (ESEs or ISEs) and silencers (ESSs or ISSs), depending on whether they enhance or repress exon inclusion (Kornblihtt et al., 2013). In fact, many splicing events are controlled by combinatorial interplay of multiple *cis*-regulatory elements (Nasrin et al., 2014; Qian & Liu, 2014). The group of *trans*-acting splicing factors includes serine–arginine repeat (SR) splicing factors, heterogeneous nuclear ribonucleoproteins (HNRNPs) and other splicing factors such as helicases (Cordin & Beggs, 2013). These *trans*-acting factors direct the spliceosome assembly across exons to mediate exon definition (Conti et al., 2013). *Trans*-acting factors can inhibit or activate cognate splice site usage and their activity depends on their binding site position within the pre-mRNA (Erkelenz et al., 2013). Generally, HNRNP proteins inhibit splicing when bound to an exon (Rothrock et al., 2005; Mauger et al., 2008), but activate splicing when bound to an intron (Hui et al., 2003; Xiao et al., 2009), while exon-bound SR proteins activate splicing (Cartegni & Krainer, 2002; Shen et al., 2004) and intron-bound SR proteins repress splicing (Ibrahim et al., 2005; Buratti et al., 2007; Shen & Mattox, 2012). A position-dependent dual regulatory function as repressors or activators of splicing has also been shown for other *trans*-acting factors that are not part of the HNRNP- or SR-protein family, e.g. NOVA1 (Licatalosi et al., 2008) or RBFOX1 (Sun et al., 2012).

Most RNA binding domains of *trans*-acting splicing factors recognize between four and eight nucleotides of target RNA and thus define typical lengths of *cis*-regulatory elements. Longer *cis*-regulatory elements are however possible, since motifs are sometimes bound by multiple RNA binding domains of a single factor (Maris et al., 2005; Auweter et al., 2006; Valverde et al., 2008; Dominguez et al., 2010). These binding sites can be short and degenerate.

Conversely, a single nucleotide change in a *cis*-regulatory element may be sufficient to cause loss of binding of a *trans*-acting factor, while at the same time this single nucleotide change attracts another factor with even opposing impact on splicing regulation (Rahman et al., 2015).

Besides the contribution of linear sequences to *cis*-regulatory elements, additional alternative splicing regulation is mediated through pre-mRNA secondary structure (Oberstrass et al., 2005; Warf et al., 2009). For instance, the RNA G-quadruplex represents a secondary structure element that is frequently involved in splicing regulation (Conlon et al., 2016; Huang et al., 2017; Weldon et al., 2018). Guanine-rich sequences can fold into RNA G-quadruplexes by forming multiple stacks of four guanines that organize into a planar arrangement via Hoogsteen hydrogen bonding (Cammas & Millevoi, 2017). These structures are prevalent in the human transcriptome and evolutionary conserved (Kwok et al., 2016). Multiple RBPs were shown to interact with RNA G-quadruplexes (von Hacht et al., 2014; Liu et al., 2017) including members of the HNRNPH protein family (Decorsière et al., 2011; Fisette et al., 2012; Conlon et al., 2016).

1.4 Mechanisms of alternative splicing regulation

Four commonly observed mechanisms explain the positive- and negative regulation of exon inclusion mediated by SR- and HNRNP proteins (**Figure 3**). (1) Recruitment and stabilization of core splicing factors by SR- and HNRNP proteins mediates splicing regulation via exon definition (Graveley, 2000; Caputi & Zahler, 2002). (2) Alternatively, competition of SR- and HNRNP proteins with splicing repressors (Zhu et al., 2001; Paradis et al., 2007) or direct competition of HNRNP proteins with core splicing factors (Heiner et al., 2010; Zarnack et al., 2013) can activate or repress splicing, respectively. (3) Moreover, SR proteins may juxtapose the 5' and 3' splice site through bridging interactions between U1 snRNP and U2AF (Wu & Maniatis, 1993). In contrast, negative regulation of HNRNP proteins can be achieved by looping out entire exons (Martinez-Contreras et al., 2006; König et al., 2010). (4) Furthermore, multimerization and cooperative binding of HNRNP proteins along the exon can

lead to displacement of SR proteins and thus promotes exon skipping (Okunola & Krainer, 2009). Conversely, cooperative HNRNP assemblies were also shown to promote exon inclusion, although the mechanism is not yet known (Gueroussov et al., 2017).

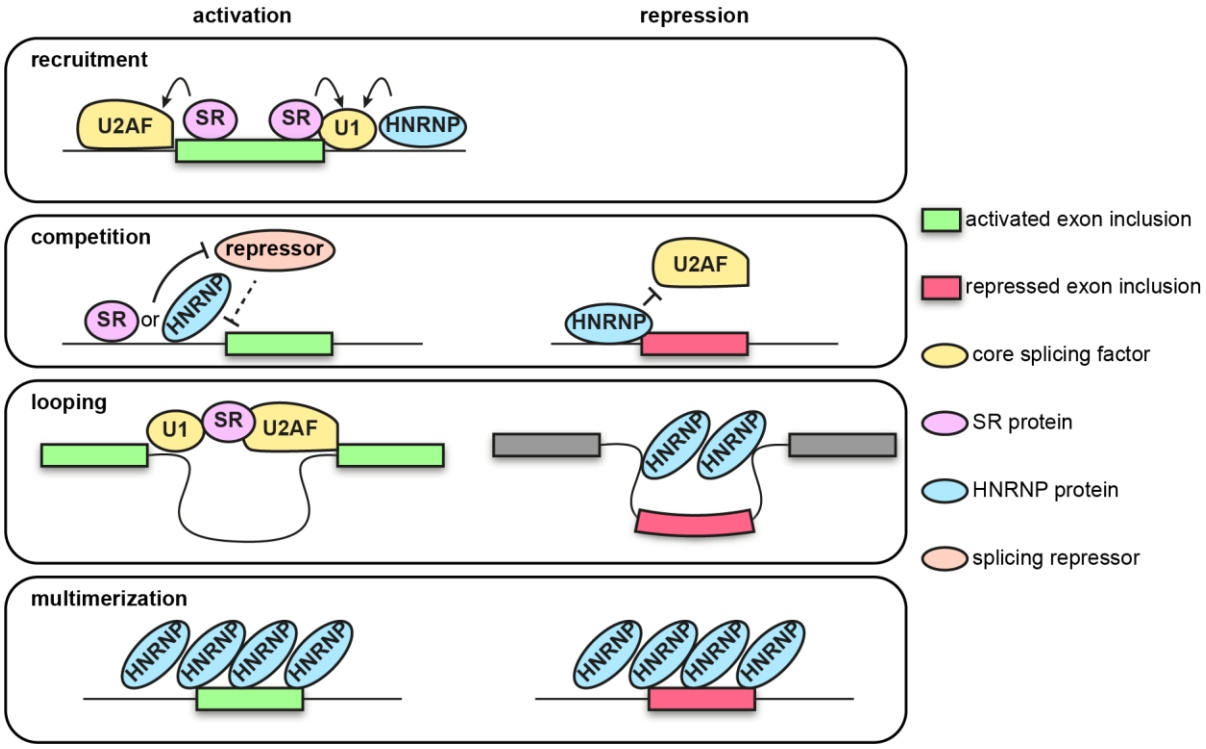


Figure 3: Mechanisms of SR- and HNRNP protein mediated splicing regulation. Recruitment of core splicing factors by SR- and HNRNP proteins activates exon inclusion. Competition of SR- or HNRNP proteins with splicing repressors activates exon inclusion, while competition of HNRNP proteins with core splicing factors represses exon inclusion. SR- or HNRNP protein mediated looping of pre-mRNA activates or represses exon inclusion, respectively. Multimerization of HNRNP proteins across pre-mRNA can both activate and repress exon inclusion. Exons and introns are depicted as boxes and lines, respectively.

1.5 The heterogeneous nuclear ribonucleoprotein H protein group

The HNRNP protein family is composed of at least 20 major members in mammals. They are not functionally defined but rather classified based on their efficient crosslinking with

nascent pre-mRNA. Thus, HNRNP proteins are functionally diverse and regulate various steps of RNA metabolism, including pre-mRNA processing, mRNA transport, translation regulation, and others. More than half of the HNRNP proteins including the members of the ubiquitously expressed HNRNPH group are however known to be involved in splicing regulation (Martinez-Contreras et al., 2007; Chaudhury et al., 2010; Katz et al., 2010). The HNRNPH group consists of the five members HNRNPH1, HNRNPH2, and HNRNPF, as well as the more distant relatives HNRNPH3a and HNRNPH3b (**Figure 4**) (Nazim et al., 2017).

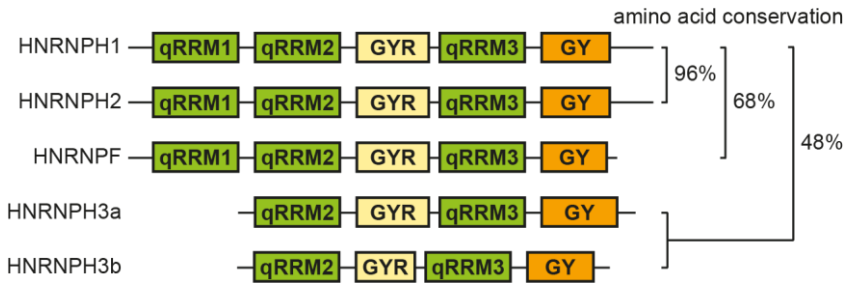


Figure 4: Domain structure and conservation of HNRNPH family members. The quasi-RNA recognition motifs (qRRM) mediate RNA binding, while the glycine-tyrosine-arginine-rich (GYR) domains control nuclear localization (van Dusen et al., 2010). A glycine-tyrosine-rich (GY) domain at the C-terminus mediates protein-protein interactions (Gueroussov et al., 2017). Adapted from (Nazim et al., 2017).

HNRNPH1 and HNRNPH2 share 96% amino acid identity and are henceforth commonly referred to as HNRNPH. While HNRNPF is similar to HNRNPH on the amino acid level, HNRNPH3a and HNRNPH3b are less conserved group members that also lack a quasi RNA recognition motif (qRRM) that the other group members have in common. The RNA binding motif of HNRNPH proteins consists of guanine-rich sequences (G-runs) (Romano et al., 2002; Marcucci et al., 2007; Camats et al., 2008; Masuda et al., 2008; Rahman et al., 2015; Uren et al., 2016; Nazim et al., 2017) and splicing changes upon reduced levels of HNRNPH correlate with the length of the G-run in the binding site (Xiao et al., 2009). RNA binding of HNRNPH and HNRNPF proteins is mediated via three qRRM domains. In contrast to the classical RRM domain of HNRNPA1 that specifically recognizes only the first guanine within a binding motif, each of the three guanines of a binding site are recognized by the qRRM (Ding et al., 1999; Dominguez et al., 2010). Depending on whether the binding motif is located in an

intron flanking an exon or the exon itself, HNRNPH either activates or represses exon inclusion, respectively (Xiao et al., 2009; Katz et al., 2010). Notably, HNRNPH related splicing defects have been linked to several diseases including cancer (Rauch et al., 2010; Lefave et al., 2011; Stark et al., 2011).

1.6 Alternative splicing in cancer

Aberrant splicing is commonly observed in cancer and splicing defects occur in transcripts coding for genes associated with hallmarks of cancer including invasion, metastasis and angiogenesis (Amin et al., 2011; Bonomi et al., 2013; Sveen et al., 2016). Moreover, cancer-associated genes are particularly susceptible to splicing defects (Sterne-Weiler & Sanford, 2014). Notably, several splicing factors are considered as tumor suppressors or proto-oncogenes, since dysregulated expression of splicing factors may cause splicing changes in downstream target mRNAs that are involved in cancer-related processes (Dvinge et al., 2016). For instance, HNRNPH controlled splicing changes were shown to promote survival and metastasis in gliomas (Lefave et al., 2011). Furthermore, splicing-factor mutations may cause collapsing of physiological splicing and induce transcriptome-wide changes favorable for tumor cells (Seiler et al., 2018).

1.7 Receptor tyrosine kinase RON

Recepteur d'origine nantais (RON) is a receptor tyrosine kinase encoded by the proto-oncogene *MST1R* (also referred to as *RON*). Receptor tyrosine kinases share common structural features, including an extracellular ligand-binding domain, an intracellular protein tyrosine kinase domain and a transmembrane helix. Ligand binding promotes receptor dimerization and stimulation of the tyrosine kinase activity. Subsequent auto-phosphorylation of the protein tyrosine kinase domain eventually triggers downstream signaling transduction (Schlessinger, 2000). Under physiological conditions, enzymatic cleavage maturates a RON single-chain precursor into a 185 kDa heterodimer of a 35 kDa (α) and a 150 kDa (β) disulfide-linked chain

(Gaudino et al., 1994). The RON ligand macrophage-stimulating protein (MSP or MST1) is circulating through the blood in a biologically inactive form and requires proteolytic maturation before the RON receptor can be bound and activated (Wang et al., 1994a; Wang et al., 1994b). Physiological RON signaling regulates the immune response during inflammation and wound healing in humans (Yao et al., 2013; Faham & Welm, 2016). Although RON and MSP are evolutionary highly conserved, their function varies in different species (Quantin et al., 1995; Gaudino et al., 1995; Nakamura et al., 1996; Bassett, 2003).

RON is transcribed in epithelial cells of most tissues, although expressed at very low levels (Gaudino et al., 1994; Gaudino et al., 1995): Only 600 – 1,000 RON receptor proteins were estimated to be present per keratinocyte, whereas highly expressed receptor tyrosine kinases were estimated to be present with more than 70,000 molecules per cell (Roque et al., 1992; Wang et al., 1996). RON is nevertheless an essential protein, since homozygous knockout is embryonic lethal in mice (Muraoka et al., 1999).

Several RON isoforms with varying oncogenic properties are known. For instance, skipping of *RON* alternative exon 11 results in the isoform RONΔ165 and on the protein level causes an in-frame deletion of 49 amino acids. As a result, the receptor remains as a single-chain precursor in the cytoplasm. Additionally, enabled by an uneven number of cysteine residues, spontaneous, ligand-independent oligomerization of RONΔ165 leads to constitutive activation (Collesi et al., 1996; Zhou et al., 2003). In contrast to other *RON* splicing isoforms, RONΔ165 overexpression triggers gene expression changes similar to those induced by activation of the full-length protein, suggesting that RONΔ165 is a major source of carcinogenic effects mediated by aberrant RON activation (Chakedis et al., 2016). However, RONΔ165 does not induce tumor formation in nude mice, but is rather involved in tumor progression towards malignancy (Zhou et al., 2003). Consequently, RONΔ165 is particularly upregulated in metastatic cancer, and contributes to tumor 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). Moreover, RONΔ165 is frequently upregulated in solid tumours, including ovarian, pancreatic, breast and colon cancers (Ghigna et al., 2005;

Mayer et al., 2015; Chakedis et al., 2016). *RON* exon 11 splicing is extensively regulated by *trans*-acting factors including HNRNPH (Lefave et al., 2011; Bonomi et al., 2013; Moon et al., 2014a). Furthermore, several *cis*-regulatory elements that regulate *RON* exon 11 exon inclusion were already identified (Lefave et al., 2011; Moon et al., 2014b).

Given their prevalence across cancers, *RON* isoforms are an attractive target for therapeutic intervention (Yao et al., 2013) and likewise, monoclonal antibodies that bind RON and block MSP-RON signaling have been assessed in clinical trials (ClinicalTrials.gov Identifier: NCT01119456; antibody RON8, Narnatumab, ImClone; phase-I discontinued) (O'Toole et al., 2006). However, since antibody-mediated interference with MSP-RON signaling does not affect constitutive activation of RONΔ165, tumors expressing this isoform escape these kind of therapies (Chakedis et al., 2016). Therefore, knowledge about the impact of mutations on RONΔ165 splicing in patients might have enormous diagnostic value and promises to allow targeted therapy in the future.

1.8 Approaches to study splicing regulation

Trans-acting factors bind to distinct *cis*-regulatory sequence elements to help the spliceosome select introns from the nucleotide sequence of the pre-mRNA. The modular organization of these interactions suggests the existence of a 'splicing code', i.e. a code of pre-mRNA features that is decoded by specific RBPs. Extensive bioinformatics studies *in silico* predict splicing outcomes by integrating RNA feature information and RNA-sequencing (RNA-seq) data, thus aiming to decipher the splice code (Barash et al., 2013; Xiong et al., 2015). These algorithms can be used for inference of global trends or they may guide experimental mutation analysis. However, their predictive power may be low when evaluating the consequences of individual single-nucleotide changes on alternative splicing (Soukarieh et al., 2016). In order to study a certain splicing event in detail, experimental validation of the *in silico* predictions is therefore necessary and commonly involves mutagenesis of minigene reporters (Cooper, 2005). While elaborate serial mutagenesis experiments of entire exons

within minigenes have successfully enabled *cis*-regulatory element detection, they are time-consuming and do not allow assessment of all possible single-nucleotide changes or detection of more complex regulation involving epistasis between two or more mutations (Nasrin et al., 2014; Ahsan et al., 2017). Nevertheless, three recent high-throughput studies helped progressing towards a better understanding of the splicing code by combining minigene reporter mutagenesis with deep sequencing. This enabled accurate prediction of the impact of sequence variants on splicing, provided an almost complete mutational screening of a human exon, or allowed deduction of general splicing mechanisms (Rosenberg et al., 2015; Julien et al., 2016; Ke et al., 2018). In these studies, however, mutations were restricted to exonic regions or synthetic spacers, and *trans*-acting factors were not considered. These limitations raise the need for alternative approaches that enable studying splicing regulation in greater detail.

1.9 Aim of the project

The aim of this PhD project is to establish a high-throughput random mutagenesis screening to comprehensively characterize the regulatory landscape of exon 11 splicing from the *MST1R* gene. First, a library of mutated minigenes will be generated by mutagenic PCR, which will subsequently be transfected into human cells. The mutations in the minigene library and the splicing outcome will be characterized by next-generation DNA- and RNA-sequencing (DNA- and RNA-seq), respectively. Since the minigene contains *RON* exon 11, the neighbouring introns, and constitutive exons, the screening is expected to allow assessing the contribution of the exon surrounding to the regulation and thus will extend beyond the aforementioned studies. Patient's clinical data will be used to compare the determined mutation effects with *RON* splicing in cancer patients bearing the same mutations. Accordingly, the study will enable assessment of mutation effects potentially relevant in clinical diagnostics. Extending beyond the comprehensive characterization of *cis*-regulatory elements, knockdown experiments will provide information on the protein factors involved in the regulation. The study is therefore expected to explain the *RON* exon 11 splicing decision in unprecedented detail.

Materials and Methods

2.1 Materials

2.1.1 Plasmids

Table 1: List of plasmids used in this study. All plasmids share the pcDNA 3.1 (+) vector backbone (Invitrogen).

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_027	pcDNA-DEST53-GFP-SMU1	N-terminally tagged SMU1-GFP fusion protein
pS_028	pcDNA-DEST53-GFP-SMU1	C-terminally tagged SMU1-GFP fusion protein
pS_056	pcDNA_HNRNPH1	*HNRNPH1* overexpression construct

Table 2: List of plasmids containing point mutations or small insertions or deletions.

Plasmid ID	Mutation
1	G272A
2	G305A
3	C307G
4	C339T
5	A356G
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
29	G471C
30	G480T
31	G500A

Plasmid ID	Mutation
32	G555C

2.1.2 Oligonucleotides

Table 3: Oligonucleotides used in this study. Oligonucleotides were purchased 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

No.	Name	Sequence (5'-3')	Purpose
oS237	RT-PCR-endo_fwd2	GGGCAGTGGAAAGCAGGTGTGAG	RT-PCR for endogenous *RON* gene in minigene transfected cells
oS118	RON A	CAAGCAGAAGACGGCATACGAGATC GGTCTCGGCATTCCTGCTGAACCGC TCTTCCGATCTNNNNNNNNNNCTAT AGGGAGACCCAAGCTT	Illumina fwd sequencing primer for DNA-seq
oS119	RON B	CAAGCAGAAGACGGCATACGAGATC GGTCTCGGCATTCCTGCTGAACCGC TCTTCCGATCTNNNNNNNNNNGTTC CACTGAAGCCTGAG	Illumina fwd sequencing primer for DNA-seq and RNA-seq
oS120	RON C	CAAGCAGAAGACGGCATACGAGATC GGTCTCGGCATTCCTGCTGAACCGC TCTTCCGATCTNNNNNNNNNNAGCT GCCAGCACGAGTTC	Illumina fwd sequencing primer for DNA-seq
oS138	RON D	CAAGCAGAAGACGGCATACGAGATC GGTCTCGGCATTCCTGCTGAACCGC TCTTCCGATCTNNNNNNNNNNGAAT CTGAGTGCCCGAGG	Illumina fwd sequencing primer for DNA-seq
oS105	RON E	CAAGCAGAAGACGGCATACGAGATC GGTCTCGGCATTCCTGCTGAACCGC TCTTCCGATCTNNNNNNNNNNCTAC TGGCTGGTCCTCATGA	Illumina fwd sequencing primer for DNA-seq
oS106	P5 SOLEXA RON	AATGATACGGCGACCACCGAGATCT ACACTCTTTCCCTACACGACGCTCT TCCGATCTNNNNNNNNNNATAGAAT AGGGCCCTCTAGA	Illumina rev sequencing primer for DNA-seq and RNA-seq

No.	Name	Sequence (5'-3')	Purpose
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
oS227	CRNKL1_qPCR_fwd	ACAACGCTGCCTCGAGTTAA	RT-qPCR validation of *CRNKL1* KD
oS228	CRNKL1_qPCR_rev	CTCCATCCAGCGCTCAAACA	RT-qPCR validation of *CRNKL1* KD
oS229	PRPF8_qPCR_fwd	TTGGGAGCAGATTCGGGATG	RT-qPCR validation of *PRPF8* KD

No.	Name	Sequence (5'-3')	Purpose
oS230	PRPF8_qPCR_rev	TCGGCGCATCATAATCCACA	RT-qPCR validation of *PRPF8* KD
oS231	IK_qPCR_fwd	CCCAGGAAGAATACAGCGAGT	RT-qPCR validation of *IK* KD
oS232	IK_qPCR_rev	TCTTCCACTGGCGATCAAGC	RT-qPCR validation of *IK* KD
oS233	eIF4A3_qPCR_fwd	TGATCTCCCTAATAACAGAGAATTGT	RT-qPCR validation of *eIF4a3* KD
oS234	eIF4A3_qPCR_rev	TCTCTGAGGATGCGGATGTC	RT-qPCR validation of *eIF4a3* KD
oS241	CDC5L_qPCR_fwd	CAGGATCTTGATGGGGAGCTAA	RT-qPCR validation of *CDC5L* KD
oS242	CDC5L_qPCR_rev	AAATTCAGAAACACCACTAGTTTGA	RT-qPCR validation of *CDC5L* KD
oS243	SRSF1_qPCR_fwd	AGGCGGTCTGAAAACAGAGT	RT-qPCR validation of *SRSF1* KD
oS244	SRSF1_qPCR_rev	ACAAACTCCACGACACCAGT	RT-qPCR validation of *SRSF1* KD
oS245	RBM22_qPCR_fwd	ATCCCAGGCAGCCAGAGG	RT-qPCR validation of *RBM22* KD
oS246	RBM22_qPCR_rev	CAGAGGCTTCTTCTTCTGCT	RT-qPCR validation of *RBM22* KD
oS249	DEK_qPCR_fwd	ATGCATCCAAAGCCTTCTGG	RT-qPCR validation of *DEK* KD
oS250	DEK_qPCR_rev	GCCTTCCTTGCCATTCCAGA	RT-qPCR validation of *DEK* KD

No.	Name	Sequence (5'-3')	Purpose
oS251	EWSR1_qPCR_fwd	TATAGCCAACAGAGCAGCAG	RT-qPCR validation of *EWSR1* KD
oS252	EWSR1_qPCR_rev	CTCATGCTCCGGTTCTCTCC	RT-qPCR validation of *EWSR1* KD
oS253	HNRNPH2_qPCR_fwd	GTGGTGGTTATGGAGGTGGT	RT-qPCR validation of *HNRNPH2* KD
oS254	HNRNPH2_qPCR_rev	GTGCTCCTTCTCTACCTAAGCA	RT-qPCR validation of *HNRNPH2* KD
oS379	DHX9_qPCR_f	GCTGCCAGAGACTTTGTTAACTAT	RT-qPCR validation of *DHX9* KD
oS380	DHX9_qPCR_r	TGTTGGTAAATCTCCTTCAGCATT	RT-qPCR validation of *DHX9* KD
oS381	DHX36_qPCR_f	ACCCACCATCAAATGAGGCA	RT-qPCR validation of *DHX36* KD
oS382	DHX36_qPCR_r	TGTGGCTCAACGGGTAATCG	RT-qPCR validation of *DHX36* KD
oS289	HNRNPF_qPCR_f	GCCTGGTAGCAACAGAAACC	RT-qPCR validation of *HNRNPF* KD
oS290	HNRNPF_qPCR_r	GTGATCTTGGGTGTGGCTTT	RT-qPCR validation of *HNRNPF* KD
oS428	BamHI_HNRNPH1_fwd	CATGGATCCACCATGATGTTGGGCACGGAAGG	Cloning of HNRNPH1 overexpression construct
oS429	XbaI_HNRNPH1_rev	CATTCTAGACTATGCAATGTTTGATTGAAAATC	Cloning of HNRNPH1 overexpression construct

2.1.3 siRNAs

Table 4: List of siRNAs used in this study. All siRNAs were purchased from Sigma Aldrich.

No.	Target	Sequence
18	No target	UGGUUUACAUGUCGACUAA[dT][dT]
20	*RBM22*	CCAGUAAAUCUUGGAAUAA[dT][dT]
21	*EWSR1*	CACUGAGACUAGUCAACCU[dT][dT]
22	*PRPF6*	GAGAAGAUUGGGCAGCUUA[dT][dT]
23	*DEK*	GAAGAUGACUCGUUCCCAU[dT][dT]
24	*HNRNPH*	GGAGCUGGCUUUGAGAGGA[dT][dT]
25	*SRSF1*	ACGAUUGCCGCAUCUACGU[dT][dT]
26	*SRSF2*	AAUCCAGGUCGCGAUCGAA[dT][dT]
27	*CDC5L*	UUGACGUGCAAUUUCACUCGCUUGG[dT][dT]
28	*CDC5L*	AAUUUGUUAUGCCAGAUUCCUCGGC[dT][dT]
29	*SMU1*	GCACGAGAAGGAUGUGAUU[dT][dT]
30	*PUF60*	GCAGAUGAACUCGGUGAUG[dT][dT]
31	*CRNKL1*	CACAGUUUGAAAUACGACA[dT][dT]
32	*PRPF8*	CACGUAUCAAGAUUGGACU[dT][dT]
33	*IK*	CAUAUGAGCGGAAUGAGUU[dT][dT]
34	*eIF4A3*	AGCCACCUUCAGUAUCUCA[dT][dT]
52	*HNRNPF*	UGAGAAGGCUCUAGGGAAA[dT][dT]
67	*DHX9*	CAAACCUUGAGCAACGGAA[dT][dT]
68	*DHX36*	CAGCUAUUAUAGACUUGAU[dT][dT]

2.1.4 Antibodies

Table 5: List of antibodies used in this study.

Antigen	Dilution	Origin	Product number	Supplier
GFP	1:500	mouse	sc-9996	Santa Cruz
GFP trap	20 µl slurry/ sample	lama	gtma-100	Chromotek
SMU1	1:1,000	mouse	sc-100896	Santa Cruz
HNRNPA1	1:20,000	mouse	R4528	Sigma Aldrich
HNRNPH	1:10,000	rabbit	AB10374	Abcam
HNRNPF	1:500	mouse	3H4	Santa Cruz
mouse IgG	1:5,000	horse	7076	Cell Signalling
rabbit IgG	1:5,000	goat	7074	Cell Signalling

2.1.5 Buffers

Table 6: 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

Table 7: Ponceau red staining solution.

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

Table 8: Buffer A+ composition.

Component	Final concentration
Hepes KOH pH 7.6	10 mM
$MgCl_2$	1.5 mM
KCl	10 mM
add immediately before use:	
IGEPAL CA-630	0.1% (v/v)
cOmplete™ Protease Inhibitor Cocktail	1 tablet/ 10 ml
DTT	0.5 mM

Table 9: Buffer C+ composition.

Component	Final concentration
Hepes KOH pH 7.6	20 mM
$MgCl_2$	2 mM
NaCl	420 mM
Glycerol	20% (v/v)
add immediately before use:	
IGEPAL CA-630	0.1% (v/v)
cOmplete™ Protease Inhibitor Cocktail	1 tablet/ 10 ml
DTT	0.5 mM

Table 10: Binding buffer composition.

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

Table 11: Washing buffer B composition.

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

Table 12: 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

Table 13: 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

Table 14: 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)
Triton X-100	25 µl	0.05% (vol/vol)
Mineral oil	to 50 ml	

2.2 Methods

2.2.1 Generation of recombinant plasmid DNA by cloning

The *RON* wt plasmid was generated via 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 65 °C annealing temperature with human genomic DNA (Promega) as a template. 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 raise AE inclusion in the *RON* wt minigene comparable to endogenous levels, the first nucleotide of the alternative exon was exchanged to a guanine. Plasmids containing point mutations were generated using the Q5 Site-Directed Mutagenesis Kit (NEB) according to the manufacturer's instructions.

The N- and C-terminal tagged SMU1 expression vectors were cloned using the Gateway™ cloning technology (Liang et al., 2013) starting from a pENTR™ vector containing a SMU1 open-reading frame provided by the IMB core facilities.

The *HNRNPH1* open-reading-frame was PCR-amplified from a plasmid kindly provided by Dr. Davor Lessel from the Institute of Human Genetics at the University Medical Center Hamburg-Eppendorf using oS428 and oS429 (**Table 3**) 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 verified by Sanger sequencing.

2.2.2 SDS PAGE and Western Blotting

To analyze protein samples by Western Blot, samples were first separated by SDS-PAGE using precast 4-12% NuPAGE Bis-Tris gels (Invitrogen) and MOPS SDS running buffer (Invitrogen). Proteins were subsequently transferred to a 0.45 µm pore size nitrocellulose membrane (GE Healthcare). Prior to Western Blot analysis, loading was controlled by staining membranes using Ponceau red staining solution (**Table 7**). Antibodies used for Western Blot analysis are listed in **Table 5**.

2.2.3 Characterization of SMU1 interactome by co-immunoprecipitation and mass spectrometry

To prepare samples for co-immunoprecipitation of proteins interacting with SMU1-GFP fusion protein and subsequent mass spectrometry analysis, two strategies were applied. First, to allow detection of RNA-dependent interactions, a triple labelling strategy of the stable isotope labeling using amino acids in cell culture (SILAC) method was carried out with three co-immunoprecipitations using either GFP-, SMU1-GFP with-, or SMU1-GFP without subsequent RNase treatment as bait proteins (Ong et al., 2002). In a complementary mixing experiment, co-immunoprecipitations of GFP and SMU1-GFP were carried out using light and heavy labelled HEK293T cells, respectively, and in contrast to the triple labelling strategy, the light and heavy labelled cell lysates were mixed before co-immunoprecipitation to allow swapping of transient interactors between the bait proteins GFP and SMU1-GFP (Hildebrandt et al., 2017).

2.2.3.1 Preparation

For metabolic labelling, HEK293T cells were grown for at least five doublings in RPMI 1640 (−Arg, −Lys) medium containing 10% dialyzed fetal bovine serum (PAA) and either light-, medium-, or heavy amino acids by supplementation of L-arginine and L-lysine, L-arginine ($^{13}C_6$), and L-lysine ($^{2}H_4$) or L-arginine ($^{13}C_6$ $^{15}N_4$) and L-lysine ($^{13}C_6$ $^{15}N_2$)

(Cambridge Isotope Laboratories), respectively. For transfection of the fusion proteins, 1.3 x 10^6 HEK293T cells resulting in ~ 70% confluence cells for transfection were seeded in 10 cm dishes. The next day, GFP or SMU1-GFP overexpression plasmids were transfected by incubating 10 µg plasmid DNA and 100 µg polyethylenimine MW ~ 2500 transfection reagent (Polysciences, Inc.) in 500 µl RPMI 1640 (−Arg, −Lys) medium for 15 min prior to addition to the cells.

2.2.3.2 Cell harvest and lysis

Two days after transfection, cells were washed twice using 10 ml of cold PBS and for the mixing experiment, cells were combined at this step. Next, cells were lysed in 300 µl lysis buffer (**Table 6**) per sample and subsequently collected using a cell scraper. For the mixing experiment, cells were incubated in 600 µl lysis buffer (**Table 6**) for 5 min on ice. Samples of both experiments were then further processed by centrifugation at 16,000xg for 15 min at 4 °C and subsequent collection of the supernatants. Next, protein concentrations were determined using the Bradford assay (Bio-Rad).

2.2.3.3 Co-immunoprecipitation

Co-immunoprecipitations of GFP or SMU1-GFP fusion proteins were carried out using GFP-Trap® agarose beads (Chromotek). First, beads were equilibrated by washing them three times with 1 ml lysis buffer (**Table 6**). Next, 20 µl of bead slurry (40 µl for the mixing experiment) was rotated with 1.3 mg of lysate for 1 hour at 4 °C. Following the incubation, beads were washed 3x with 1 ml lysis buffer (**Table 6**; 6x for the mixing experiment).

2.2.3.4 RNase digest and combining of samples for mass spectrometry analysis

The RNase treated sample was incubated with 7 µl RNase A (7 U/µl, Qiagen) and 2 µl RNase T1 (1,000 U/µl, Thermo Scientific) in 100 µl lysis buffer (**Table 6**) for 30 min at 4 °C.

Next, all samples including the RNase treated sample were washed 3x using 1ml lysis buffer (**Table 6**). The three samples of the triple SILAC experiment and the two samples of the mixing experiment were combined, and then both combined samples were washed again using 1 ml lysis buffer (**Table 6**). Subsequently, samples were prepared for mass spectrometry analysis as described in (Hildebrandt et al., 2017). Mass spectrometry analysis was carried out by the Beli laboratory at the Institute of Molecular Biology, Mainz.

2.2.4 RNA-pulldown coupled mass spectrometry

To identify interactors of specific sequences within the *RON* minigene transcript, 59 nt RNA of target regions with *RON* wild type sequence or a single point mutation in the center were produced as *in vitro* transcripts containing a streptavidin aptamer at their 3'-ends. In order to analyze specific RNA-protein complexes of either sequence pair, the wild type RNA and the point mutation containing RNA were used as bait molecules in light and heavy SILAC labeled HEK293T nuclear extracts. For the metabolic labeling, HEK293T cells were grown for at least five doublings in RPMI 1640 (−Arg, −Lys) medium containing 10% dialyzed fetal bovine serum (PAA) and either light- or heavy amino acids by supplementation of L-arginine and L-lysine or L-arginine (13C6 15N4) and L-lysine (13C6 15N2) (Cambridge Isotope Laboratories), respectively.

2.2.4.1 Preparation of nuclear extract

Metabolically labeled HEK293T cells were harvested by washing off using PBS and subsequent centrifugation of the cell suspension at 400xg for 5 min at 4 °C. Next, the cell pellets were washed with cold PBS and pellets corresponding to light or heavy labeled cells were combined. The pellet was then washed again using cold PBS and collected by centrifugation at 400xg for 5 min. Next, the volume of the pellet was determined (5 ml pellet from cells grown to confluence in 40x 15 cm plates). Now the cell pellet was resuspended in five volumes of cold Buffer A+ (**Table 8**) and incubated for 10 min on ice. Next, the cell pellet was collected by centrifugation at 400xg for 5 min and subsequently the cell pellet volume was

determined again (5 ml from both light and heavy labelled cells). The cell pellet was then resuspended in two volumes of cold Buffer A+ (**Table 8**) before being transferred to a dounce homogenizer. Following 40 strokes with a type B pestle (tight), cells lysis was microscopically confirmed using trypan blue (<4% living cells). The suspension was next separated into a cytoplasmic fraction and a fraction containing the nuclei by centrifugation at 3,900 rpm for 15 min. Following determination its volume (2 ml from both light and heavy labelled cells), the cell pellets were resuspended in two volumes of buffer C+ (**Table 9**) and subsequently rotated for 60 min at 4 °C. Using an ultracentrifuge (Heraeus Fresco 21), the nuclear extract was collected by centrifugation at 14,000 rpm for 60 min and then snap frozen in liquid nitrogen and stored at -80 °C.

2.2.4.2 Pulldown using nuclear extract and RNA as bait

For equilibrating the beads, Dynabeads MyOne Steptavidin C1 beads were washed twice with RNA binding buffer. To allow binding of the bait RNA to the beads, 52.5 µl of *in vitro* transcribed RNA was rotated with 105 µl of beads in 92.5 µl RNA binding buffer (**Table 12**) for 30 min at 4 °C. Next, the beads were washed three times with RNA binding buffer. Then, 400 µg of HEK293T nuclear extract was mixed with 20 µg of Yeast tRNAs (Butter laboratory) as a competing RNA from a different species in a total volume of 200 µl filled with RNA wash buffer. 50 µl of RNA-coupled beads were then rotated with 200 µl of the mixed extract for 30 min at 4 °C. Next, the supernatant was discarded and the beads were washed 3x with 200 µl RNA wash buffer. Interacting proteins were then eluted by incubating the beads in 25 µl NuPAGE LDS buffer containing DTT (Invitrogen) at 70 °C for 10 min. Further sample processing including mass spectrometry analysis of interacting proteins was carried out by the Butter laboratory at IMB.

2.2.5 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. To this end, reverse transcription was carried out 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 condition as follows: 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 used to amplify the minigene derived isoforms were oS066 (forward primer) and oS067 (reverse primer) and anneal to the upstream constitutive exon and a region located downstream of the random barcode but upstream of the polyadenylation site. The primers to amplify endogenously derived isoforms were oS044 (forward primer) and oS045 (reverse primer). To analyze 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.

2.2.6 Quantification of mRNA levels by RT-qPCR

To quantify mRNA levels of siRNA treated cells, cDNA was analyzed 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). 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 **Table 3**.

2.2.7 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.

2.2.7.1 siRNA-mediated knockdown of target mRNA

RNA interference induced knockdown of target mRNAs was carried out using single small interfering RNAs (siRNAs) at a final concentration of 20 nM synthesized by Sigma-Aldrich. For gradual *HNRNPH* KD, the siRNA concentration was varied between 0.05 nM and 10 nM. 1 d prior to 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 3 d 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 analyzes. Sequences of employed siRNAs are listed in **Table 4**.

2.2.7.2 Transfection of minigenes

Transfection of 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.

2.2.7.3 TGF-beta-induced epithelial to mesenchymal transition

To compare splicing changes between cells with epithelial characteristics and cells that underwent epithelial to mesenchymal transition (EMT), MDA-MB-435S were treated for six days with either 2.5 - 20 ng/ml TGF-beta (Peprotech GmbH) or non-TGF-beta containing normal growth medium. MCF10A cells were treated with 20 ng/ml TGF-beta or non-TGF-beta containing normal growth medium for nine days. Respective media were replaced every two days and MDA-MB-435S and MCF10A cells were split 1:4 and 1:5 when reaching confluence, respectively. EMT-induced morphological changes were microscopically controlled and expression of EMT target genes was analyzed by semi quantitative RT-PCR. Primers used for RT-PCR are listed in **Table 3**.

2.2.8 Preparation of mutated *RON* minigene library

Mutagenesis of the *RON* minigene was based on error-prone PCR amplification of the wild type minigene using the GeneMorph II Kit (Agilent). The target mutation frequency was controlled via the input DNA amount and the number of PCR cycles. When lower input DNA amounts are used, more target duplications, i.e. PCR cycles are required to reach saturated PCR product levels and thus mutation frequencies are increased.

Aiming for a mutation rate of three to four mutations per minigene, three independent libraries were generated by mutagenic PCR with 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 tag 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.

To allow formation of single bacterial colonies with 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.

2.2.8.1 Library quality controls

Mutation frequencies of individual minigenes 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.

2.2.8.2 Library amplification

To amplify preexisting 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).

2.2.9 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 (**Table 14**) 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 (**Table 3**), 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.

2.2.10 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 analyzed 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.

2.2.11 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 (**Table 11**). 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 abovementioned 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 analyzed 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.

2.3 Supplementary Methods

The methods described in the following sections were mainly conducted by F. X. Reymond Sutandy (iCLIP experiments), Mariela Cortés-López, Dr. Anke Busch, Samarth Thonta Setty, and Dr. Kathi Zarnack (bioinformatics analyzes), Bernardo P. de Almeida and Dr. Nuno L. Barbosa-Morais (TCGA and GTEx analyzes), and Dr. Mihaela Enculescu and Dr. Stefan Legewie (mathematical modelling). They are included in this thesis for reasons of clarity and comprehension.

2.3.1 iCLIP experiment and data processing

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). Sequencing reads were first filtered for quality in the experimental and random barcode, and then the adaptor sequences were trimmed. 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. In order to quantitatively compare HNRNPH iCLIP data for the *RON* wt and point mutation minigenes,

crosslink events were normalized to the total number of crosslink events within the minigene region excluding *RON* exon 11. Normalized counts were averaged between replicates, counted into 5-nt sliding windows and then subtracted between conditions to determine differences in HNRNPH crosslinking.

2.3.2 DNA-seq data processing and mutation calling

The DNA-seq library was sequenced on Illumina MiSeq (300-nt paired-end) with a total of 40 million reads and analyzed with a custom Python pipeline (version 2.7.9: Anaconda 2.2.0, 64-bit). In detail, FastQC (fastqc_v0.11.3; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for quality control, whereas Trimmomatic (version 0.33; parameters HEADCROP:20 SLIDINGWINDOW: 7:10 MINLEN:0) (Bolger et al., 2014) was used for removal of excess sequence and trimming of low-quality bases (average Phred score < 10 in 7-nt window). After trimming, reads were filtered for a minimum length of 130 nt (read #1) and 90 nt (read #2). In order to extract the 15-nt barcode (read #1) which assigns the read pairs to an individual minigene variant, matchLRPatterns() from the R/Bioconductor package 'Biostrings' were used to search for the flanking restriction sites (Lpattern="TCTAGA", Rpattern="GGATCC", allowing one mismatch). Read pairs were only retained when having a Phred score ≥ 30 at all barcode positions. For each minigene variants with at least 640 read pairs, reads were mapped to the sequence of the *RON* wt minigene 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. 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. 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. The identified mutations include 18,948 point mutations as well as

608 short insertions and deletions. The latter were taken into account as independent sequence variants in the mathematical splicing model, but ignored in the subsequent analyzes. 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 further false-negatives.

2.3.3 RNA-seq data processing and splicing isoform quantification

RNA-seq libraries were sequenced on Illumina MiSeq (300-nt paired-end), yielding 17-22 million reads per sample and analyzed with a custom Python pipeline similar to DNA-seq (see above). Briefly, 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 above. Only reads originating from the 5,791 minigene variants that were recovered from the DNA-seq library were considered for further analyzes. 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 input of prior knowledge of existing splice junctions. 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. 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 remaining 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.

2.3.4 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 relation to the total read pairs for all detected isoforms for this particular minigene variant. All kept non-canonical isoforms derived from cryptic splice site activation were collect in the isoform category 'other'.

2.3.5 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 $dx_0/dt = c - (r_1 + r_2 + r_3 + r_4 + r_5 + r_6)x_0$. 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 $dx_i/dt = r_i x_0 - d_i x_i$, where d_i are RNA degradation rates.

The measured isoform frequencies correspond in the model to the concentration of transcripts x_i normalized by the total RNA concentration. These fractions can be calculated analytically from the steady state of the system. As a result, the frequency p_i of a certain isoform *i* has the form $p_i = K_i/(K_1 + K_2 + K_3 + K_4 + K_5 + K_6)$. Here, the splicing rates $K_j = r_j/d_j$, $j = 1,2,4,5,6$ are the ratios of production and degradation rates for the isoforms involving splicing, and $K_3 = 1 + r_3/d_3$ 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, it was instructive to consider the isoform ratio relative to the inclusion isoform ($p_i/p_1 = K_i/K_1$), as this no longer depends on all splice rates, and relates to K_i in a linear fashion.

2.3.6 Calculation of single mutation effects by linear regression

For the estimation of 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 amount of minigene variants that were detected in the RNA-seq replicates.

To support the assumption of additive mutation effects, it was analyzed how single mutation effects 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 analyzed. For the majority of these mutations, the combined mutational effects on the splicing rates K_i were multiplicative, e.g. $K_i\left(m_{1,}m_2\right)/K_i\left(WT\right) = K_i\left(m_1\right)/K_i\left(WT\right) * K_i\left(m_2\right)/K_i\left(WT\right)$, where $K_i(WT), K_i(m_1), K_i(m_2)$ 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 $K_i\left(m_{1,}\dots m_n\right)/K_i\left(WT\right)$ 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.

As an alternative approach to estimate the single mutation effects, the median isoform frequency across all minigene variants that harbour a given mutation was calculated, and 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 this particular mutation.

2.3.7 Estimation of the inference accuracy of the model

The training dataset contained about ~600 mutations that were measured also as single mutations in individual minigenes. 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 $n-1$, with$1 < m \leq n-1$. Finally, the standard deviation of the prediction errors for all mutations was estimated for each measured frequency.

2.3.8 Definition of significant single mutation effects and synergistic interactions

The estimated single mutation effects on splice isoform ratios as obtained by linear regression could be 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: $\frac{({p_i}^{mutation} - mean({p_i}^{wt}))}{standard\ deviation\ ({p_i}^{wt})}$. The z-scores were

independently calculated per replicate and later averaged. Only mutations present in all three replicates were kept for further analyzes.

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 standard-normally distributed metric 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 ($\Delta IF \geq 5\%$), 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). 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.

2.3.9 Characterization of splicing-effective positions

Splice site strengths were predicted using the sequence analysis software MaxEntScan (Yeo & Burge, 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 (Pollard et al., 2010) were retrieved from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables; table: Mammal Cons, PhyloP46wayPlacental) for the genomic coordinates corresponding to the *RON* minigene (chr3:49933134 – 49933840, human genome version hg19).

2.3.10 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 32**, 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 HNRNPH's sequence specificity, 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 33** compares the median splicing effect (average of three biological replicates) of all G-run disrupting versus extending mutations for the 22 predicted HNRNPH SRBS.

2.3.11 Analysis of gene expression and alternative splicing across human healthy and cancer tissues

Normalized gene expression data for 11,688 *post mortem* samples from 30 human tissues, collected from 714 non-diseased human donors, were retrieved from the Genotype-

Tissue Expression (GTEx) project (v7) (GTEx Consortium, 2015). Normalized gene expression data from The Cancer Genome Atlas (TCGA) tumour samples (https://cancergenome.nih.gov/) were retrieved from Firebrowse (http://firebrowse.org/). Alternative splicing for both datasets was quantified using *psichomics* (version 1.2.1, https://github.com/nuno-agostinho/psichomics), using the default minimum coverage to calculate *RON* exon 11 percent spliced-in (PSI) values. *RON* gene expression and *RON* exon 11 PSIs were quantified for 2,743 normal samples, from 24 healthy human tissues, and 4,514 tumour samples, from 27 cancer types.

2.3.12 Calculation of single mutation effects in cancer

Exome sequencing data from TCGA tumour samples were downloaded from Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). A total of 153 patients bearing 55 different mutations within the region of our *RON* minigene were identified. The impact on splicing of each mutation in the TCGA tumour samples was quantified, per cohort, as the difference of *RON* exon 11 skipping (calculated as 1 - PSI) between mutated and non-mutated tumour samples. These differences were correlated with those derived from the skipping isoform frequencies observed in the screen for each mutation. Since observed correlations were affected by the minimum read coverage used to calculate PSIs, the correlation analysis was restricted to cohorts with an average of more than 24 reads mapping to the involved splice junctions (resulting in 117 patients from 14 cohorts harbouring 36 different mutations). The intrinsic variability of *RON* exon 11 inclusion levels in TCGA patient samples was calculated as the standard deviation of *RON* exon 11 PSI in 'unmutated' TCGA tumour samples (i.e. without a given mutation) from considered cohorts and with more than 24 reads mapping to the involved splice junctions.

2.3.13 Identification of candidate RBPs

A recent large-scale RBP KD screen tested the KD effect of >200 RBPs on splicing of *RON* exon 11 and other alternative exons in HeLa cells (Papasaikas et al., 2015). The study

used z-scores calculated from the percent spliced-in (PSI) upon siRNA treatment and the median absolute PSI deviation, divided by its standard deviation. A positive z-score indicates more AE inclusion upon RBP KD. Using a cut-off of |z-score| > 1.5, 125 RBPs showed a substantial effect on *RON* exon 11 splicing. These include 17 RBPs that also have predicted SRBS in the *RON* minigene.

In order to identify potential regulators of *RON* exon 11 splicing in humans, RBPs were searched whose expression correlated with *RON* exon 11 splicing in cancer. The correlation analysis was performed with 190 pre-selected RBPs, consisting of 65 identified via ATtRACT, 108 identified in the previously published RBP KD screen (Papasaikas et al., 2015), and 17 common to both approaches. The mRNA expression levels of the RBPs were Spearman-correlated with *RON* exon 11 inclusion levels across TCGA tumour samples. The significance of those correlations (ranked by minus base-10 logarithm of the associated p-value) was tested against those of all RBPs retrieved 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 first restricted 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 gene expression ranges. Moreover, linear regression analyzes were performed between the expression of each of the 190 pre-selected RBPs and *RON* exon 11 PSI in TCGA tumour samples, using the resulting slopes to quantitatively assess the relative magnitude of association between each RBP and *RON* exon 11 splicing.

2.3.14 Analysis of cooperativity and switch-like splicing behaviour

Changes 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 39**) were fitted using the Hill function

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

with x and y being vectors of experimentally determined HNRNPH levels and corresponding splicing outcomes (ΔPSI), respectively (**Fig. 6b**). $y_{min}, y_{max}, EC50, n_H$ are fitted parameters. Fitting was done by minimising the residual cost function

$$\chi^2 = \big(\Delta PSI - y(HNRNPH)\big)/\sigma_{\Delta PSI},$$

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 $y_{min} \in [-0.5,0], y_{max} \in [0,0.5], EC50 \in [0.1,2], n_H \in [1,20]$. The optimal parameter values found were

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

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

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

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

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:

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

Results

3.1 The *RON* minigene

In order to comprehensively characterize all point mutations that control *RON* exon 11 splicing, a high-throughput screening for splicing regulatory regions was established. The technique is based on random mutagenesis of minigene reporter constructs. The minigene construct comprised a segment of the *MST1R* gene including the alternative exon (AE) 11 together with the complete flanking introns and the constitutive exons 10 and 12 (**Figure 5**).

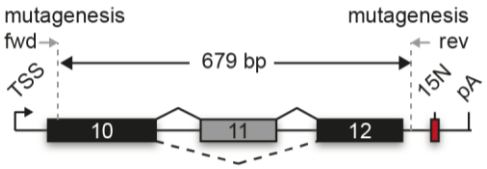


Figure 5: The *RON* minigene harbours the alternative exon 11 of the *MST1R* gene. The *RON* minigene consists of the AE 11 and the upstream and downstream constitutive exons 10 and 12, respectively. Mutagenesis of a 679 bp region was performed using error-prone PCR and the indicated forward and reverse primers (mutagenesis fwd and rev, respectively). TSS, transcriptional start site; pA, polyadenylation site; 15 N, the 15-nt barcode.

Upon mutagenic PCR based construction of the minigene library, the mutations were identified by high-throughput DNA-seq (**Figure 6**). Next, the library was transfected as a pool into human cells and the resulting alternatively spliced transcripts were analyzed by high-throughput RNA-seq. Importantly, a unique barcode sequence was used to tag each minigene and is also part of each transcript. This enabled unambiguous assignment of mutated minigenes and corresponding splicing outcomes. The barcode was located downstream of exon 12 with a

50-nt spacer of intronic sequence of the corresponding genomic segment in the *MST1R* gene. This prevented interference of the barcode sequence with splicing of the minigene reporter.

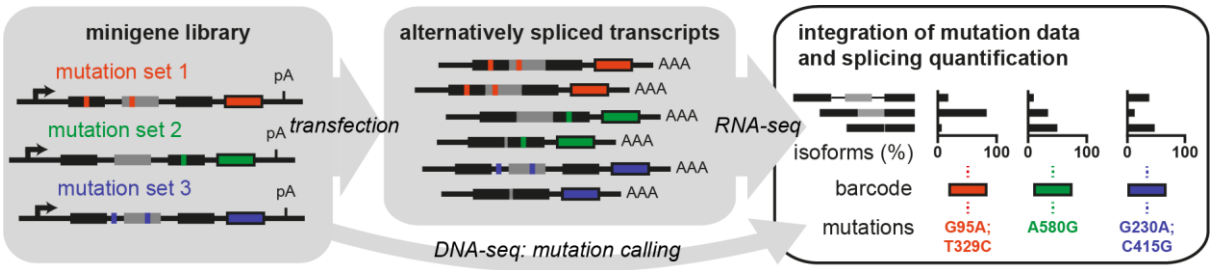


Figure 6: High-throughput screening for mutations affecting *RON* exon 11 splicing. A library of mutated minigenes is generated by mutagenic PCR (left). Upon transfection into human cells, the minigenes are transcribed and alternatively spliced (middle). Mutations in the library and corresponding splicing products are characterized by next-generation DNA-seq and RNA-seq, respectively. A unique molecular barcode that labels each minigene variant links mutations to their corresponding isoform quantification. Black and grey boxes indicate constitutive and alternative exons, respectively.

The minigene reporter system was chosen based on the following selection criteria:

i. Sequence length of the minigene reporter

ii. The minigene reporter is properly spliced

iii. Prior knowledge of *cis*-regulatory elements and *trans*-acting factors

iv. Clinical relevance of the splicing event

A detailed explanation of these criteria is provided in the following sections.

3.1.1 Sequence length of the minigene reporter

The Illumina next-generation sequencing technique was employed to characterize the mutations in the minigene library and the associated splicing outcome. However, the method is restricted by the sequence length of the minigene reporter. This is, because the experimental barcode and all relevant splice junctions must be part of the paired-end RNA-seq read to allow

unambiguous assignment of minigenes and their corresponding splicing products. The longest read lengths currently available for the Illumina MiSeq platform are 2 x 300 bp with the 600-cycle kit. Thus, 300 bp define the maximum distance between the 3'-end of the barcode and the downstream exon 3' splice site on one side, and the distance between the start of the opposite read and the 5' splice site of the alternative exon on the other side (**Figure 7**).

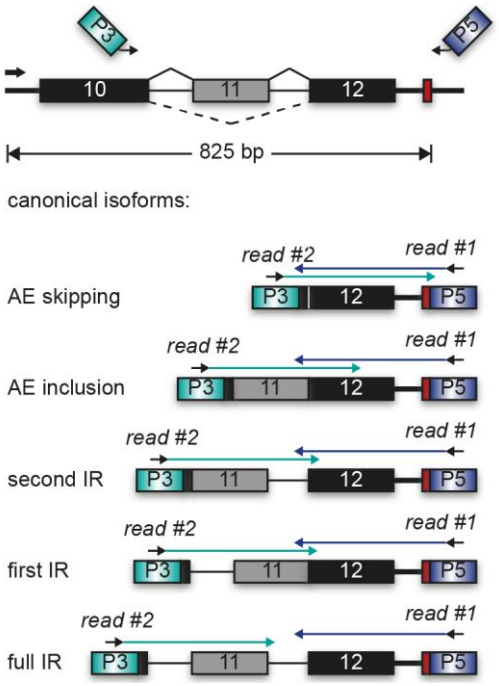


Figure 7: The five canonical isoforms are unambiguously identified by paired-end RNA-seq. Read #1 starting from the P5 adaptor provides the 15-nt barcode information and the splice junction upstream of exon 12, while read #2 from P3 reads the splice junction downstream of exon 10. For partial or full IR isoforms, both reads extend into the respective intron.

In addition, the amplicon should not exceed a sequence length of 1 kb, since longer fragments frequently fail during bridge amplification on the Illumina flow cell and do not form clusters required for high-throughput sequencing. Finally, the quality of the sequence reads drops with increasing read lengths and typically not the entire 300 bp read can be exploited for downstream analyzes (Schirmer et al., 2015).

3.1.2 The minigene reporter is properly spliced

Splicing of the minigene reporter had to be compared to endogenous *RON* splicing first, as minigene splicing patterns may differ from the endogenous splicing outcome. In the *RON* minigene, the thymidine at the alternative exon start was replaced with a guanine to increase the 3' splice site strength of the alternative exon (Yeo & Burge, 2004), a necessary step to obtain increased AE inclusion levels as observed from the endogenous *MST1R* gene (**Figure 8A**). RT-PCR based splicing analyses in different cell lines furthermore confirmed that the minigene splicing patterns are comparable to the endogenous locus. Therefore, the minigene allowed studying the natural regulation of *RON* exon 11 splicing.

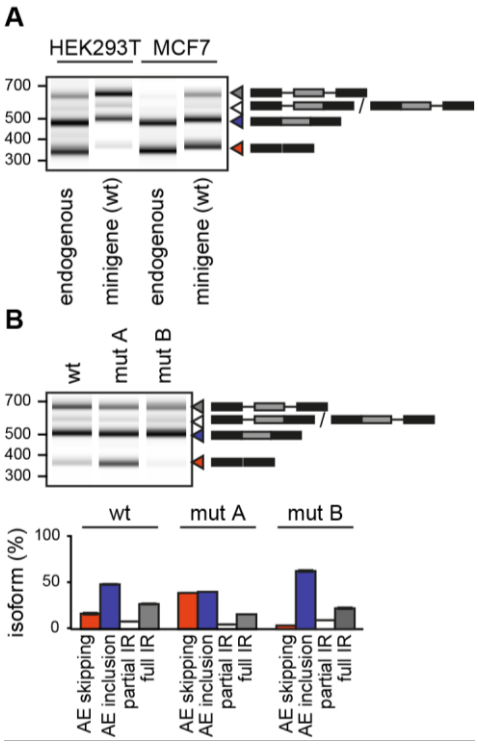


Figure 8: The *RON* minigene gives rise to identical isoforms as its endogenous counterpart and is regulated like expected. (**A**) RT-PCR analysis shows the *RON* minigene gives rise to the same splicing isoforms as the endogenous *MST1R* gene in HEK293T and MCF7 cells. Gel-like representation of capillary 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 (**Table 3**). (**B**) 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. 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.

3.1.3 Prior knowledge of *cis*-regulatory elements and *trans*-acting factors

The selection of an alternative splicing event for the random mutagenesis approach with several already known *cis*-regulatory elements and *trans*-acting factors allowed validation of mutation effects that are quantified from the screening. Furthermore, the presence of *cis*-regulatory elements evidences a regulated alternative splicing event and suggests additional, yet undiscovered *cis*-regulatory elements and *trans*-acting factors that are also involved in the regulation. Moreover, available mutation information can be used to assay correct regulation of the minigene reporter. Accordingly, two mutations that disrupt a HNRNPA1 (Bonomi et al., 2013) or HNRNPH binding site (Lefave et al., 2011) were validated to affect splicing in the direction of increased or decreased AE skipping, respectively (**Figure 8B**).

3.1.4 Clinical relevance of the splicing event

The AE skipping isoform RONΔ165 was previously shown to promote increased cell motility during epithelial to mesenchymal transition (EMT); (Ghigna et al., 2005). In order to test if endogenous *RON* splicing is altered upon EMT activation, EMT was induced in MDA-MB-435S cancer cells using TGF-beta and *RON* splicing was subsequently monitored by RT-PCR analysis (**Figure 9A**). Triggered EMT activation was microscopically confirmed by loosened cell-cell-contacts and altered cell morphology. This effect was accompanied by an upregulation of the mesenchymal marker *Fibronectin*, but not *N-cadherin*. Furthermore, *RON* splicing was altered towards the RONΔ165 isoform. To highlight the specificity of this splicing change, the experiment was repeated under similar cell confluence between TGF-beta-treated cells and untreated cells and with lower doses of TGF-beta for the EMT induction (**Figure 9B**). Again, *RON* splicing showed increased levels of AE skipping, suggesting that RONΔ165 is involved in the TGF-beta-induced EMT program in MDA-MB-435S cells.

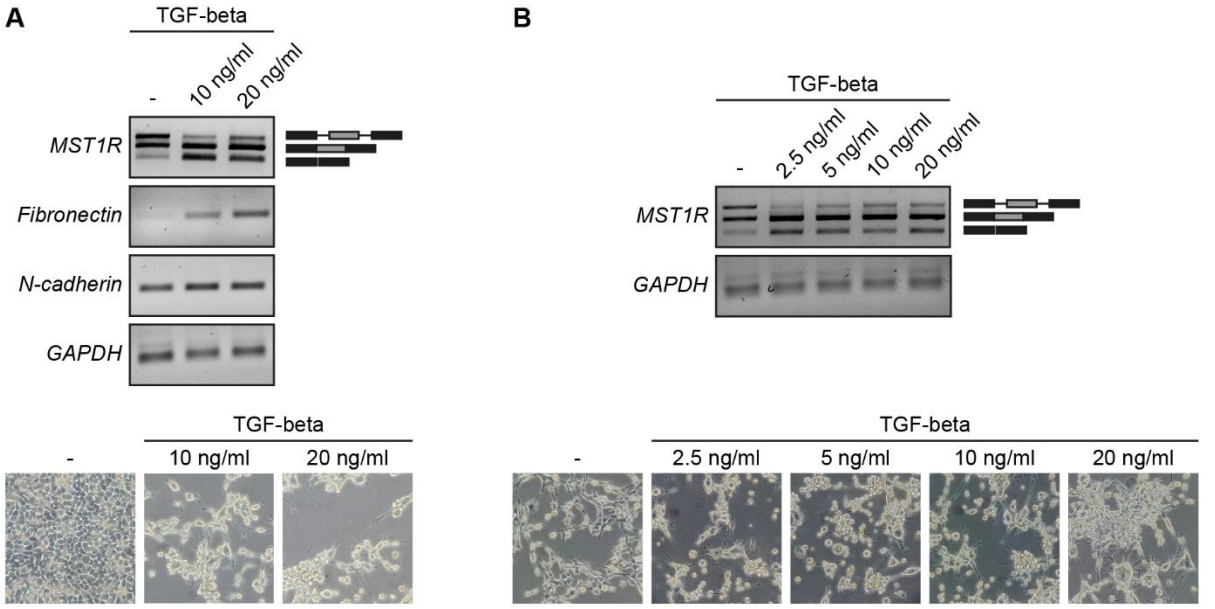


Figure 9: TGF-beta induced EMT in MDA-MB-435S cells. (**A**) MDA-MB-435S cells were treated with medium (10 ng/ml) or high dose (20 ng/ml) of TGF-beta for six days and splicing of *MST1R* (endogenous *RON* gene) and expression of the mesenchymal markers Fibronectin and N-cadherin were monitored by RT-PCR analyzes, while *GAPDH* expression was used as a reference (top). Microscope view of morphological changes associated with EMT-induction (100-fold magnification, bottom). (**B**) MDA-MB-435S cells were treated with varying doses of TGF-beta (2.5 - 20 ng/ml) for six days and *MST1R* (endogenous *RON* gene) splicing was monitored by RT-PCR analyzes, while *GAPDH* expression was used as a reference (top). Microscope view of morphological changes associated with EMT-induction (100-fold magnification, bottom).

To next assess, whether *RON* splicing is also altered upon EMT activation in other cell lines, EMT was induced in the non-tumorigenic epithelial cell line MCF10A (**Figure 10**). While nine days of TGF-beta treatment induced characteristic cell-morphology changes, the EMT process in MCF10A cells was not accompanied by a downregulation of the epithelial marker *E-cadherin* or changes in endogenous *RON* splicing. Together, these findings suggest that increased levels of *RON* AE skipping are produced upon TGF-beta mediated EMT-activation in MDA-MB-435S, but not MCF10A cells.

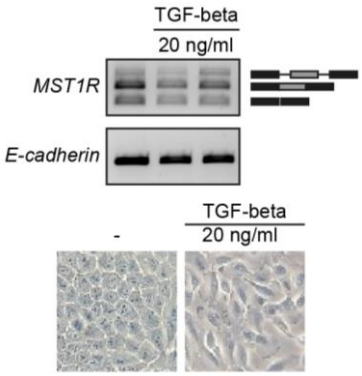


Figure 10: TGF-beta induced EMT in MCF10A cells. MCF10A cells were treated with high dose (20 ng/ml) of TGF-beta for nine days in biological duplicates and splicing of *MST1R* (endogenous *RON* gene) and expression of the epithelial marker *E-cadherin* were monitored by RT-PCR analyzes (top). Microscope view of morphological changes associated with EMT-induction (400-fold magnification, bottom).

3.2 Preparation of the *RON* minigene library

In order to generate a *RON* mutant minigene library with a target mutation rate of three to four mutations per minigene, variable amounts of template DNA and PCR cycles were used for an error-prone PCR amplification of the *RON* wt minigene (**Figure 11A**). 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 11B**). Sanger Sequencing of ten minigenes of each of the three libraries revealed comparable mutation frequencies with an average of three to four mutations per minigene and thus the libraries were pooled to result with a single *RON* minigene library of ~6,000 minigenes. In addition, the library was supplemented with a spike-in of 200 *RON* minigenes with the wt sequence, but variable barcodes for internal reference. RT-PCR analysis of a random selection of mutated minigene variants from the library revealed considerable variability in the splicing patterns, suggesting that the selected mutation frequency was reasonable (**Figure 11C**).

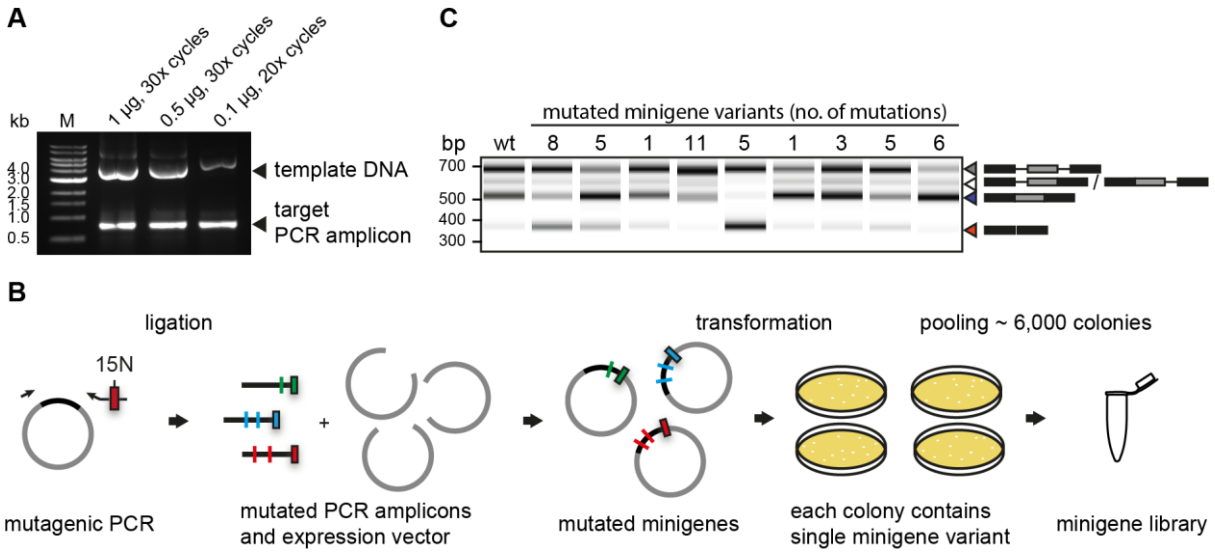


Figure 11: Preparation of a mutated *RON* minigene library. (**A**) Analytical agarose gel electrophoresis of fragments resulting from error prone PCR using varying amounts of *RON* wild type plasmid as template and indicated PCR cycles. The target PCR amplicon amount is saturated and equal for each 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 to result with a minigene library. (**C**) Mutated minigene variants display variable 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.

3.3 Mapping of mutations in the *RON* minigene library by high-throughput DNA-seq

In order to characterize the mutations of the minigene library, next-generation sequencing was carried out with 300-nt paired-end reads and five overlapping amplicons (**Figure 12A** and **Figure 12B**). Since the quality of the sequencing reads drops towards the end of the read and bigger amplicons are negatively biased during cluster formation on the flow

cell, five amplicons were required to result with sufficient coverage across the minigene after quality filtering (**Figure 12C**).

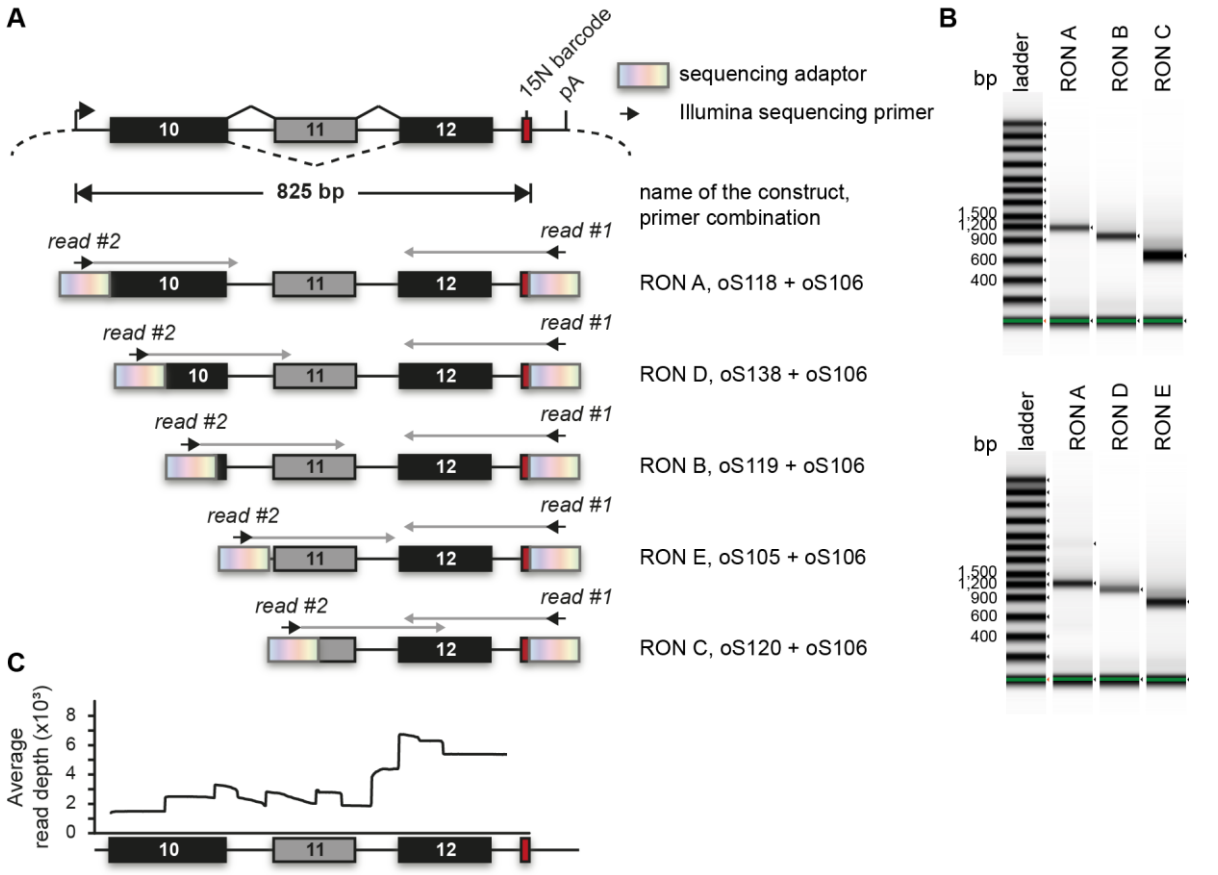


Figure 12: DNA-seq strategy employing five overlapping amplicons ensures high read coverage across the *RON* minigene sequence space. (**A**) Overview of the five overlapping amplicons generated for paired-end DNA-seq. The reverse primer binds downstream of the 15-nt barcode (15N, red box) and introduces 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. The graph is aligned with the overview shown in (**A**).

Importantly, the 15-nt barcode was included in each read pair and enabled reconstruction of the complete sequence of all minigene variants in the library. In total, 5,791 unique variants were captured including 5,200 with randomly introduced mutations and 591

with the wt sequence. 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 13A** and **Figure 13B**). 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 additional false-positives. Therefore, it was possible to screen mutation effects across the entire length of the *RON* minigene sequence.

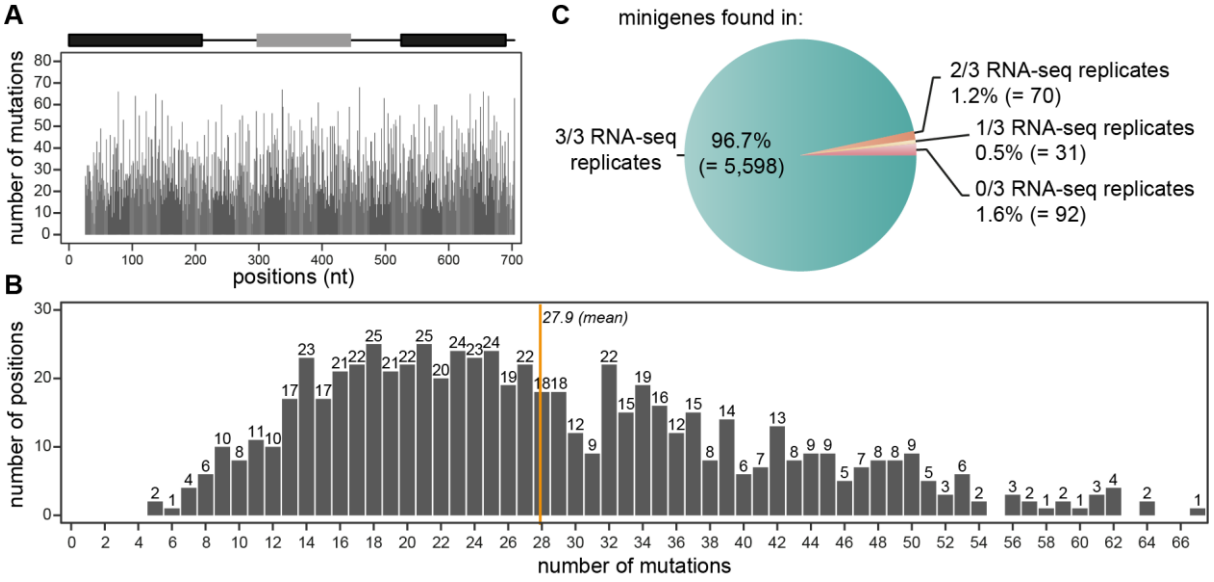


Figure 13: Mutations are equally distributed across the *RON* minigene region and the majority of minigenes is recovered in three RNA-seq replicates. (**A**) Mutations evenly distribute along the *RON* minigene positions. The bar chart shows the distribution of 18,948 point mutations across 5,791 minigenes. (**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 minigenes identified by DNA-seq is retrieved 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.

3.4 Quantification of alternative splicing of the *RON* minigene library by RNA-seq

In order to measure the splicing products of the *RON* minigene library, the library was transfected as a pool in HEK293T cells. 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 7** and **Table 3**). Again, each read pair contained the 15-nt barcode to allow assignment of the respective source minigenes. Three independent RNA-seq replicates were generated and 97% of minigenes were found in all three replicates (**Figure 13C**). While the canonical isoforms 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 constituted by 'other' isoforms originating from cryptic 3' – and 5' splice site usage (**Figure 14A** and **14B**).

Albeit low overall abundance, non-canonical 'other' isoforms can prevail the splicing products of 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 14A** and **14Figure 14F**). The wt minigenes showed little splicing variance, thus indicating the barcode does not impinge on the measurement of mutation effects (**Figure 14C**). In contrast, 45% of mutated minigenes were spliced with more than 10% deviation from the wt average AE inclusion, evidencing that the introduced mutations cause diverse splicing outcomes that can be explored in further analysis.

To validate the splicing quantifications, splicing of minigenes containing AE splice site mutations was monitored (**Figure 14C**). As expected, splice site mutations completely prevented AE inclusion, illustrating that the screening allows reliable detection of strong splicing changes. Furthermore, the precision of the splicing quantification was tested by comparison of the RNA-seq derived quantifications with splicing quantifications from RT-PCR measurements of 59 randomly selected minigene variants (**Figure 14D**). 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* sequencing space.

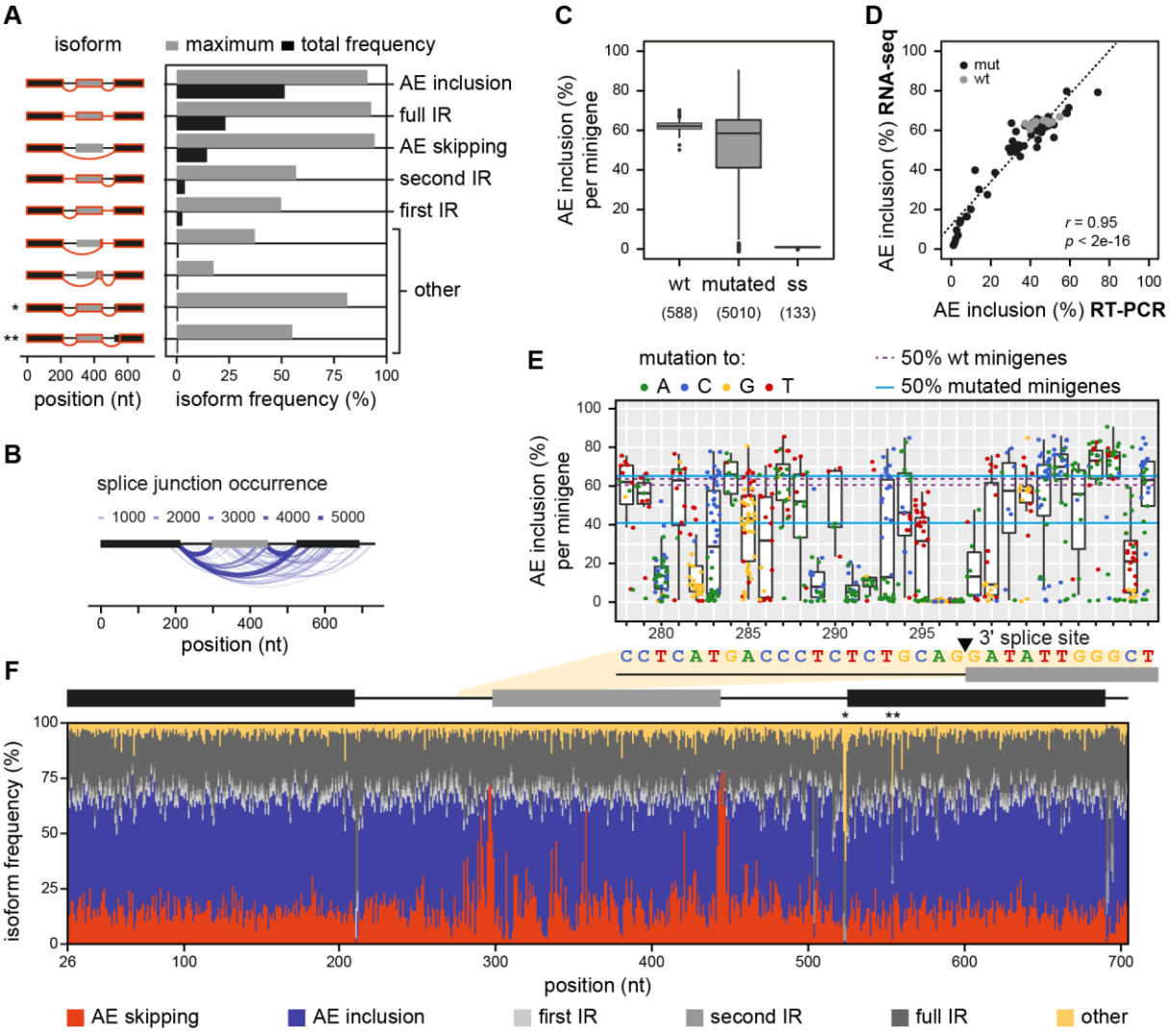


Figure 14: High-throughput mutagenesis screen of a minigene library provides quantitative splicing data across the *RON* minigene. (**A**) Canonical isoforms compose the most frequent splicing variants in the library. Bar diagram shows the nine most frequent isoforms sorted by total frequency in the RNA-seq library (black) along with the maximum frequency for an individual minigene variant (grey). (**B**) Splice junction occurrence. Line thickness and color 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**) High dynamic splicing range from 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. Number of minigenes in each category is given below. Whiskers correspond to 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 surrounding the AE 3'-splice site. Boxplots represents the isoform ratios displayed as AE inclusion (%) for minigenes harbouring a mutation at the respective site with colors indicating the inserted nucleobase (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 highlight positions that upon mutation cause increased formation of non-canonical isoforms depicted in (**B**).

3.5 Dissecting individual mutation effects with mathematical modelling

Minigenes of the library contain 3.6 mutations on average. Thus, the splicing outcome of a minigene variant is frequently a result of multiple mutations and a given mutation may display variable splicing effects depending on co-occurring mutations present in the same minigene (**Figure 14E**). In order to dissect individual mutation effects, linear regression modelling was applied (see Methods). In brief, a dynamic splicing model was formulated, that allowed calculation of the linear regression based on splice isoform ratios (isoform production versus isoform degradation relative 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. Non-linearity arises from different starting isoform frequencies of each minigene and the natural boundary of isoform frequencies ranging between 0% and 100%. Assuming that mutation effects are additive, the linear regression calculated single mutation effects (log-fold changes relative to the wt) from the sum of multiple mutation effects in a minigene. 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 15**). Indeed,

the sum of the single mutation effects agreed well with the combined mutation effects, supporting the assumption of the linear regression. Consequently, the regression model allows precise description of the experimentally measured isoform frequencies for each mutated minigene variant (**Figure 16**).

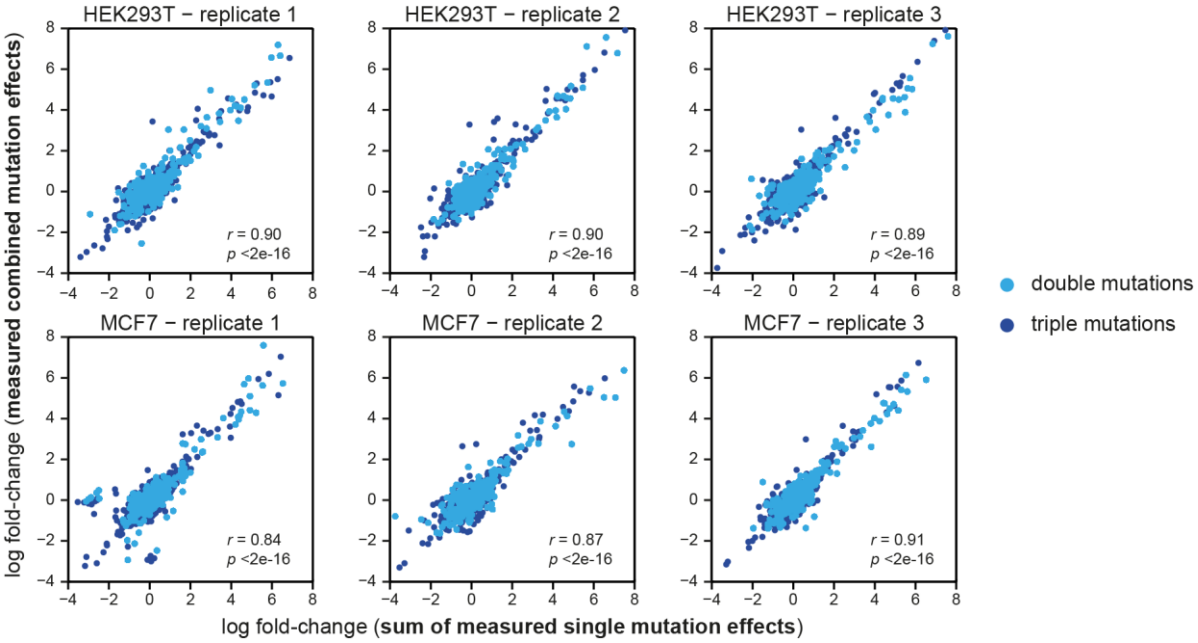


Figure 15: Mutation effects are additive. Scatterplots showing the correlation of mutation effects (log-fold changes relative to the wt) for the sum of single mutation effects and measured combined mutation effects from minigenes containing combinations of two (light blue) or three (dark blue) of individually measured mutations. Top and bottom panels show analyzes for three replicates in HEK293T and MCF7 cells, respectively. Pearson correlation coefficients *r* and associated *p*-values are provided in each panel.

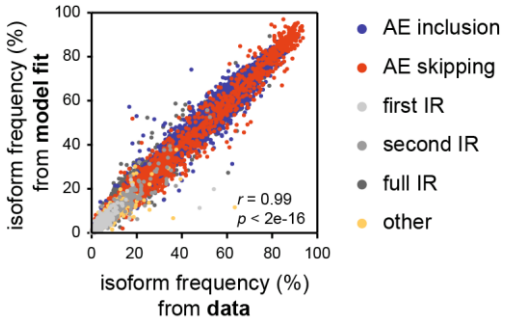


Figure 16: Inferred isoform frequencies from linear regression modelling correlate with measured isoform frequencies. Scatterplot showing the frequencies of splice isoforms for combined mutations calculated from the fitted model against the measured data of one biological replicate in HEK293T cells. Colors indicate distinct splice isoforms (see legend). Pearson correlation coefficient *r* and associated *p*-value are provided.

In order to validate the regression model, inferred single mutation effects were compared with RT-PCR derived splicing measurements of newly created minigenes harboring single mutations (**Figure 17**).

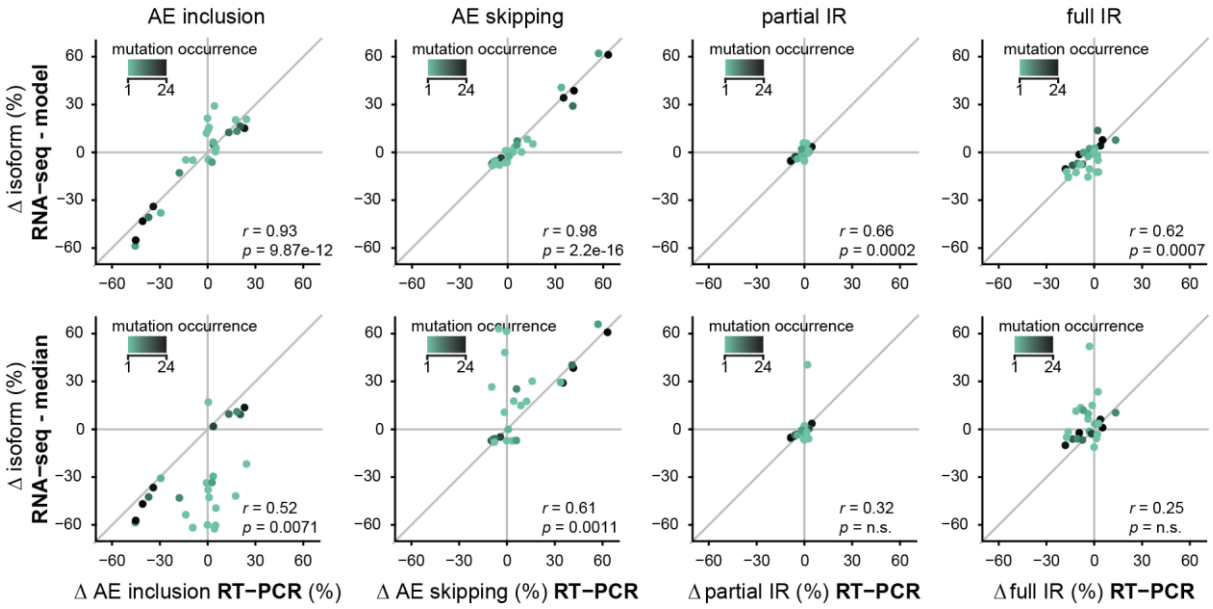


Figure 17: Linear regression provides more accurate estimations of single mutation effects than the median isoform frequency of minigenes that share a mutation at a given position. Effects of mutations that rarely occur in the library (color coded) correlate better with the model-inferred than the median-based 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 semiquantitative RT-PCR measurements (x-axes) of targeted minigenes harboring 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.

For comparison with the model performance, the median isoform frequency across all minigene variants that harbor a given mutation was additionally calculated as a simpler approach to estimate single mutation effects (**Figure 17**). However, this median-based estimation was outperformed by the linear regression derived single mutation effects. In particular, mutations that occurred with low frequency in the library, i.e. if only few minigenes shared a certain mutation, confounding mutations can easily shift the median away from the actual single mutation effect. Taken together, employing a linear regression model that assumes

additive mutation effects allowed accurate estimation of single mutation effects from measured combined mutation effects for ~1,800 mutations.

3.6 The regulatory landscape of *RON* exon 11 splicing

Altogether, the model predictions for HEK293T cells allowed estimation of mutation effects for ~1,800 single point mutations on five canonical isoforms. Splicing of at least one isoform was significantly altered by 778 mutations (henceforth called splicing-effective mutations; ≥ 5% change in isoform frequency, 5% false discovery rate, FDR; **Figure 18A**). Splice site disrupting mutations at the alternative exon almost completely prevented AE inclusion (**Figure 18B**). In contrast, effects at the poly-pyrimidine tract upstream of the AE 3' splice site were base-specific: While pyrimidine-maintaining transitions acted neutrally, transversions to purines reduced inclusion (e.g. position 290; **Figure 18B**). Similarly, degenerate positions of the branch point consensus sequence (positions 279, 281, and 283) were more robust against mutation than the strictly conserved U or A at positions 280 and 282, respectively (Gao et al., 2008; **Figure 18B**). Mutations at the splice sites of the flanking constitutive exons caused increased full IR and lowered AE inclusion levels (**Figure 18C**). Notably, the 5' splice site of the downstream constitutive exon affected AE inclusion, suggesting that exon definition at flanking exons may be required for splicing of the central exon. If exon definition at the upstream constitutive exon also contributed to splicing regulation of the AE could not be assessed, since the 3' splice site of the upstream constitutive exon was not mutagenized (**Figure 5**).

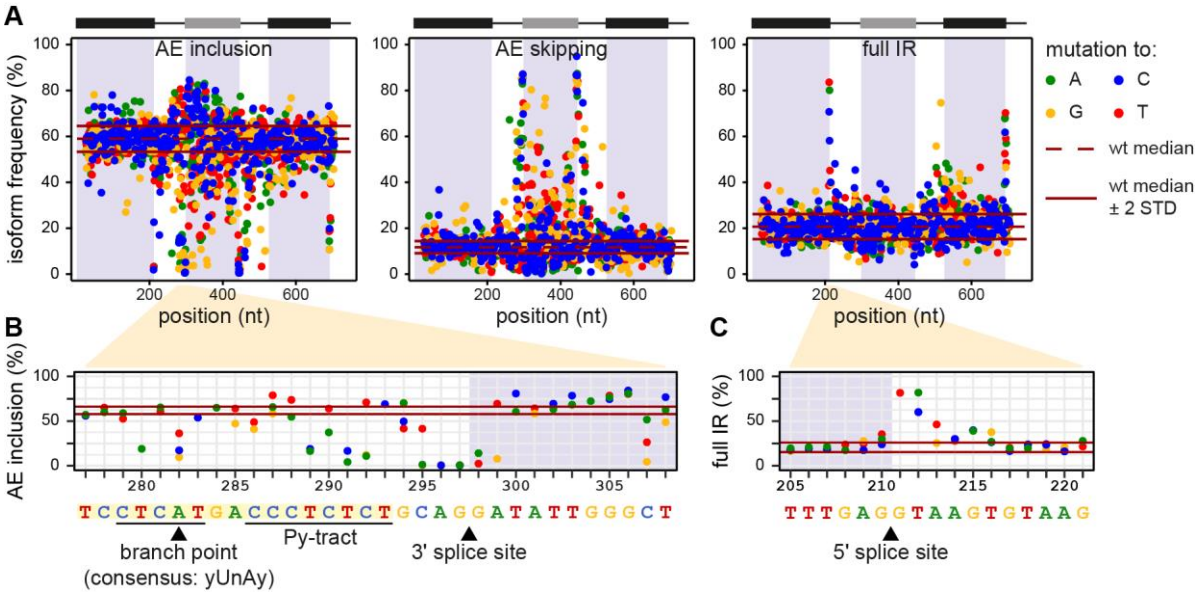


Figure 18: Effects of 1,800 point mutations on RON exon 11 splicing. (**A**) Linear regression modelling derived quantifications of single point mutations on AE inclusion- (left), AE skipping- (middle), and full IR isoforms (right) across the *RON* minigene positions in HEK293T cells. (**B**) Mutation effects on AE inclusion surrounding the AE 3'-splice site. Horizontal black lines and arrowheads mark core splicing signals, including branch point, polypyrimindine tract (Py-tract) as well as the 3' splice site. The plot shows the same region as in **Figure 14E**. (**C**) Mutation effects on full IR surrounding the 5'-splice site (black arrowhead) of the upstream constitutive exon.

To test whether the AE inclusion levels correlated with the splice-site strength of respective mutants at the 5' splice site, *in silico* splice site strength prediction of target input sequences was carried out using the MaxEntScan software (Yeo & Burge, 2004). The observed correlation suggested that *in silico* predictions of splice site strength can capture mutation effects at the 5' splice site of *RON* exon 11 (Spearman correlation coefficient, $r = 0.89$, p-value = 2.36e-08; **Figure 19A**). Predictions for 3' splice site strength, however, were less correlated (Spearman correlation coefficient, $r = 0.62$, p-value = 4.02e-07; **Figure 19B**), illustrating that splice site strength predictions are less accurate at this site. This might be attributed to the higher density of regulatory signals at the 3' splice site compared to the 5' splice site.

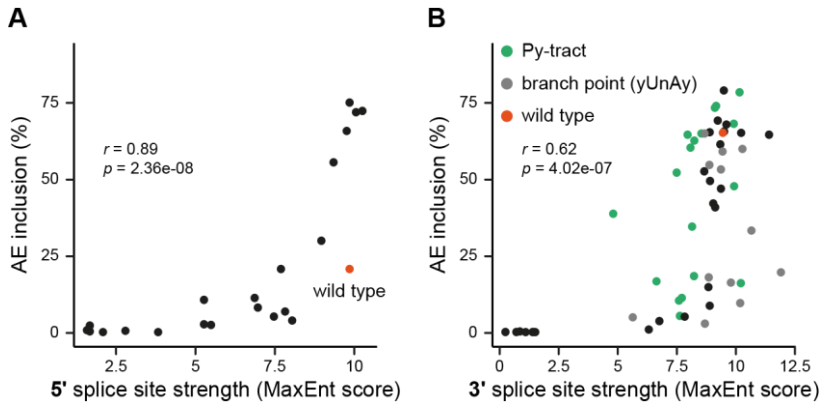


Figure 19: Quantified mutation effects at AE splice sites correlate with *in silico* predictions of splice site strengths in HEK293T cells. (**A**, **B**) AE inclusion correlates with the *in silico* predictions of 5'- (**A**) or 3' (**B**) splice site strength of respective sequences by the MaxEntScan software. Spearman correlation coefficients *r* and associated *p*-values are provided.

The highest density of splicing effective positions, i.e. positions with at least one splicing effective mutation, was found in the alternative exon. In fact, 91% of all positions within *RON* exon 11 (134/147 nt) were splicing effective (**Table 15**). Nevertheless, also the upstream and downstream introns contained 77% and 82% splicing effective positions, respectively, and about half of the positions in the surrounding constitutive exons were involved in the splicing regulation. This evidences that (1) the investigated sequence space is densely packed with splicing regulatory regions and (2) that splicing regulation is not only conferred by the nucleotides within an alternative exon but also contributed by the neighboring introns and constitutive exons. Particularly, the widespread occurrence of splicing effective positions across exons highlights the dual role of protein-coding- and splicing-regulatory function converging at exonic positions.

Table 15: Splicing effective mutations and positions per region in the *RON* minigene in HEK293T cells. Splicing-effective mutations (top) and positions (bottom) in HEK293T cells. The total number of possible and measured mutations/positions are indicated first, followed by the number of significant effects when considering any isoform at three cutoffs (>5%, >10%, and >20%). Additionally, the significant mutations (>5%) are given for each individual isoform. AE - alternative exon; IR - intron retention.

		Exon 10	Intron 10	Exon 11	Intron 11	Exon 12	Intron 12	Total
HEK239T	**Mutations**	555	261	441	240	498	42	2037
	Measured	487 (87.7%)	224 (85.8%)	381 (86.4%)	190 (79.2%)	430 (86.3%)	35 (83.3%)	1747 (85.8%)
	Any isoform > 5%	**117** (24%)	**118** (52.7%)	**270** (70.9%)	**108** (56.8%)	**144** (33.5%)	**21** (60%)	**778** (44.5%)
	AE inclusion	100	111	263	87	92	19	672
	AE skipping	20	67	185	53	29	9	363
	First IR	2	6	3	0	4	2	17
	Second IR	0	1	0	3	6	10	20
	Full IR	70	74	107	79	113	16	459
	Other	0	0	2	4	1	0	7
	Any isoform > 10%	**26** (5.3%)	**66** (29.5%)	**159** (41.7%)	**54** (28.4%)	**45** (10.5%)	**12** (34.3%)	**362** (20.7%)
	Any isoform > 20%	**2** (0.4%)	**32** (14.3%)	**59** (15.5%)	**25** (13.2%)	**9** (2.1%)	**9** (25.7%)	**136** (7.8%)
	Positions	185	87	147	80	166	14	679
	Measured	184 (99.5%)	87 (100%)	147 (100%)	77 (96.2%)	166 (100%)	14 (100%)	675 (99.4%)
	Any isoform > 5%	**92** (50%)	**67** (77%)	**134** (91.2%)	**64** (83.1%)	**99** (59.6%)	**13** (92.9%)	**469** (69.5%)
	Any isoform > 10%	**25** (13.6%)	**42** (48.3%)	**97** (66%)	**33** (42.9%)	**39** (23.5%)	**7** (50%)	**243** (36%)
	Any isoform > 20%	**2** (1.1%)	**18** (20.7%)	**45** (30.6%)	**16** (20.8%)	**9** (5.4%)	**4** (28.6%)	**94** (13.9%)

3.7 Distribution of splicing-effective mutations related to evolutionary conservation

In order to investigate how natural selection shaped the splicing-regulatory landscape of *RON* exon 11, evolutionary conservation was compared between splicing-effective- and non-effective positions (**Figure 20**). Within introns, splicing-effective positions differed by greater evolutionary conservation from positions that do not affect splicing, hence demonstrating evolutionary selection pressure towards maintenance of splicing-effective positions (Xing & Lee, 2006). In contrast, greater conservation of splicing-effective positions within exons was not observed, reflecting that selection pressure generated by the splicing-function of a nucleotide is overruled by protein-coding constraints. Notably, intronic conservation increases with the increasing splicing impact of positions such that strongest splicing-effective positions, including core-splicing elements, display similar conservation as exonic positions.

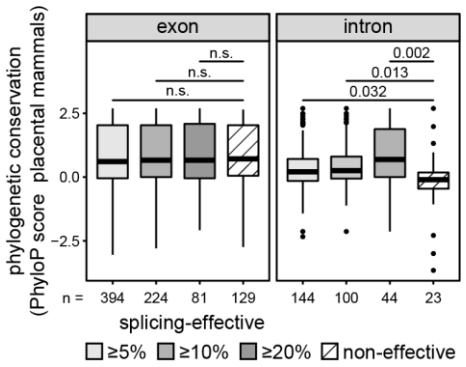


Figure 20: Intronic splicing effective positions are more conserved than non-effective positions in MCF7 cells. Splicing effective positions within introns but not in exons are significantly more conserved (PhyloP score across 46 placental mammals) compared to splicing non-effective positions. Splicing-effective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

Besides protein functional constraints, evolutionary selection of exonic nucleotide compositions is furthermore affected by the pressure to maintain splicing patterns (Warnecke et al., 2009). This is particularly evident for *RON*, since the majority of positions in exon 11

mediate splicing regulation (**Table 15**). To test the extent to which synonymous mutations are implicated in splicing regulation, the mutation effects were compared between synonymous- and non-synonymous mutations (**Figure 21**). In-line with previous results for FAS/CD95 exon 6 (Julien et al., 2016), synonymous mutations in *RON* exons 10, 11, and 12 mediate splicing-regulation with similar effect sizes as non-synonymous mutations. Thus, demonstrating evolutionary pressure of co-evolution splicing regulation and preservation of the genetic code (Mueller et al., 2015). Moreover, the 135 splicing-effective mutations that are synonymous do not alter the RON protein sequence, yet they still may contribute to disease by changing the protein amount and -function via alternative splicing (Xing & Lee, 2005; Shabalina et al., 2013; **Table 16**).

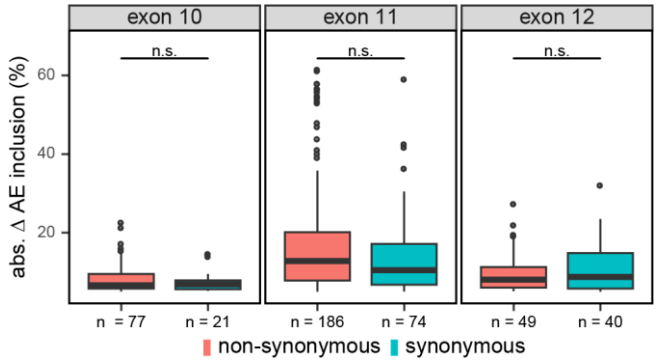


Figure 21: Synonymous and non-synonymous mutations display similar effect sizes with regard to altered AE inclusion in HEK293T cells. Boxplots show the distribution of absolute changes in AE inclusion in HEK293T cells for synonymous and non-synonymous mutations in exons 10-12. Number of mutations in each box indicated below. *p*-values correspond to two-sided Mann-Whitney-Wilcoxon test. n.s., not significant.

Table 16: Significant splicing-regulatory effects are observed with equal frequency among synonymous and non-synonymous mutations. Table summarizes the coincidence of significant splicing effects in HEK293T cells and synonymous/non-synonymous mutations across the three exons of the *RON* minigene.

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

Taken together, the alternative splicing of *RON* exon 11 is controlled by numerous intronic and exonic positions and the significance of splicing regulatory positions within introns is reflected by their evolutionary conservation. Moreover, splicing regulatory mutations frequently coincided with synonymous mutations. Since these mutations do not alter the sequence of the encoded protein, they might erroneously be interpreted as non-pathogenic, while in fact, they may impair protein function through alternative splicing.

3.8 Pathophysiological relevance of mutations in *RON*

Since *RON* is a proto-oncogene and skipping of exon 11 is relevant in cancer pathophysiology, mutation effects were quantified in the non-invasive human breast cancer cell line MCF7. Despite increased AE skipping levels observed for the minigene splicing in MCF7 compared to HEK293T cells (**Figure 8A**), the overall mutation effects were reproducible in terms of fold-change (**Figure 22**). Moreover, the number and distribution of splicing effective mutations and positions is remarkably similar between the two cell lines (**Table 15** and **Table 17**), indicating that regulation of *RON* splicing is largely consistent across different cell types.

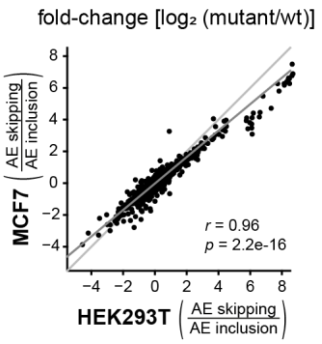


Figure 22: Mutation effects are conserved between MCF7 and HEK293T cells. Scatterplot compares changes in splice isoform ratios predicted for mutations assayed in MCF7 and HEK293T cells. Light and dark grey lines correspond to diagonal and linear regression line, respectively. *r*, Pearson correlation coefficient and associated *p*-value.

Table 17: Splicing effective mutations and positions per region in the *RON* minigene in MCF7 cells. Splicing-effective mutations (top) and positions (bottom) in HEK293T cells. The total number of possible and measured mutations/positions are indicated first, followed by the number of significant effects when considering any isoform at three isoform change cutoffs (>5%, >10% and >20%). Additionally, the significant mutations (>5%) are given for each individual isoform. AE - alternative exon; IR - intron retention.

		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 >5%	**150** (29.9%)	**149** (65.1%)	**300** (77.7%)	**137** (69.9%)	**264** (60%)	**22** (62.9%)	**1022** (57.2%)
	AE inclusion	81	115	260	99	91	16	662
	AE skipping	86	125	271	102	217	18	819
	First IR	5	14	5	3	6	0	33
	Second IR	1	2	12	7	15	11	48
	Full IR	79	63	62	82	185	16	487
	Other	3	2	8	14	13	0	40
	Any isoform > 10%	**41** (8.2%)	**88** (38.4%)	**202** (52.3%)	**76** (38.8%)	**100** (22.7%)	**14** (40%)	**521** (29.2%)
	Any isoform > 20%	**6** (1.2%)	**39** (17%)	**86** (22.3%)	**32** (16.3%)	**16** (3.6%)	**10** (28.6%)	**189** (10.6%)
	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 > 5%	**108** (58.4%)	**74** (85.1%)	**139** (94.6%)	**70** (89.7%)	**147** (88.6%)	**12** (85.7%)	**550** (81.2%)
	Any isoform > 10%	**36** (19.5%)	**52** (59.8%)	**112** (76.2%)	**48** (61.5%)	**74** (44.6%)	**8** (57.1%)	**330** (48.7%)
	Any isoform > 20%	**6** (3.2%)	**22** (25.3%)	**61** (41.5%)	**22** (28.2%)	**14** (8.4%)	**5** (35.7%)	**130** (19.2%)

The pathophysiological relevance of the mutations was next assessed by analysis of cancer patient data. Initially, the Catalogue of Somatic Mutations in Cancer (COSMIC) database was screened for mutations within the *RON* minigene region. The database contains manually curated somatic mutations from peer-reviewed publications and databases. The region of the *RON* minigene contained 33 COSMIC entries, of which 20 coincided with splicing-effective mutations (**Figure 23**). Notably, seven of these were synonymous with respect to the encoded RON protein, suggesting that their role in cancer arises from their impact on alternative splicing regulation rather than their protein-coding function.

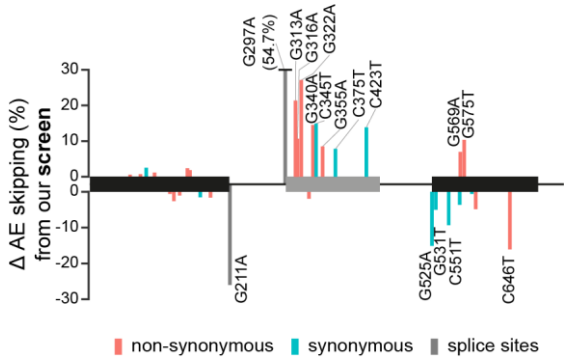


Figure 23: *RON* somatic mutations in cancer are enriched for splicing-effective mutations. Bar diagram shows the AE skipping changes quantified from the screening in MCF7 cells for 33 mutations from the COSMIC database available for the *RON* minigene region. Mutation identity is labelled for splicing-effective mutations. Orange, blue and grey indicate non-synonymous, synonymous and splice site mutations, respectively.

In order to further investigate the role of mutations affecting *RON* splicing in cancer, patient data from The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/) was analyzed next. In total, the database contained 51 mutations in the *RON* minigene region from 19 different cohorts (representing different cancer types) that occur specifically in the tumor- but not in the corresponding normal samples (**Table 18**). In contrast to the COSMIC database, TCGA provides additional RNA-seq data (i.e. splicing information) associated with the mutation information, thus enabling the comparison of *in vivo* mutation effects with the mutagenesis screening quantifications. To this end, the change of AE skipping levels between mutant-bearing tumor and matched normal tissue were compared with the AE skipping change of respective mutations quantified from the mutagenesis screening (**Figure 24A**). The observed

correlation (Pearson correlation coefficient, r = 0.62, p-value = 4.8e-05) indicated that the screening allows estimation of strong *in vivo* mutation effects in cancer.

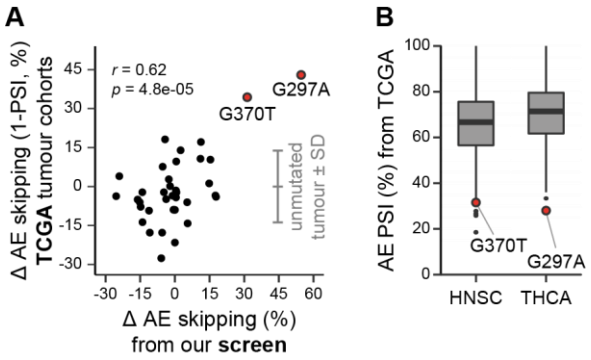


Figure 24: *RON* splicing in mutation-bearing cancer patients correlates with mutation effect quantifications from the screen. (**A**) Scatterplot comparing AE skipping changes of mutations found in TCGA database with screening derived quantifications. Highlighted mutations are detailed in (**B**). Grey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

The strongest mutation effects were caused by G370T and G297A mutations in Head-Neck Squamous Cell Carcinoma and Thyroid Carcinoma, respectively (**Figure 24B**). Both mutations induced elevated AE skipping levels, either via disrupting the AE 3' splice site (G297A) or by alteration of a putative ESE in *RON* exon 11 (G370T). Taken together, the mutagenesis screening allows evaluating the pathophysiological relevance of cancer mutations and suggests that implications of non-synonymous mutations in cancer may not only arise from amino acid alterations but also through splicing changes.

Table 18: Mutation effects on *RON* exon 11 splicing in cancer patients. Information on 51 mutations that are present in tumors 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 tumor 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). Only mutations from cohorts with more than 24 supporting reads on average were used in **Figure 24A** (highlighted in blue). 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.

mutation	genomic position (hg19)	cohort	total patients	patients with mutation	average supporting reads	ΔAE skipping (1-PSI; TCGA)	ΔAE skipping (screening)
G43T	chr3:49933795	BLCA	242	1	27	-17,76	-5,52
G56A	chr3:49933782	COAD	281	1	77	9,64	0,58
G56T	chr3:49933782	SKCM	44	1	26	-6,06	5,32
G76T	chr3:49933762	CESC	247	1	182	-1,52	0,45
G81A	chr3:49933757	BLCA	242	1	40	13,96	2,48
G103T	chr3:49933735	HNSC	423	1	19	-	-
A118T	chr3:49933720	OV	175	1	15	-	-
C119A	chr3:49933719	ESCA	172	1	298	-17,75	-10,85
C124T	chr3:49933714	LUAD	444	1	138	2,78	-2,62
C133T	chr3:49933705	STAD	391	1	39	-3,52	-1,02
C144A	chr3:49933694	OV	178	1	24	-	-
G221T	chr3:49933617	HNSC	423	1	22	-	-
G222A	chr3:49933616	BRCA	778	1	18	-	-
G222T	chr3:49933616	COAD	281	1	88	-21,58	0,02
G242T	chr3:49933596	THYM	49	1	16	-	-
A246G	chr3:49933592	STAD	391	1	96	-3,45	17,55
T277G	chr3:49933561	BLCA	242	6	46	5,74	14,85
T277G	chr3:49933561	PAAD	159	1	63	-10,56	14,85
T277G	chr3:49933561	OV	175	8	30	3,02	14,85
T277G	chr3:49933561	BRCA	739	8	27	-3,45	14,85
T277G	chr3:49933561	LUSC	268	9	35	-2,96	14,85
T277G	chr3:49933561	CESC	247	10	47	4,12	14,85
T277G	chr3:49933561	KIRC	5	1	10	-	-
T277G	chr3:49933561	KIRP	45	1	12	-	-
T277G	chr3:49933561	THCA	271	12	23	-	-

Table continues on the next page.

mutation	genomic position (hg19)	cohort	total patients	patients with mutation	average supporting reads	ΔAE skipping (1-PSI; TCGA)	ΔAE skipping (screening)
T277G	chr3:49933561	LUAD	444	15	52	4,94	14,85
T277G	chr3:49933561	STAD	391	4	128	4,12	14,85
T277G	chr3:49933561	HNSC	423	18	45	-1,72	14,85
C278G	chr3:49933560	STAD	391	1	99	10,70	11,08
C287T	chr3:49933551	SKCM	44	1	12	-	-
G297A	chr3:49933541	THCA	271	1	25	42,96	54,65
C345A	chr3:49933493	SKCM	44	2	27	-3,98	17,82
G348T	chr3:49933490	HNSC	423	1	21	-	-
G370T	chr3:49933468	HNSC	423	1	75	34,37	31,32
C375A	chr3:49933463	CESC	247	1	116	10,35	15,38
C376A	chr3:49933462	OV	178	1	24	-	-
G381A	chr3:49933457	CESC	247	1	16	-	-
C398A	chr3:49933440	HNSC	423	1	35	-2,20	-4,75
C398A	chr3:49933440	SKCM	44	1	16	-	-
C403T	chr3:49933435	LUSC	268	1	44	7,62	-5,28
C406A	chr3:49933432	SKCM	44	1	16	-	-
C411A	chr3:49933427	HNSC	423	1	35	-2,20	-14,08
C437A	chr3:49933401	HNSC	423	1	45	-4,87	-3,78
G471A	chr3:49933367	DLBC	3	1	57	17,14	11,28
C478T	chr3:49933360	BRCA	739	1	31	-17,15	0,02
C478T	chr3:49933360	COAD	281	1	66	-0,85	0,02
G479T	chr3:49933359	HNSC	423	1	24	-	-
G479T	chr3:49933359	SKCM	44	1	12	-	-
G499T	chr3:49933339	BLCA	242	1	45	0,14	-2,08
G500T	chr3:49933338	LUSC	268	1	55	-14,19	5,52
A547T	chr3:49933291	UCEC	61	1	22	-	-
A549G	chr3:49933289	CESC	247	1	43	-2,26	0,68
G568A	chr3:49933270	BLCA	242	1	12	-	-
G579T	chr3:49933259	BLCA	242	1	36	18,15	-4,25
G582A	chr3:49933256	COAD	281	1	32	-13,73	-14,02
G582T	chr3:49933256	THCA	271	1	18	-	-

Table continues on the next page.

mutation	genomic position (hg19)	cohort	total patients	patients with mutation	average supporting reads	ΔAE skipping (1-PSI; TCGA)	ΔAE skipping (screening)
C587A	chr3:49933251	HNSC	423	1	19	-	-
G599A	chr3:49933239	HNSC	423	1	15	-	-
C602T	chr3:49933236	COAD	281	1	47	-8,94	-0,58
C615A	chr3:49933223	BRCA	739	1	20	-	-
G636T	chr3:49933202	BLCA	242	1	23	-	-
C640G	chr3:49933198	THCA	271	1	25	-29,04	-14,82
C640G	chr3:49933198	HNSC	423	1	40	13,42	-14,82
C646A	chr3:49933192	BRCA	739	1	56	-27,60	-5,88
C646A	chr3:49933192	LIHC	24	1	21	-	-
C646T	chr3:49933192	CESC	247	1	88	-5,02	-16,12
G656T	chr3:49933182	BRCA	739	1	39	-4,21	0,48
G676T	chr3:49933162	SKCM	44	1	25	-9,29	-11,18
G690T	chr3:49933148	SKCM	44	1	26	-6,06	-15,18
T692C	chr3:49933146	COAD	281	1	60	-3,73	-25,42
A694G	chr3:49933144	COAD	281	1	134	3,97	-24,05

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

In total, the *RON* minigene contained 1,022 splicing-effective mutations in MCF7 cells (**Table 17**), and hence, multiple *cis*-regulatory elements are putatively located across the minigene. To ask which *trans*-acting factors target these *cis*-regulatory elements and are therefore involved in the regulation of *RON* splicing, multiple approaches were combined:

i) individual KD of target candidate genes followed by analysis of *RON* splicing using RT-PCR

ii) correlation analysis of RBP expression and *RON* splicing from cancer patients

iii) RNA-pulldown coupled mass-spectrometry for *de novo* identification of putative regulators

iv) characterization of a target protein interactome

v) combination of the screen results with *in silico* binding site predictions

These approaches are detailed in the following sections.

3.9.1 Knockdown of *trans*-acting factors and analysis of *RON* splicing via RT-PCR

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, use was made of a previously published, large-scale RBP KD screen (Papasaikas et al., 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 25A**) and successful KDs were confirmed by RT-qPCR analysis (**Figure 25B**). The directionality of the observed splicing changes were consistent with the results of Papasaikas et al., 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 Papasaikas et al. 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.

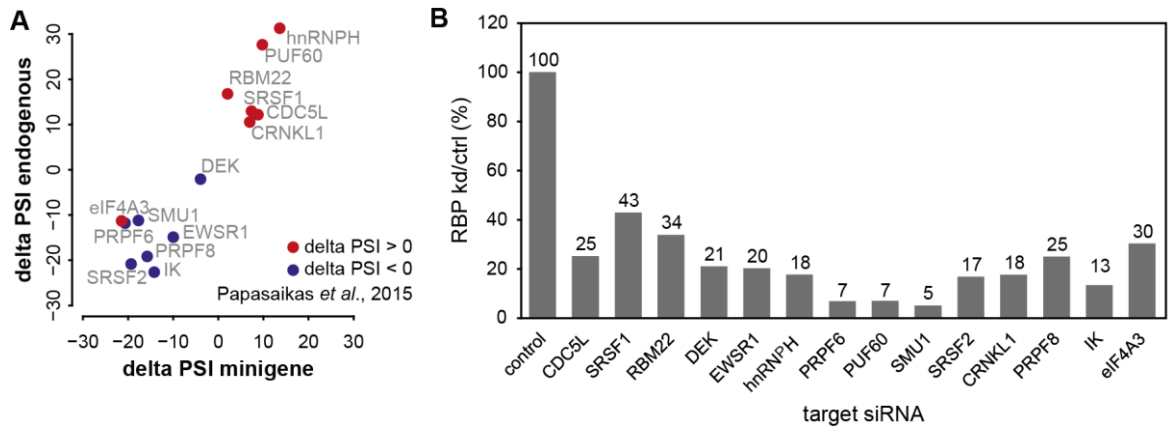


Figure 25: Several splicing activators and –repressors regulate *RON* exon 11 splicing. (**A**) Scatterplot comparing splicing changes induced by KD of selected RBPs between endogenous *RON* and *RON* minigene splicing. Red and blue color codes for positive and negative splicing changes measured by (Papasaikas et al., 2015), respectively. PSI percent-spliced-in. (**B**) Corresponding KD efficiencies analyzed by RT-qPCR.

3.9.2 Association of RBP expression and *RON* splicing in cancer patients

In order to identify novel regulators of *RON* splicing that were not considered in the previous RBP KD screen (Papasaikas et al., 2015), *in vivo* RBP expression and *RON* splicing was analyzed from TCGA. Based on the idea that the expression levels of *RON* interacting RBPs are correlated with *RON* splicing in tumor samples, the expression of 190 RBPs and corresponding *RON* AE PSI levels were analyzed in 4,514 cancer patient samples (**Table 19**). The strongest association between RBP expression levels and *RON* AE PSI was observed for *HNRNPH2* (**Figure 26**), which upon KD also induced substantial splicing changes in the aforementioned KD screening (**Table 19**). To conclude, *RON* splicing is regulated by numerous RBPs, with *HNRNPH2* as the strongest splicing determinant *in vivo*.

Table 19: List of top ten RBPs whose expression levels show strongest association with *RON* AE PSI levels in tumor samples from TCGA. The Spearman correlation coefficients of transcript expression levels of the top five positively and top five negatively correlating RBPs are shown from 190 RBPs across 4,514 tumor samples (TCGA) with *RON* exon 11 splicing along with significance of the correlations (*p*-value) and false discovery rates (FDR). Regression slopes correspond to linear regression between RBP expression and *RON* exon 11 percent spliced-in (PSI) in TCGA tumor samples. RBPs which upon KD induce substantial effects on *RON* exon 11 splicing (|z-score| > 1.5) as measured in (Papasaikas et al., 2015) are highlighted in blue.

gene	ensembl ID	Spearman correlation	regression slope	*p*-value	FDR
CDK10	ENSG00000185324	0,211	0,036	9,09E-47	8,72E-45
NXF1	ENSG00000162231	0,190	0,029	6,62E-38	4,24E-36
SNRNP70	ENSG00000104852	0,176	0,029	1,21E-32	4,65E-31
PRPF31	ENSG00000105618	0,170	0,029	1,31E-30	4,18E-29
PAXBP1	ENSG00000159086	0,162	0,026	7,29E-28	1,75E-26
PRPF4	ENSG00000136875	-0,154	-0,024	2,45E-25	3,91E-24
HNRNPF	ENSG00000169813	-0,154	-0,026	1,66E-25	2,89E-24
DHX8	ENSG00000067596	-0,165	-0,027	4,47E-29	1,23E-27
SLU7	ENSG00000164609	-0,178	-0,029	1,29E-33	6,19E-32
HNRNPH2	ENSG00000126945	-0,266	-0,044	1,03E-73	1,98E-71

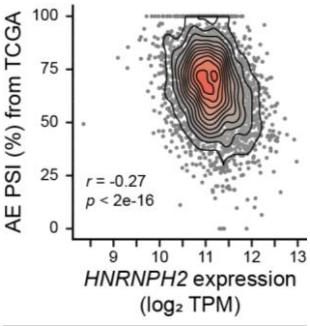


Figure 26: *HNRNPH2* expression correlates with *RON* AE PSI levels across TCGA tumour samples. Density scatter plot shows *HNRNPH2* expression (in transcripts per million, TPM) and *RON* exon 11 PSI across all TCGA tumour samples. *r*, Spearman correlation coefficient and associated *p*-value.

3.9.3 *De novo* detection of putative *RON* splicing regulators by RNA-pulldown coupled mass-spectrometry

In order to complement the list of putative regulators of *RON* splicing (**Figure 25A** and **Table 19**), RNA-pulldown coupled mass-spectrometry analyzes were performed using pairs of

in vitro transcribed RNA oligonucleotides. Each oligonucleotide pair was composed of stretches of 59-nt *RON* sequence with either wt sequence or a central point mutation that was shown by the screening to induce strong splicing changes. Interacting proteins were identified by stable isotope labelling using amino acids in cell culture (SILAC) based mass-spectrometry, which detected specific interactors of either wt or mutant RNA and furthermore allowed identification of previously unknown regulators of *RON* splicing.

Using wt/ mutant pairs of the point mutations G228T and G305A that in MCF7 cells increase AE skipping by 14% and decrease AE skipping by 20%, respectively, revealed the DEAH-Box helicase DHX36 as interactor of both wt RNAs and the DEAH-Box helicase DHX9 enriched for the wt RNA over G305A mutant (**Figure 27A**).

In order to confirm a regulatory role of DHX9 and DHX36 in *RON* splicing, siRNA-mediated KD of these factors was carried out in MCF7 cells and *RON* minigene splicing was subsequently analyzed by RT-PCR. However, neither *DHX9* and *DHX36* single- or *DHX9*/*DHX36* double KD affected *RON* splicing (**Figure 27B** and **27C**). For comparison, the effect of *HNRNPH* KD on *RON* splicing was measured, since HNRNPH was previously shown to regulate *RON* splicing (Lefave et al., 2011) and it was affecting splicing strongest in the tested set of RBPs (**Figure 25**). In accordance with the splicing shift observed previously in the set of 14 selected RBPs, *HNRNPH* KD promoted increased AE inclusion (**Figure 27B**). To test whether the combined action of the helicases DHX9 and DHX36 and HNRNPH was required for splicing regulation of *RON*, a triple KD of *DHX9*, *DHX36*, and *HNRNPH* was performed. The splicing phenotype of *HNRNPH* single KD was partially reverted if *DHX9* and *DHX36* levels were also reduced in the triple-KD, suggesting that combined action of these factors might be required for *RON* splicing regulation (**Figure 27B** and **27C**).

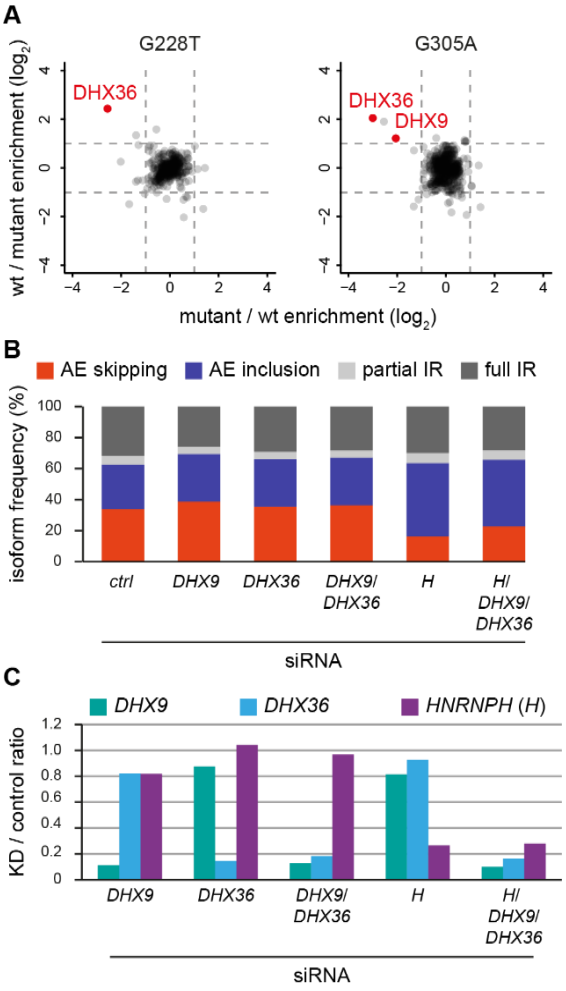


Figure 27: RNA-pulldown coupled mass-spectrometry identifies DEAH-Box helicases as putative interactors of *RON* minigene positions 228 and 305 and RT-PCR analyzes suggest they are involved in the regulation of *RON* splicing. (**A**) Scatterplots showing SILAC-pulldown results of RNA sequence stretches of 59 nucleotides from the respective region of the minigene either carrying the indicated central point mutation or not reveals DHX9 as interactor of position 305 and DHX36 as interactor of both positions 228 and 305. (**B**) Bar plot of RT-PCR results showing the isoform frequency resulting from KD of indicated factors in *RON* minigene-transfected MCF7 cells. *H* - *HNRNPH*. (**C**) Bar plot showing KD efficiencies of the experiment shown in (**B**) assessed by RT-qPCR quantification.

Additional SILAC pulldown experiments were performed using wt/ G272A mutant pair. Mutation G272A was found in the screen to decrease skipping by 25% in MCF7 cells and thus suggested the presence of a *cis*-regulatory element. The pulldown revealed specific interactions of HNRNPH1, HNRNPH2, and HNRNPF with the 59-nt *RON* wt RNA stretch, but not with the corresponding point mutant G272A *in vitro* (**Figure 28A**).

In order to validate these interactions in MCF7 cells, single and double KD of *HNRNPH* and *HNRNPF* were performed and the specificity of the KD was confirmed by Western Blot analysis (**Figure 28B**). Subsequent RT-PCR analysis of the *RON* wt minigene and G272A point mutation minigene revealed that *HNRNPH*- and *HNRNPF* KD exert opposing effects on splicing of the wt *RON* minigene, while the double KD recapitulates *HNRNPH* single KD. This

suggested a predominant role of HNRNPH over HNRNPF in the regulation of *RON* splicing (**Figure 28C** and **28D**). Splicing of the point mutant G272A was further altered upon KD of either *HNRNPH*, or *HNRNPF*, or both (**Figure 28C**), suggesting two possible explanations: (1) The *cis*-regulatory element at position 272 is not regulated by HNRNPH/F. (2) Alternatively, additional HNRNPH/F binding sites apart from position 272 exist in the *RON* minigene. Accordingly, splicing of a mutant of the putatively HNRNPH/F regulated site 272 is further altered upon *HNRNPH/F* KD, since these additional sites are still under regulation.

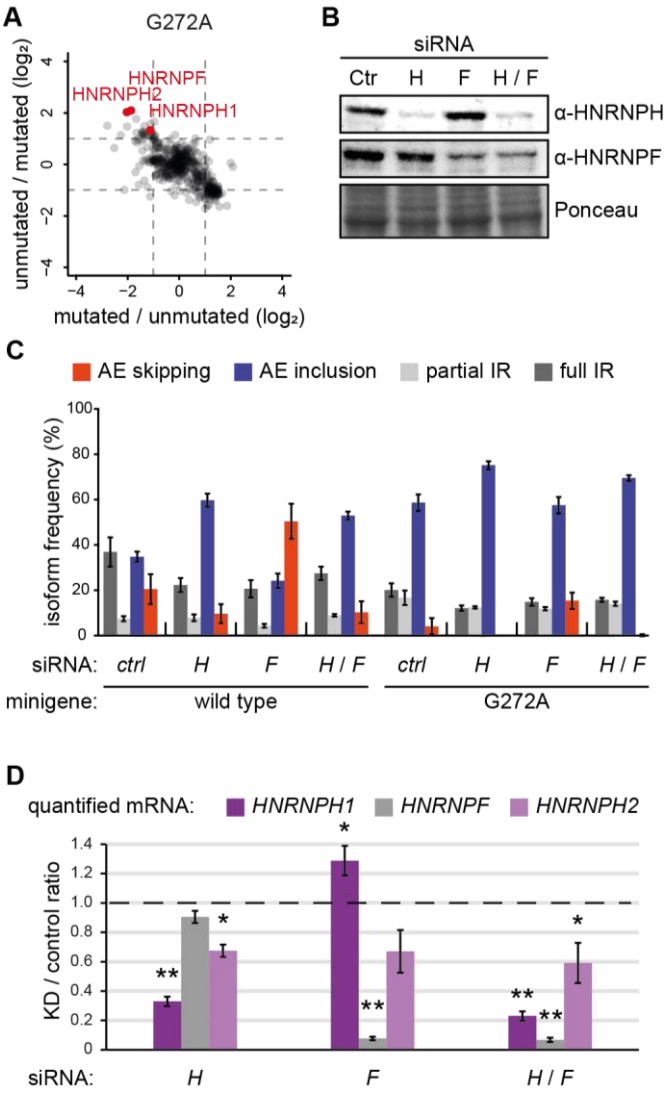


Figure 28: HNRNPH and HNRNPF putatively interact with position 272. (**A**) Scatterplot showing SILAC-pulldown results of RNA sequence stretch surrounding minigene position 272 either carrying a point mutation or not reveals HNRNPH1, HNRNPH2, and HNRNPF as interactors of the wt RNA. (**B**) Western Blot confirms KD specificity in MCF7 cells by using HNRNPH- and HNRNPF-specific antibodies. Ctr, H, F, and H / F lysates of control, HNRNPH, HNRNPF, and HNRNPH/HNRNPF siRNA-treated cells. Ponceau-staining used for loading control. (**C**) RT-PCR measurements of *RON* wt minigene and point mutant G272A in control (*ctrl*), *HNRNPH* (*H*), *HNRNPF* (*F*), or *HNRNPH* and *HNRNPF* (*H* / *F*) siRNA-treated MCF7 cells. (**D**) RT-qPCR measurement of KD efficiency for *HNRNPH1*, *HNRNPH2*, and *HNRNPF* in samples described in (**C**). * p-value < 0.05, ** p-value <0.001, unpaired two-tailed t-test. Error bars denote standard deviation of three independent replicates.

It should be noted though, that *HNRNPF* KD significantly elevates *HNRNPH1* mRNA levels (**Figure 28D**), which is also reflected by increased HNRNPH protein level in the *HNRNPF* KD (**Figure 28B**). Elevated HNRNPH1 levels may cause splicing changes that are erroneously attributed to the *HNRNPF* KD.

Together, DHX9 and DHX36 were identified as novel putative regulators of *RON* splicing and RNA-pulldown coupled mass-spectrometry data furthermore supports a role of HNRNPF, HNRNPH1, and HNRNPH2 as important factors involved in the regulation of *RON* splicing.

3.9.4 Characterization of the SMU1 interactome

The auxiliary spliceosomal protein SMU1 regulates *RON* splicing by activation of alternative exon inclusion (**Figure 25**), consistent with the presumed function of regulating alternative splice site selection in a small set of target pre-mRNAs (Sugaya et al., 2006; Fournier et al., 2014; Papasaikas et al., 2015). In order to gain further insight into the regulation of *RON* splicing by SMU1, the SMU1 interactome was characterized by co-immunoprecipitation of interacting proteins and subsequent mass-spectrometry analysis in two complementary experiments (**Figure 29A**).

In the 'co-IP + RNase experiment', RNA-independent interactions of SMU1 were detected by excluding RNA-dependent interactions through RNase digestion after the co-immunoprecipitation step (see Methods). Via this approach, the AAA ATPase VCP and the phosphatase PPP1R8 were identified as RNA-independent interactors of SMU1 (**Figure 29B**). Since splicing-related functions of these proteins are not yet described in the literature, these interactions might rather be driven by the high cellular abundance of VCP and PPP1R8 than by the interaction's specificity. Functional and physical coupling of the auxiliary spliceosomal protein IK and SMU1 has been described before (Ulrich et al., 2016). Since IK and SMU1 form RNA-independent dimers (Ulrich et al., 2016) but IK was not enriched in the 'co-IP + RNase experiment', a complementary 'mixing experiment' was performed. The setup of this co-immunoprecipitation experiment reduces transient interactors of the GFP-tagged SMU1 target

protein and recovers stable interactions between SMU1 and interacting proteins that have already been established in the cell (Hildebrandt et al., 2017). Indeed, the interaction of IK and SMU1 was recovered in the mixing experiment, while VCP and PPP1R8 were not enriched (**Figure 29C**), supporting the reliability of this dataset in contrast to the 'co-IP + RNase experiment'. In addition, several RBPs that were previously shown to regulate *RON* splicing were found in the SMU1-interactome (**Figure 29D**).

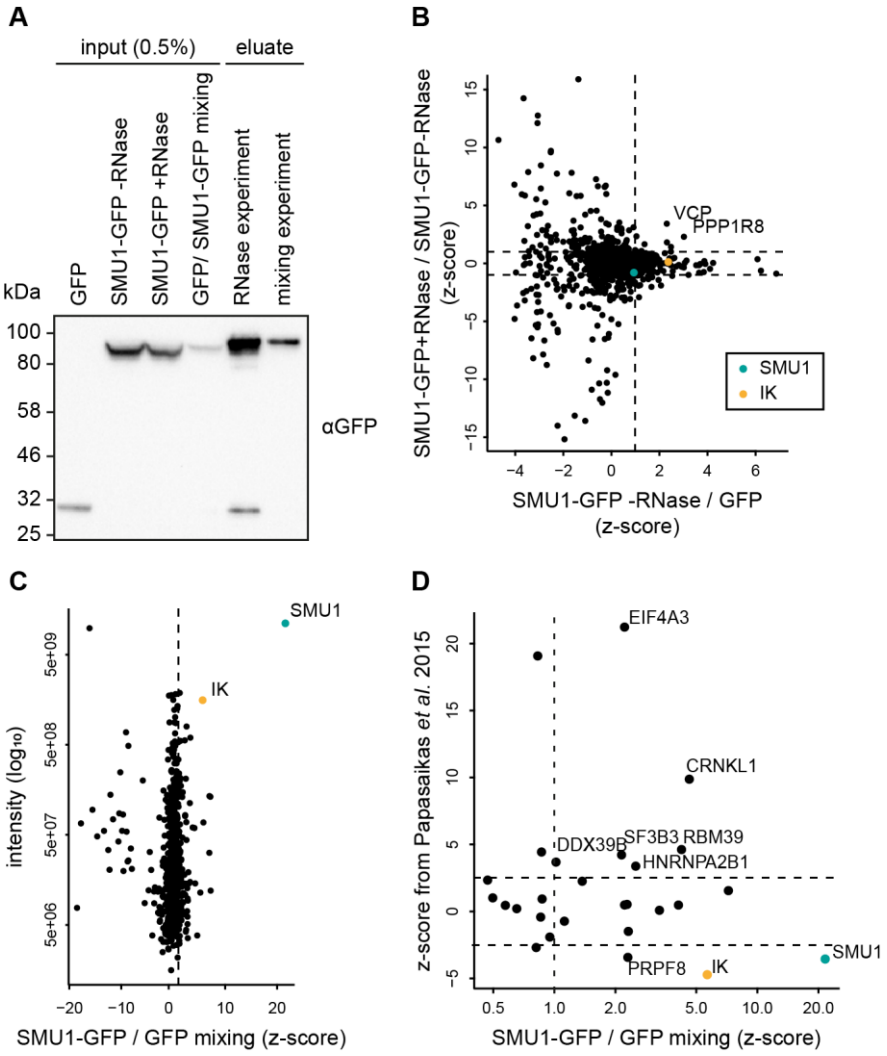


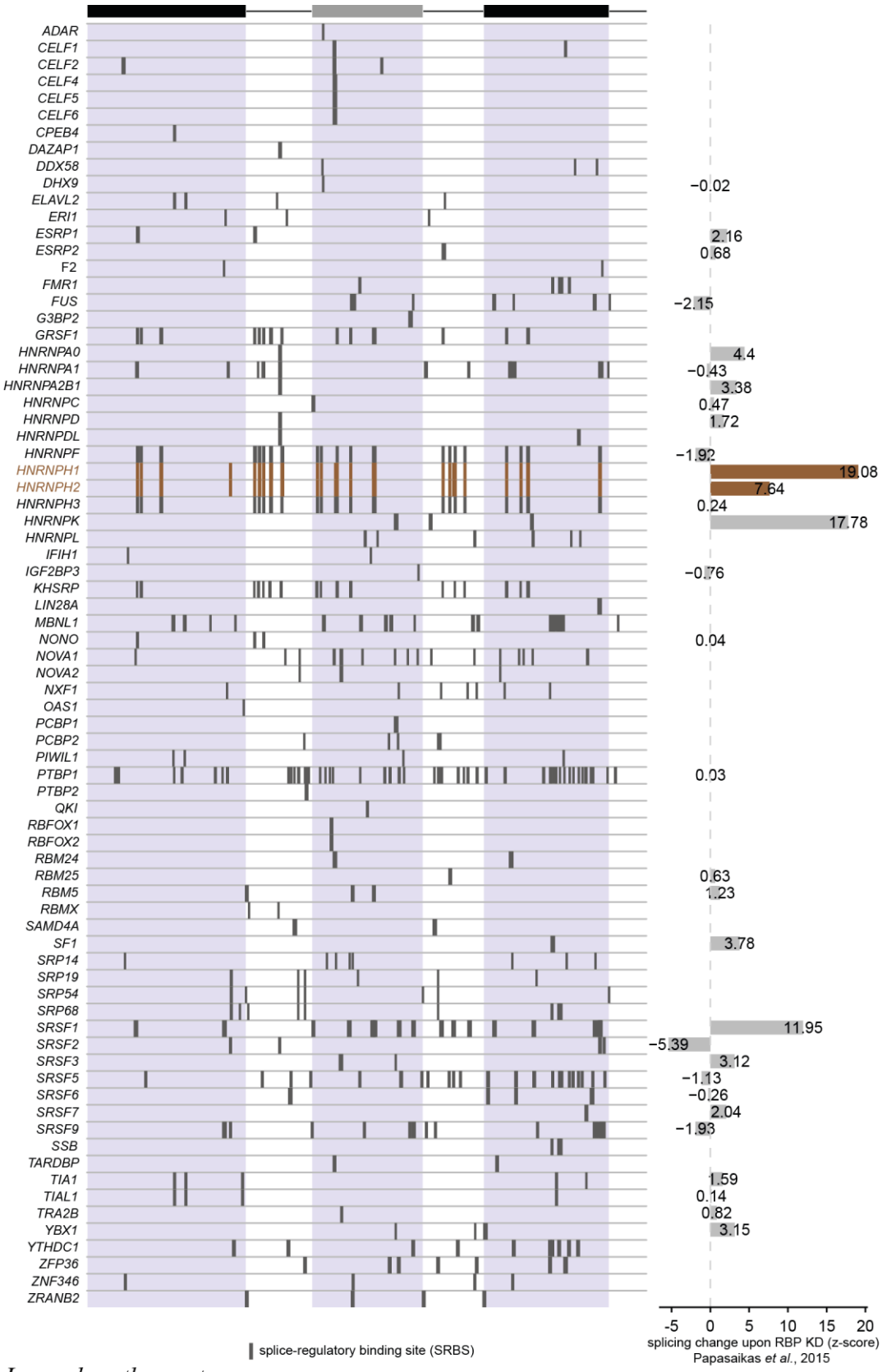
Figure 29: SMU1 interactome reveals factors that were previously shown to regulate *RON* splicing. (**A**) GFP and SMU1-GFP were expressed in HEK293T cells and co-immunoprecipitated with a GFP-specific antibody in a co-IP + RNase- and mixing experiment (see text). GFP-specific antibody was used for Western Blot analysis. (**B**)

Scatterplot showing RNase dependence of SMU1 interactions. SMU1-GFP versus GFP SILAC ratios are shown after z-score normalization (x-axis) against z-score normalized SILAC ratios of SMU1-GFP RNase treated versus SMU1-GFP non-RNase treated (y-axis). Dashed lines indicate z-score cut-offs of >|1| (y-axis) and >1 (x-axis). SMU1 interactor IK is shown in yellow. (**C**) Ratio-intensity plot of co-immunoprecipitated interactors of SMU1-GFP versus GFP. SMU1 interactor IK is shown in yellow. (**D**) Influence on *RON* splicing measured in (Papasaikas et al., 2015) (y-axis) against enrichment of interactors in SMU1-GFP versus GFP co-immunoprecipitation. Dashed lines indicate z-score cut-offs of >|2.5| (y-axis) and >1 (x-axis). SMU1 interactor IK is shown in yellow.

Taken together, multiple *RON* splicing regulators stably interact with each other, suggesting the presence of a regulatory network of *trans*-acting factors regulating *RON* splicing, including SMU1, IK, PRPF8 as splicing activators and EIF4A3, CRNKL1, DDX39B, SF3B3, RBM39, and HNRNPA2B1 as -repressors (Papasaikas et al. z-scores <0 and >0, respectively).

3.9.5 Combination of *in silico* binding site predictions with splicing quantification from the screen

In order to comprehensively collect a set of *trans*-acting factors that are potentially involved in *RON* splicing regulation, the ATtRACT database was used for *in silico* binding site predictions across the *RON* minigene (Giudice et al., 2016). To exploit the results from the screen and focus on sites that are actively involved in splicing regulation, RBP motif predictions were filtered for the presence of at least 60% splicing-effective positions (termed splice-regulatory binding sites, SRBS). The analysis revealed 76 putative regulators of *RON* splicing and recovered previously published binding sites of HNRNPH and SRSF1 (Ghigna et al., 2005; Lefave et al., 2011; **Figure 30**). The KD effect of a subset of 31 RBPs from the 76 potential regulators was previously tested in the RBP KD screen (Papasaikas et al., 2015). Notably, 17 RBPs substantially affected *RON* splicing with SRFS2 and HNRNPH as strongest activator and repressor, respectively. In sum, *RON* splicing is extensively regulated by multiple RBPs and HNRNPH is crucial for *RON* splicing both *in vivo* (**Figure 26**) and in cells (**Figure 25** and **Figure 30**).



Legend on the next page.

Figure 30: Putative *RON* splicing regulators identified by combination of *in silico* binding site predictions and splicing quantification from the screen. 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). Bar chart on the right provides the splicing change measured upon KD of the indicated RBPs from published data (z-scores; z-scores >1 indicate increased AE inclusion upon KD, z-scores <1 indicate decreased AE inclusion upon KD). SRBS for HNRNPH1 and HNRNPH2 are highlighted in brown.

3.10 Regulation of *RON* splicing by HNRNPH

In total, the *RON* minigene contains 22 SRBS for HNRNPH located in all transcript regions including both introns, the constitutive exons, and the alternative exon (**Figure 30**). These SRBS arrange into five clusters, each containing three to five SRBS. In order to confirm site-specific interactions of HNRNPH with the *RON* minigene, individual-nucleotide resolution cross-linking and immunoprecipitation (iCLIP) was performed in HEK293T cells (**Figure 31**).

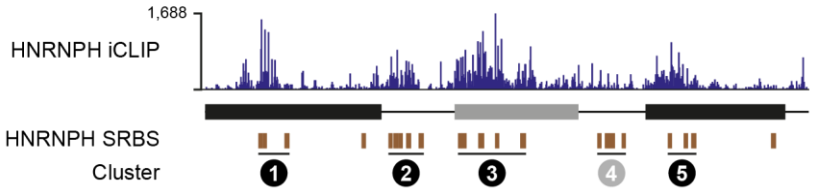


Figure 31: HNRNPH binds to multiple regions across the *RON* minigene. Binding of HNRNPH to the predicted HNRNPH SRBS is validated for four out of five HNRNPH SRBS clusters. Bar diagram shows HNRNPH iCLIP crosslinking events per position across the *RON* minigene (top). Brown boxes indicate HNRNPH SRBS that were assigned to five clusters (circled numbers; bottom).

Indeed, binding signal of HNRNPH was significantly enriched in four out of five HNRNPH SRBS clusters (**Figure 32**). Cluster 4 lacked significant iCLIP signal, suggesting that it is bound by a different RBP with a similar binding motif. The described HNRNPH binding motif is guanine-rich (Uren et al., 2016) and consequently, HNRNPH SRBS frequently harbor guanine-rich sequences (G-runs).

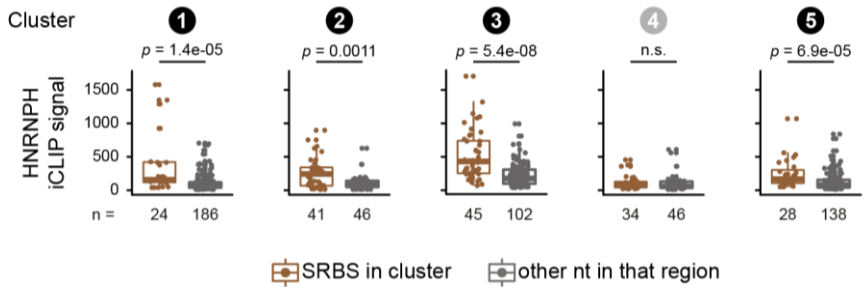


Figure 32: HNRNPH iCLIP signal is enriched in four out of five HNRNPH SRBS clusters. Boxplots compare the HNRNPH crosslink events within HNRNPH SRBS ±2 nt (brown) with HNRNPH crosslink events for other positions within the same intron/ exon region (grey). Number of positions per boxplot is given below. *p*-values correspond to two-sided Wilcoxon Rank-Sum test.

Accordingly, splicing was altered upon disruption or extension of G-runs by an additional guanine, corresponding to reduced or enhanced HNRNPH binding, respectively (**Figure 33**).

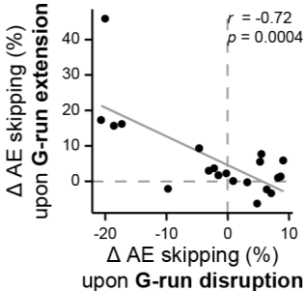


Figure 33: Opposite effects on splicing by G-run extension and –disruption. Scatterplot showing correlation of median AE skipping induced by mutations that extend or disrupt G-runs within HNRNPH SRBS. *r*, Spearman correlation-coefficient; *p*, associated *p*-value.

The directionality of HNRNPH-mediated splicing changes depended on the HNRNPH binding site position within the minigene. HNRNPH binding in the alternative exon mediated splicing repression, whereas binding of HNRNPH in the upstream intron activated splicing. Mutations in cluster 1 and 5, located in the up- and downstream constitutive exons, respectively, reduced intron retention and in cluster 5, this was accompanied by increased AE skipping (**Figure 34A**). Such context-dependent regulation has been described before for HNRNPH and other splicing regulators (Xiao et al., 2009; Katz et al., 2010). The strongest splicing effects were mediated by mutations in HNRNPH SRSBS cluster 3 that is located in the alternative exon (**Figure 34A**). Strikingly, *HNRNPH* KD induces a similar splicing pattern as mutations in cluster 3, underlining the significance of HNRNPH SRBS cluster 3 in the HNRNPH-mediated

regulation of *RON* splicing (**Figure 34B**). In conclusion, HNRNPH acts simultaneously as an activator and repressor of *RON* splicing and mediates complex regulation of different isoforms via multiple binding sites across the minigene region.

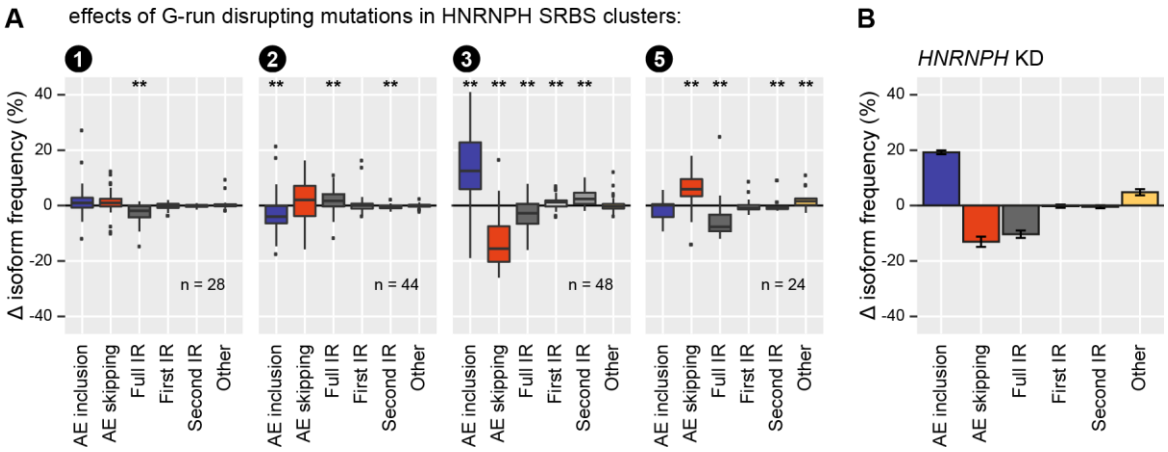


Figure 34: HNRNPH-mediated regulation of *RON* splicing depends on the HNRNPH binding site location. (**A**) Boxplots show the isoform frequencies resulting from G-run disrupting mutations within different HNRNPH SRBS clusters (circled numbers) in MCF7 cells (average of three biological replicates). Number of mutations for each cluster is given below. * 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.

3.11 Synergistic interactions between mutations and *HNRNPH* knockdown

Although HNRNPH pervasively binds across the entire minigene (**Figure 31**), there were qualitative and quantitative differences of mutation effects between different HNRNPH binding sites (**Figure 34**). This suggests that not all binding sites equally contribute to *RON* splicing regulation. In an effort to identify the functionally most relevant sites of HNRNPH-regulation, splicing of the minigene library was analyzed under *HNRNPH* KD conditions. Here, mutations that affect HNRNPH binding were expected to display positive- or negative synergy with the *HNRNPH* KD, i.e the combined effects of mutation and knockdown are greater or

smaller than expected from their independent contributions, respectively. For instance, mutations disrupting HNRNPH binding sites, would display reduced KD response compared to wt minigenes (negative synergy), while mutations strengthening HNRNPH binding sites would display enhanced KD response compared to wt minigenes (positive synergy).

In fact, *HNRNPH* KD globally promoted increased AE inclusion (**Figure 35**). In-line with the idea of synergy, subsets of minigenes displayed weaker or stronger KD responses. For instance, minigenes that harbor mutations within HNRNPH SRBS in cluster 3, such as G305A or G310A, were consistently less affected by the KD than the majority of minigenes with comparable ctrl AE inclusion. In contrast, minigenes with C307G mutations extend a pre-existing HNRNPH binding motif by an additional guanine. The increased HNRNPH binding strength was reflected by the low ctrl AE inclusion of C307G-containing minigenes and concordant with positive synergy, the *HNRNPH* KD promoted a stronger-than-average increase of AE inclusion in these minigenes.

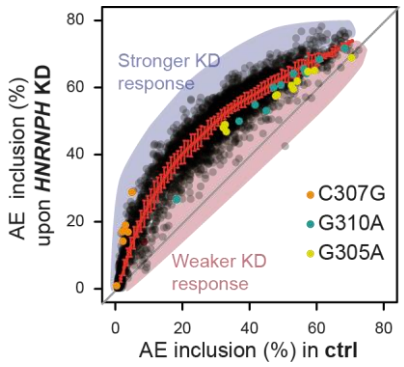


Figure 35: Minigenes in the library are differentially affected by *HNRNPH* KD. Scatterplot showing splicing change of minigenes upon *HNRNPH* KD compared to ctrl. Several minigenes are stronger (blue shade) or weaker (red shade) affected by the *HNRNPH* KD compared to the majority of minigenes (described by the running mean ± standard deviation, red line). Minigenes harboring indicated point mutations C307G, G310A, and G305A are highlighted in orange, blue, and yellow, respectively.

The synergy was quantitatively assessed for each mutation by calculating the difference of the mutation effects between ctrl and *HNRNPH* KD. Normalization of this difference by the experimental variation of the wt minigenes provided a z-score as a quantitative measure of the

synergy (**Figure 36**). If the z-score is close to zero, i.e. the mutation effect is the same between ctrl and *HNRNPH* KD, mutations and *HNRNPH* KD are not synergistic, while z-scores greater or smaller than zero indicate positive or negative synergy, respectively.

In total, 358 mutations in 281 positions (41%) displayed significant synergy for at least one splice isoform (|*z*-score|>2, adjusted *p*-value < 0.001, Stouffer's test). Synergistic interactions accumulated within the alternative exon and strikingly, 93% of positions within SRBS cluster 3 were synergistic (**Figure 37A**). Accordingly, the alternative exon represents the key region for HNRNPH-mediated splicing regulation of *RON*.

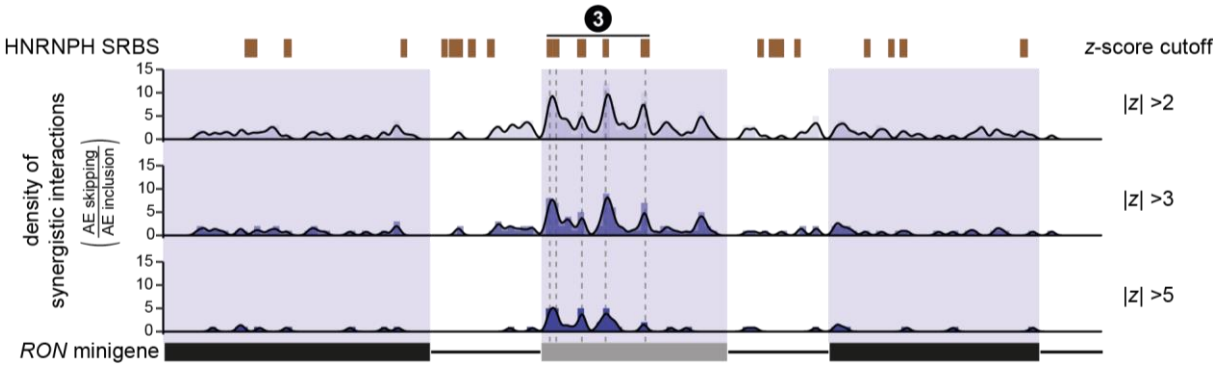


Figure 36: Synergistic interactions between mutations and *HNRNPH* KD overlap with HNRNPH SRBS in the alternative exon. Bar diagrams quantify significant synergistic interactions affecting AE skipping-to-inclusion isoform ratio using different z-score cutoffs in adjacent 5-nt windows. Line indicates density in 5-nt sliding window. Splice sites ± 2 nt were excluded. Predicted HNRNPH SRBS (brown) are given above.

The relevance of SRBS cluster 3 was validated by splicing quantification under control and *HNRNPH* KD conditions of ten minigene variants that each harbor point mutations in one of the five clusters (**Figure 37B**). Indeed, all four point mutations located in cluster 3 showed strong reduction of *HNRNPH* KD response. Splicing of minigenes harboring point mutations in other clusters was contrarily still altered upon *HNRNPH* KD, which agreed with the reduced synergy quantified for other clusters than cluster 3.

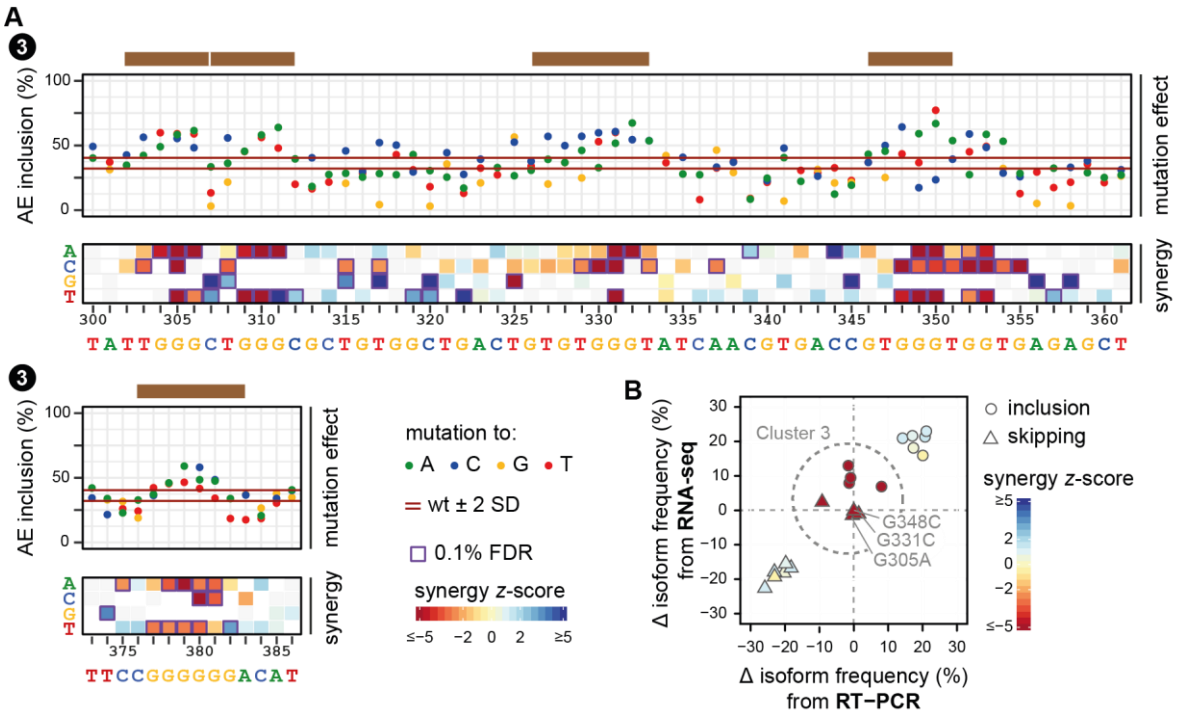


Figure 37: SRBS cluster 3 is most important for HNRNPH-mediated regulation of RON splicing. (**A**) Dot plots (top) display single mutation effects (inserted nucleobase, see legend) on AE inclusion (mean, n=3). Red lines indicate median isoform frequency of wt minigenes ± 2 standard deviations (SD). HNRNPH SRBS (brown) are given above. Heatmaps (bottom) show z-scores as measure of synergy (mean, n=3) per inserted nucleobase. 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 harboring 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.

Overall, KD followed by RNA-seq of a mutant minigene library with subsequent calculation of synergy between mutations and *HNRNPH* KD provided a quantitative measure that enabled evaluation of the functional importance of binding sites.

3.12 Cooperative HNRNPH binding establishes a splicing switch

The finding that single point mutations in cluster 3 are sufficient to almost completely release HNRNPH-mediated splicing repression (**Figure 37B**) suggested that HNRNPH binding occurs interdependent and mutations that compromise binding of HNRNPH also impair HNRNPH-binding at neighboring SRBS. In order to test this idea, iCLIP experiments were repeated in the context of minigenes harboring point mutations in cluster 3 (**Figure 38**). In-line with cooperative binding, the reduction of iCLIP signal was not limited to the site of the point mutation but extended across the entire alternative exon.

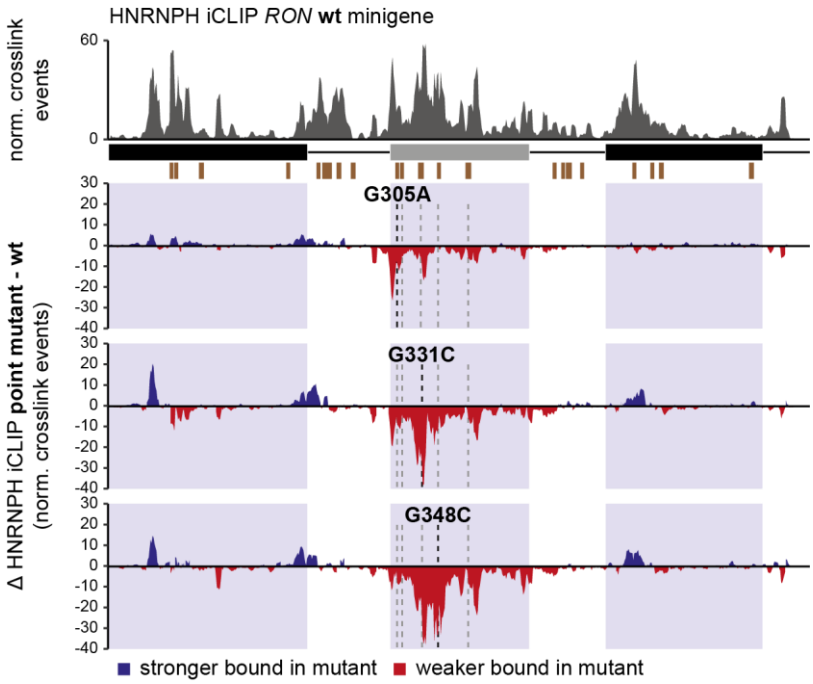


Figure 38: HNRNPH iCLIP signal reduction in point mutant minigenes extends towards neighboring HNRNPH binding sites. Bar diagram shows HNRNPH iCLIP normalized crosslinking events per position across the wt *RON* minigene in a sliding 5-nt window (top). Bar diagram showing the difference of normalized crosslinking events of indicated point mutants and the wt *RON* minigene in 5-nt sliding windows (bottom). Stronger or weaker 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.

Cooperative regulation of *RON* splicing by HNRNPH would imply that *RON* isoform levels are sensitive towards HNRNPH protein level alterations and hence describe a steep, sigmoidal dose-response curve. In order to test this, *RON* minigene and endogenous *RON* splicing was measured upon gradually changed HNRNPH levels achieved by *HNRNPH* KD- and *HNRNPH1* overexpression titration (**Figure 39A** and **39C-H**).

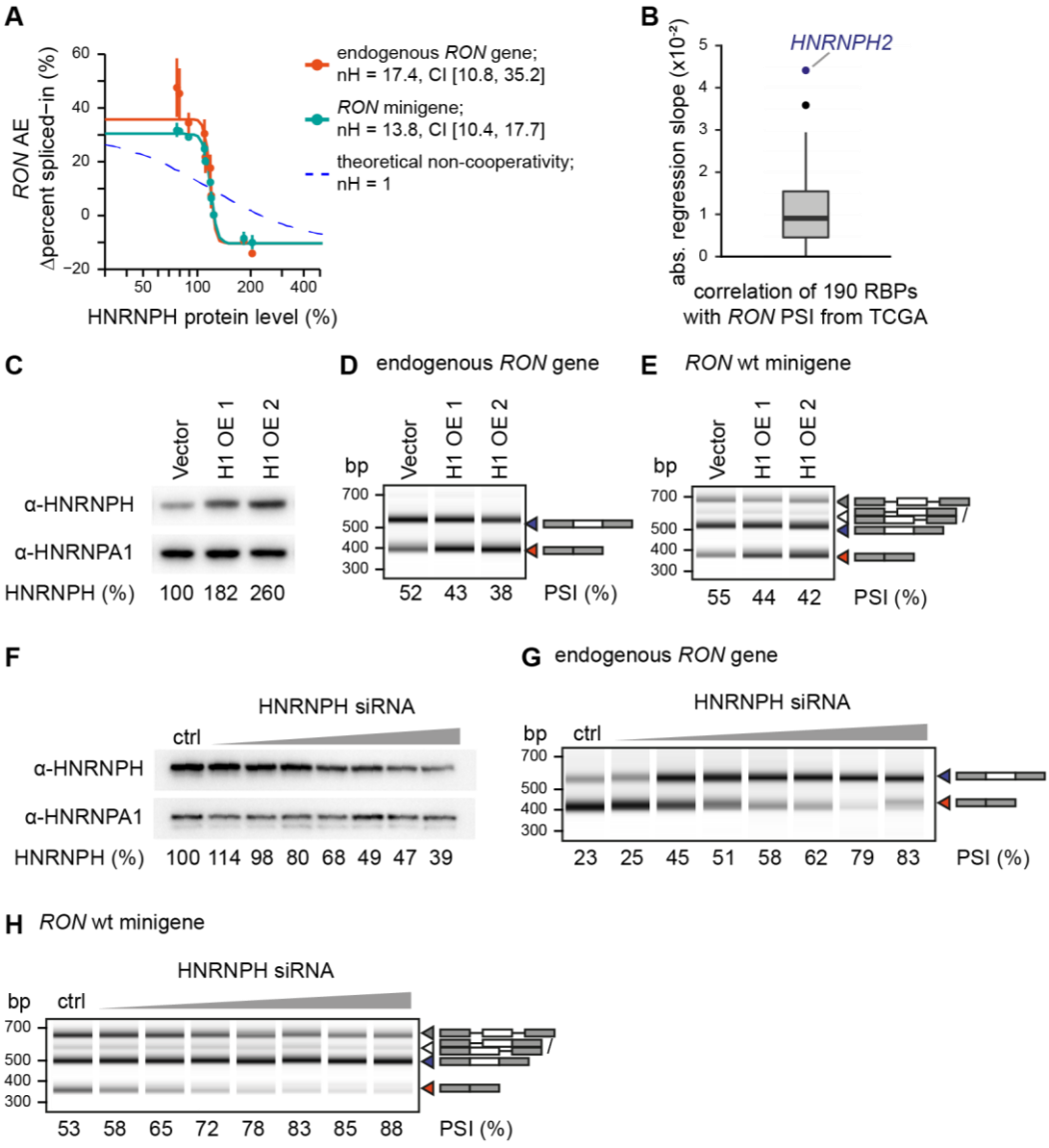


Figure 39: HNRNPH establishes a splicing switch of *RON* exon 11. (**A**) Dose-response curve showing the change of *RON* AE percent-spliced-in for differing HNRNPH protein levels for endogenous *RON* (orange) and the *RON* minigene (blue) from biological triplicates. Degree of cooperativity is quantified by fitting the Hill

equation (solid lines) and compared to the theoretical fit for non-cooperativity (dashed line). Error bars denote standard deviation. CI confidence intervals. (**B**) Boxplot shows the distribution of regression slopes from the correlation of 190 RBPs with *RON* splicing in cancer patients. *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. PSI percent-spliced-in. (**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.

Indeed, a switch-like splicing response of *RON* exon 11 from the minigene as well as the endogenous *RON* gene was observed for changing HNRNPH concentrations, which can be quantitatively evaluated using the Hill equation. Resulting Hill-coefficients of 13.8 and 17.4 for the *RON* minigene and the endogenous *RON*, respectively, indicated strong cooperativity. Moreover, *HNRNPH2* showed the steepest regression slope among the 190 RBPs tested for expression correlation in the TCGA data, which was consistent with cooperative regulation also *in vivo* (**Figure 39B**). Taken together, HNRNPH establishes a splicing switch of *RON* exon 11 through cooperative binding, which results in large splicing changes caused by small changes of HNRNPH protein level.

Discussion

In this PhD work, a novel high-throughput screening approach was established that enabled decoding the complete *cis*-regulatory landscape of *RON* exon 11 splicing. By systematically linking the effective mutations to the activity of the *trans*-acting factor HNRNPH, it extends beyond the sequence-centric view and takes a first step towards deciphering the complete regulatory network. Importantly, the results recapitulated the splicing regulation in cancer patients and provided mechanistic insights into the molecular mechanisms of *RON* alternative splicing.

4.1 Widespread occurrence of splicing-effective positions

Splicing regulation has long been believed to be mediated by discrete units of splicing enhancers and silencers that are serially arranged and estimations of their prevalence across pre-mRNAs are controversial (Savisaar & Hurst, 2017b). For instance, a 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 & Chasin, 2004). Other studies estimated that 30% - 60% of the sequence of human coding exons contains putative *cis*-regulatory elements (Parmley et al., 2006; Savisaar & Hurst, 2017a). In contrast, the current study showed that within *RON* exon 11, more than 90% of positions affect splicing and hence, splicing regulation is conferred by almost every nucleotide within the alternative exon. In this respect, *RON* exon 11 does not seem to be an exceptional case, as Lehner and colleagues presented a similar density of splicing-effective mutations in their mutagenesis study of *FAS*/*CD95* exon 6 (Julien et al., 2016). The gap between experimental

and computational estimates of the prevalence of splicing-effective positions has been discussed before (Savisaar & Hurst, 2017b). In a possible explanation for the discrepancy, the atypically short exons of the minigenes used in previous systematic mutagenesis studies were highlighted (87 bp, 54 bp, and 63 bp for *CFTR* exon 12, *SMN1* exon 7, and *FAS*/*CD95* exon 6, respectively; Pagani et al., 2005; Mueller et al., 2015; Julien et al., 2016). Hurst and colleague hypothesized that splicing regulatory information is more likely to be located at exon ends, while exon cores are depleted of splicing information. Human exons have a median length of 134 bp and thus, the previous minigene studies employing shorter exons were considered to overestimate the amount of splicing information, while the computational approaches included the full range of exon sizes (Savisaar & Hurst, 2017b). However, in the current study, investigation of the 147 bp *RON* exon 11 demonstrated that splicing information may densely spread also across exons of the length of an average human exon. In sum, it seems that splicing regulation is more complex than previously anticipated since it is mediated by the entire nucleotides of alternative exons rather than being scattered in individual *cis*-regulatory elements.

So far, previous mutagenesis studies focused on exons but splicing regulatory elements within introns have not been systematically studied, mostly because of technical limitations (Mueller et al., 2015; Julien et al., 2016; Ke et al., 2018). This study, however, enabled a systematic quantification of mutation effects in natural introns and showed that in addition to exons, also a large number of positions within the introns of the *RON* minigene mediate splicing regulation. Of note, though, are the 87 bp and 80 bp short introns up- and downstream of *RON* exon 11, respectively. In comparison to averaged-sized human introns of 3.3 kb (Lander et al., 2001), these are atypically short, which renders a more concentrated arrangement of splicing signals more likely.

Taken together, splicing of *RON* exon 11 is extensively regulated by numerous *cis*-regulatory elements and positively and negatively regulating sequences frequently overlap. This challenges the view of how splicing regulation is organized and explains why mutation effects are often difficult to predict (Grodecká et al., 2017). In fact, *in silico* prediction tools are

considered to be not useful in clinical diagnostics and instead laborious mutagenesis studies are required to better understand splicing regulation (Baralle & Buratti, 2017).

4.2 *RON* splicing is regulated by numerous *trans*-acting factors

In this work, multiple approaches were carried out in an effort to identify *trans*-acting factors that putatively regulate *RON* exon 11 splicing. As a targeted method, individual KD of selected *trans*-acting factors was carried out, which confirmed a previous high-throughput candidate screening (Papasaikas et al., 2015). While this approach did not allow *de novo* detection of putative regulators, it nevertheless provided a list of strong regulators. In contrast, the unbiased *in silico* prediction using the ATtRACT database or the RNA-pulldown coupled mass-spectrometry allowed identification of novel regulators, but both approaches required further validation. In addition, characterization of the SMU1 interactome revealed that several proteins that regulate *RON* splicing also physically interact with each other (**Figure 29**).

In sum, the combination of various approaches revealed that numerous *trans*-acting factors regulate *RON* splicing. In particular, the large number of 76 putative regulators coming from the ATtRACT database prediction raises the question of whether or not all of these *trans*-acting factors are actively involved in *RON* splicing regulation. In other words, if multiple *trans*-acting factors putatively share overlapping *cis*-regulatory elements, how are their cognate *cis*-regulatory elements selected in a given context? 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 & Ares, 2014). Accordingly, several putative regulators suggested in this work by the ATtRACT database analysis are cell type- or tissue-specifically expressed and therefore a given binding site prediction might be irrelevant under the experimental conditions. Moreover, several binding sites provided in the database were generated from *in vitro* based methods like SELEX and thus might be inaccurate *in vivo* (Wu & Kwon, 2016). Furthermore, it was previously shown that splicing factors only bind a subset of their natural binding motifs

present in the transcriptome and that their apparent binding is shaped by cofactors (Sutandy et al., 2018). Nevertheless, already the large number of validated regulators in the current study (**Figure 25**) highlights that *RON* splicing is extensively regulated by numerous *trans*-acting factors. Accordingly, Valcárcel and colleagues showed that splicing regulatory networks of functionally and physically associated splicing factors mediate alternative splicing regulation (Papasaikas et al., 2015).

Taken together, combining multiple approaches for the identification of putative splicing regulators yields extensive lists of candidate proteins that oftentimes require downstream validation. Furthermore, multiple *trans*-acting factors are involved in the regulation of *RON* exon 11 splicing, and several of these factors physically interact with each other. This suggests that *RON* splicing regulation is complex and regulated by a network of splicing factors rather than isolated interactions between *cis*-regulatory elements and individual *trans*-acting factors.

4.3 Clinical significance of the mutagenesis data

RON mRNA is abundantly expressed in many epithelial cells from most tissues (Gaudino et al., 1994; Gaudino et al., 1995). The expression levels are however very low (Wang et al., 1996), and oftentimes *RON* is not covered in publicly available, genome-wide sequencing data, which raises demand for alternative strategies to assess the disease relevance of single nucleotide variants. This information is crucial, since clinical management of cancer patients may be highly influenced by knowledge of mutation pathogenicity (Grodecká et al., 2017). For instance, monoclonal antibodies targeting RON to block MSP-RON signaling were in clinical trial (ClinicalTrials.gov Identifier: NCT01119456, (O'Toole et al., 2006)). However, since RONΔ165 is constitutively active and independent from MSP activation, RONΔ165-expressing tumors escape this kind of therapies (Chakedis et al., 2016). Therefore, identification of effective mutations in patients is of diagnostic value and may allow targeted therapy.

This study provided quantitative splicing information on ~1,800 mutations of which 778 and 1022 significantly affected splicing in HEK293T and MCF7 cells, respectively (**Table 15**). The observed correlation between mutation effects in cancer and the screening data (**Figure 24**) suggests that the screening can recapitulate strong *in vivo* splicing effects and may therefore aid the assessment of a mutation's pathogenicity. For instance, mutations G297A and G370T occurred in cancer patients and increased AE skipping on average by 43% and 34%, respectively (55% and 31% in the screening). Since G297A is a splice site mutation, its strong impact on increased AE skipping could have been already estimated without the information from the mutagenesis screen. In contrast, mutation G370T is a non-synonymous mutation that on the amino acid level introduces a premature termination codon (PTC) in the alternative exon. PTC-containing mRNAs were previously shown to be degraded via nonsense-mediated mRNA decay (Nicholson & Mühlemann, 2010). However, this work showed that in consequence of the mutation AE skipping is increased and therefore the alternative exon is not included in the mature mRNA. This inverts the physiological consequences of this mutation: instead of less functional receptor, the cells face increased RONΔ165 levels, which was previously shown to lead to constitutive receptor activation (Collesi et al., 1996; Zhou et al., 2003).

In this study, it was shown that splicing-effective mutations distribute proportionally between synonymous and non-synonymous mutations (**Table 16**). Due to the lack of comprehensive and quantitative splicing data, a recent study aiming to evaluate the selection of synonymous mutations in cancer evolution was restricted to essential splice site mutations when classifying mutations to exclude from their neutral background model (Martincorena et al., 2017). Moreover, synonymous mutations were often considered as silent mutations (Greenman et al., 2006) and their impact on altered protein isoforms through alternative splicing-induced in-frame deletions was frequently overlooked (Gotea et al., 2015). This study and others, however, highlighted the potentially deleterious impact of synonymous mutations in cancer (Gartner et al., 2013; Supek et al., 2014). In fact, it was previously proposed that loss of HNRNPH2 binding sites through synonymous mutations in oncogenes leads to tumor progression (Supek et al., 2014). In *RON* splicing, however, mutation-induced loss of HNRNPH binding sites exclusively in SRBS cluster 2 and 5 promoted increased levels of the pathogenic

AE skipping isoform, while global loss of HNRNPH led to reduced AE skipping instead (**Figure 34**). In sum, synonymous mutations can impair binding of HNRNPH and other splicing factors, which may lead to deleterious changes in alternative splicing.

Here, a splicing-effective position was defined by showing significant difference to the wt splicing and by additionally setting an arbitrary minimum of 5% change in isoform frequency. However, splicing effectiveness does not directly compare to functional relevance and it is not trivial to estimate the extent of aberrant splicing that can lead to a disease phenotype, e.g. tumor metastasis. In fact, the AE skipping levels in lower-stage tumors compared to metastatic tumors from TCGA were not generally higher (personal communication by Nuno Barbosa-Morais), demonstrating that besides *RON* splicing, other molecular determinants of cancer metastasis exist. At least in introns, the functional importance of the splicing-effective positions is implied by their increased evolutionary conservation (**Figure 20**). Accordingly, even the 5% change in isoform frequency mediates a phenotype that is visible for evolutionary selection, suggesting that mutations that were defined as splicing-effective are indeed functionally relevant.

In conclusion, the mutagenesis screening presented here provides quantitative splicing data for almost all possible mutations affecting *RON* splicing. The data may help to assess the disease relevance of mutations in clinical diagnostics and may additionally be used to train *in silico* splicing prediction tools in the future.

4.4 HNRNPH cooperatively regulates *RON* splicing

In this study, HNRNPH was shown to cooperatively regulate *RON* splicing. The presented evidence included that (i) mutating individual binding sites can resemble the full effect of *HNRNPH* KD on AE inclusion, (ii) single point mutations alleviate HNRNPH binding along the complete cluster 3 in iCLIP experiments, (iii) AE inclusion shows an amplified response to small changes in HNRNPH concentration, and (iv) *RON* exon 11 inclusion shows a particularly steep correlation with *HNRNPH2* expression levels across TCGA cancer samples.

This mode of regulation raises two questions: how is cooperativity achieved and what is the functional benefit of cooperative regulation? With regard to the former, it was previously shown that most HNRNPH pre-mRNA targets harbor multiple binding sites, which is a prerequisite for cooperative binding (Dominguez et al., 2010). Cooperativity might be mediated by oligomerization of HNRNPH via its glycine/tyrosine (GY)-rich domain (van Dusen et al., 2010). Indeed, other HNRNP proteins were recently shown to form multimeric assemblies via GY-rich domains to regulate splicing and it has been proposed that splicing factors tend to bind cooperatively (Akerman et al., 2009; Gueroussov et al., 2017). Moreover, it was previously suggested that the structural context of HNRNPH-regulated mRNA is crucial for HNRNPH function (Jablonski et al., 2008). Specifically, a G-quadruplex structure in *p53* pre-mRNA and a stem-loop structure in *H-Ras* pre-mRNA were shown to regulate splicing through interactions with HNRNPH/F and HNRNPH, respectively (Camats et al., 2008; Decorsière et al., 2011). Accordingly, the cooperative assembly of HNRNPH on *RON* exon 11 might allow remodeling of RNA secondary and tertiary structures, such as G-quadruplexes, to maintain them in single-stranded conformation. In fact, a two-state model of a stable G-quadruplex on the one hand and an open conformation mediated by an HNRNPH assembly on the other hand, would explain the observed switch-like splicing response upon altered HNRNPH levels (**Figure 39**).

However, literature is controversial about whether HNRNPH binds already folded G-quadruplexes or maintains otherwise G-quadruplex-forming RNA in single-stranded conformation. While Conlon et al. showed that G-quadruplexes are bound by HNRNPH, von Hacht et al. proposed that HNRNPH binds and dissolves G-quadruplexes (von Hacht et al., 2014; Conlon et al., 2016). Therefore, different modes of regulation seem to be target-dependent and increasing evidence highlights the importance of G-quadruplex-resolving helicases in HNRNPH-mediated splicing regulation. For instance, it was previously proposed that DHX36 unwinds G-quadruplexes to allow access of HNRNPH/F and similarly, it was previously shown that DDX5 and DDX17 interact with HNRNPH/F at G-quadruplex-forming sites to mediate splicing (Dardenne et al., 2014; Newman et al., 2017). Consistently, DHX9 and DHX36 were found in this study *in vitro* to bind to regions within *RON* that correspond to HNRNPH SRBS clusters 2 and 3, although DHX9 or DHX36 KD did not affect *RON* splicing

(**Figure 27**). Taken together, cooperativity could be achieved through homotypic HNRNPH assemblies on *RON* exon 11. Furthermore, 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 experimental evidence is required to validate this hypothesis.

After all, cooperative regulation of *RON* splicing translates into a switch-like splicing response, but what is the functional benefit of switch-like splicing over gradual splicing changes resulting from changed HNRNPH levels? In human embryonic development, a switch from rapid cell proliferation to organ development occurs between the fourth and ninth week post-fertilization (Yi et al., 2010). *HNRNPH2* was found amongst the 10% of the genes that are differentially regulated in exactly that time window (Yi et al., 2010). The downregulation of *HNRNPH2* reported in the study of Yi et al. would cause increased *RON* AE inclusion and lower RONΔ165 levels, which would in turn translate to reduced cell proliferation – in-line with the developmental switch in human embryogenesis. In a complementary study that measured transcriptomic profiles of human embryogenesis at earlier time points, *HNRNPH1*, but not *HNRNPH2*, was found significantly upregulated between the third and the fourth week post-fertilization (Fang et al., 2010). During this phase, the embryo might benefit from enhanced cell-proliferation activity induced by higher RONΔ165 levels, which are in turn caused by an upregulation of *HNRNPH1*. Cellular differentiation during embryonic development is associated with dramatic phenotypic changes. It is therefore tempting to speculate that a tightly controlled splicing switch better supports the rapid phenotypic transitions between proliferative- and stationary cell states than gradual splicing changes would do.

In conclusion, switch-like splicing of *RON* may enable precise modulation of the proliferation and migration activity of cells, which is important during human embryogenesis and might be controlled through *HNRNPH* gene expression changes in early development.

4.5 Outlook

The developed high-throughput screening approach can be applied to study other alternative splicing events than *RON*. While respective minigenes ideally fulfill the criteria mentioned in section 3.1 of this thesis, the length of the minigene reporter can be flexibly increased by using other sequencing techniques that allow for longer read lengths. For instance, the boundary for maximum minigene lengths of 1 kb set by the Illumina Platform can be pushed by at least one order of magnitude using the SMRT sequencing technology of Pacific Biosciences or nanopore sequencing by Oxford Nanopore Technologies (Eid et al., 2009; Clarke et al., 2009). While the latter technology is still considered premature for routine applications, SMRT sequencing already today offers a realistic long-read alternative to the Illumina technology (Ardui et al., 2018). In particular, these techniques may enable random mutagenesis studies of minigenes containing typical human introns of multiple kb length.

This study showed that *RON* splicing is regulated by multiple *trans*-acting factors. Since the synergy analysis conceptually illustrated that KD followed by RNA-seq of the minigene library 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 *RON* splicing regulatory network. This is, because KD of factors that act in the same molecular complex should display synergies with the same set of *cis*-regulatory elements.

In order to further investigate the cooperative splicing regulation mediated by HNRNPH, transcriptome-wide RNA-seq data generated from cells with gradually altered HNRNPH levels would allow comprehensive analysis of linear and switch-like splicing responses. By incorporating transcriptome-wide information about RNA G-quadruplex positioning, either by *in vitro* data or *in silico* predictions, cooperative HNRNPH splicing regulation could functionally be linked to these RNA structure elements (Bedrat et al., 2016; Kwok et al., 2016).

References

Ahsan, K.B., Masuda, A., Rahman, M.A., Takeda, J.-I., Nazim, M., Ohkawara, B., Ito, M., and 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., and Mandel-Gutfreund, Y. (2009). A computational approach for genome-wide mapping of splicing factor binding sites. Genome biology *10*, R30.

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., and Ladomery, M.R. (2011). WT1 mutants reveal SRPK1 to be a downstream angiogenesis target by altering VEGF splicing. Cancer cell *20*, 768–780.

Ardui, S., Ameur, A., Vermeesch, J.R., and 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., Oberstrass, F.C., and Allain, F.H.-T. (2006). Sequence-specific binding of single-stranded RNA: Is there a code for recognition? Nucleic acids research *34*, 4943–4959.

Baralle, D., and Buratti, E. (2017). RNA splicing in human disease and in the clinic. Clinical science (London, England 1979) *131*, 355–368.

Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nature reviews. Molecular cell biology *18*, 437–451.

Barash, Y., Vaquero-Garcia, J., González-Vallinas, J., Xiong, H.Y., Gao, W., Lee, L.J., and Frey, B.J. (2013). AVISPA: A web tool for the prediction and analysis of alternative splicing. Genome biology *14*, R114.

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., and Blencowe, B.J. (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science (New York, N.Y.) *338*, 1587–1593.

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.

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

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

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

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., and 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., and Blencowe, B.J. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. Genome research *24*, 1774–1786.

Buratti, E., Stuani, C., Prato, G. de, and 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., and 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*, e2926.

Cammas, A., and Millevoi, S. (2017). RNA G-quadruplexes: Emerging mechanisms in disease. Nucleic acids research *45*, 1584–1595.

Caputi, M., and 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.

Cartegni, L., and 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.

Chakedis, J., French, R., Babicky, M., Jaquish, D., Mose, E., Cheng, P., Holman, P., Howard, H., Miyamoto, J., Porras, P., Walterscheid, Z., Schultz-Fademrecht, C., Esdar, C., Schadt, O., Eickhoff, J., and Lowy, A.M. (2016). Characterization of RON protein isoforms in pancreatic cancer: Implications for biology and therapeutics. Oncotarget *7*, 45959–45975.

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

Chen, L., Bush, S.J., Tovar-Corona, J.M., Castillo-Morales, A., and 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.

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

Collesi, C., Santoro, M.M., Gaudino, G., and 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.

Conlon, E.G., Lu, L., Sharma, A., Yamazaki, T., Tang, T., Shneider, N.A., and 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*.

Conti, L. de, Baralle, M., and Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. Wiley interdisciplinary reviews. RNA *4*, 49–60.

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

Cordin, O., and Beggs, J.D. (2013). RNA helicases in splicing. RNA biology *10*, 83–95.

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., and Auboeuf, D. (2014). RNA helicases DDX5 and DDX17 dynamically orchestrate transcription, miRNA, and splicing programs in cell differentiation. Cell reports *7*, 1900–1913.

Decorsière, A., Cayrel, A., Vagner, S., and 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.

Ding, J., Hayashi, M.K., Zhang, Y., Manche, L., Krainer, A.R., and 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.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) *29*, 15–21.

Dominguez, C., Fisette, J.-F., Chabot, B., and 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.

Dorak, M.T. (Ed.) (2006). *Real-time PCR*. New York.

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

Eid, 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., and Schaal, H. (2013). Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. RNA (New York, N.Y.) *19*, 96–102.

Faham, N., and 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.

Fang, H., Yang, Y., Li, C., Fu, S., Yang, Z., Jin, G., Wang, K., Zhang, J., and Jin, Y. (2010). Transcriptome analysis of early organogenesis in human embryos. Developmental cell *19*, 174–184.

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

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

Fournier, G., Chiang, C., Munier, S., Tomoiu, A., Demeret, C., Vidalain, P.-O., Jacob, Y., and Naffakh, N. (2014). Recruitment of RED-SMU1 complex by Influenza A Virus RNA polymerase to control Viral mRNA splicing. PLoS pathogens *10*, e1004164.

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

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

Garg, K., and 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., Chen, G., Elnitski, L.L., Davies, M.A., Gershenwald, J.E., Carter, H., Karchin, R., Robinson, W., Robinson, S., Rosenberg, S.A., Collins, F.S., Parmigiani, G., Komar, A.A., Kimchi-Sarfaty, C., Hayward, N.K., Margulies, E.H., and Samuels, Y. (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., Avantaggiato, V., Follenzi, A., Acampora, D., Simeone, A., and Comoglio, P.M. (1995). The proto-oncogene RON is involved in development of epithelial, bone and neuro-endocrine tissues. Oncogene *11*, 2627–2637.

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

Ghigna, C., Giordano, S., Shen, H., Benvenuto, F., Castiglioni, F., Comoglio, P.M., Green, M.R., Riva, S., and 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., and Lara-Pezzi, E. (2016). ATtRACT-a database of RNA-binding proteins and associated motifs. Database the journal of biological databases and curation *2016*.

Gotea, V., Gartner, J.J., Qutob, N., Elnitski, L., and 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 (New York, N.Y.) *6*, 1197–1211.

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

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

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (New York, N.Y.) *348*, 648–660.

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

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

Hildebrandt, A., Alanis-Lobato, G., Voigt, A., Zarnack, K., Andrade-Navarro, M.A., Beli, P., and König, J. (2017). Interaction profiling of RNA-binding ubiquitin ligases reveals a link between posttranscriptional regulation and the ubiquitin system. Scientific reports *7*, 16582.

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

Hui, J., Stangl, K., Lane, W.S., and Bindereif, A. (2003). HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. Nature structural biology *10*, 33–37.

Ibrahim, E.C., Schaal, T.D., Hertel, K.J., Reed, R., and 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., and 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., and Smith, C.W.J. (2017). Intron retention as a component of regulated gene expression programs. Human genetics *136*, 1043–1057.

Johnson, T.L., and 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., and Lehner, B. (2016). The complete local genotype-phenotype landscape for the alternative splicing of a human exon. Nature communications *7*, 11558.

Katz, Y., Wang, E.T., Airoldi, E.M., and 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., and Chasin, L.A. (2018). Saturation mutagenesis reveals manifold determinants of exon definition. Genome research *28*, 11–24.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nature structural & molecular biology *17*, 909–915.

Kornblihtt, A.R., Schor, I.E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M.J. (2013). Alternative splicing: A pivotal step between eukaryotic transcription and translation. Nature reviews. Molecular cell biology *14*, 153–165.

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

Lander, E.S., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lee, Y., and 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., and Cartegni, L. (2011). Splicing factor hnRNPH drives an oncogenic splicing switch in gliomas. The EMBO journal *30*, 4084–4097.

Liang, X., Peng, L., Baek, C.-H., and Katzen, F. (2013). Single step BP/LR combined Gateway reactions. BioTechniques *55*, 265–268.

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., and Darnell, R.B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

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

Marcucci, R., Baralle, F.E., and 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., and 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.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell *171*, 1029-1041.e21.

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

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

Masuda, A., Shen, X.-M., Ito, M., Matsuura, T., Engel, A.G., and 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., and 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., and Stickeler, E. (2015). RON alternative splicing regulation in primary ovarian cancer. Oncology reports *34*, 423–430.

Moon, H., Cho, S., Loh, T.J., Oh, H.K., Jang, H.N., Zhou, J., Kwon, Y.-S., Liao, D.J., Jun, Y., Eom, S., Ghigna, C., Biamonti, G., Green, M.R., Zheng, X., and Shen, H. (2014a). SRSF2 promotes splicing and transcription of exon 11 included isoform in Ron proto-oncogene. Biochimica et biophysica acta *1839*, 1132–1140.

Moon, H., Cho, S., Loh, T.J., Zhou, J., Ghigna, C., Biamonti, G., Green, M.R., Zheng, X., and Shen, H. (2014b). A 2-nt RNA enhancer on exon 11 promotes exon 11 inclusion of the Ron proto-oncogene. Oncology reports *31*, 450–455.

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., Altshuler, D., and Groop, L.C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature genetics *34*, 267–273.

Mueller, W.F., Larsen, L.S.Z., Garibaldi, A., Hatfield, G.W., and 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., and 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.

Nakamura, T., Aoki, S., Takahashi, T., Matsumoto, K., and Kiyohara, 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.

Nasrin, F., Rahman, M.A., Masuda, A., Ohe, K., Takeda, J.-I., and Ohno, K. (2014). HnRNP C, YB-1 and hnRNP L coordinately enhance skipping of human MUSK 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., and 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.

Newman, M., Sfaxi, R., Saha, A., Monchaud, D., Teulade-Fichou, M.-P., and 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., and 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., and 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., and Krainer, A.R. (2009). Cooperative-binding and splicing-repressive properties of hnRNP A1. Molecular and cellular biology *29*, 5620–5631.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Molecular & cellular proteomics MCP *1*, 376–386.

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., and 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., and 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.

Papasaikas, P., Tejedor, J.R., Vigevani, L., and 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., and Chabot, B. (2007). hnRNP I/PTB can antagonize the splicing repressor activity of SRp30c. RNA (New York, N.Y.) *13*, 1287–1300.

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

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

Qian, W., and 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., and 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., and 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., and Gires, O. (2010). Heterogeneous nuclear ribonucleoprotein H blocks MST2-mediated apoptosis in cancer cells by regulating A-Raf transcription. Cancer research *70*, 1679–1688.

Roca, X., Krainer, A.R., and Eperon, I.C. (2013). Pick one, but be quick: 5' splice sites and the problems of too many choices. Genes & development *27*, 129–144.

Romano, M., Marcucci, R., Buratti, E., Ayala, Y.M., Sebastio, G., and 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.

Roque, R.S., Caldwell, R.B., and 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., and 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., and Lynch, K.W. (2005). HnRNP L represses exon splicing via a regulated exonic splicing silencer. The EMBO journal *24*, 2792–2802.

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

Savisaar, R., and Hurst, L.D. (2017b). Estimating the prevalence of functional exonic splice regulatory information. Human genetics *136*, 1059–1078.

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

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

Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M.A., Valcárcel, J., and 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., and 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., and Yu, L. (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., and 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., and 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., and 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.

Soukarieh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frébourg, T., Tosi, M., and 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., and 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.

Sterne-Weiler, T., and Sanford, J.R. (2014). Exon identity crisis: Disease-causing mutations that disrupt the splicing code. Genome biology *15*, 201.

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., and 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.

Sugaya, K., Hongo, E., Ishihara, Y., and Tsuji, H. (2006). The conserved role of Smu1 in splicing is characterized in its mammalian temperature-sensitive mutant. Journal of cell science *119*, 4944–4951.

Sun, S., Zhang, Z., Fregoso, O., and Krainer, A.R. (2012). Mechanisms of activation and repression by the alternative splicing factors RBFOX1/2. RNA (New York, N.Y.) *18*, 274–283.

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

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., and König, J. (2018). In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. Genome research *28*, 699–713.

Sutandy, F.X.R., Hildebrandt, A., and König, J. (2016). Profiling the Binding Sites of RNA-Binding Proteins with Nucleotide Resolution Using iCLIP. Methods in molecular biology (Clifton, N.J.) *1358*, 175–195.

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

Treutlein, B., Gokce, O., Quake, S.R., and 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.

Ulrich, A.K.C., Schulz, J.F., Kamprad, A., Schütze, T., and Wahl, M.C. (2016). Structural Basis for the Functional Coupling of the Alternative Splicing Factors Smu1 and RED. Structure (London, England 1993) *24*, 762–773.

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

Valverde, R., Edwards, L., and 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., and 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., and 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.

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

Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The spliceosome: Design principles of a dynamic RNP machine. Cell *136*, 701–718.

Wang, D., Shen, Q., Chen, Y.-Q., and 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, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

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

Wang, M.H., Gonias, S.L., Skeel, A., Wolf, B.B., Yoshimura, T., and Leonard, E.J. (1994a). Proteolytic activation of single-chain precursor macrophage-stimulating protein by nerve growth factor-gamma and epidermal growth factor-binding protein, members of the kallikrein family. The Journal of biological chemistry *269*, 13806–13810.

Wang, M.H., Yoshimura, T., Skeel, A., and Leonard, E.J. (1994b). 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.

Warf, M.B., Diegel, J.V., Hippel, P.H. von, and Berglund, J.A. (2009). The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. Proceedings of the National Academy of Sciences of the United States of America *106*, 9203–9208.

Warnecke, T., Weber, C.C., and Hurst, L.D. (2009). Why there is more to protein evolution than protein function: Splicing, nucleosomes and dual-coding sequence. Biochemical Society transactions *37*, 756–761.

Weldon, C., Dacanay, J.G., Gokhale, V., Boddupally, P.V.L., Behm-Ansmant, I., Burley, G.A., Branlant, C., Hurley, L.H., Dominguez, C., and 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., and Lührmann, R. (2011). Spliceosome structure and function. Cold Spring Harbor perspectives in biology *3*.

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

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

Wu, Y.X., and 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., and 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., and Lee, C. (2005). Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. Proceedings of the National Academy of Sciences of the United States of America *102*, 13526–13531.

Xing, Y., and 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., and Frey, B.J. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science (New York, N.Y.) *347*, 1254806.

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., Sahni, N., Yi, S., Rodriguez, M.D., Balcha, D., Tan, G., Costanzo, M., Andrews, B., Boone, C., Zhou, X.J., Salehi-Ashtiani, K., Charloteaux, B., Chen, A.A., Calderwood, M.A., Aloy, P., Roth, F.P., Hill, D.E., Iakoucheva, L.M., Xia, Y., and Vidal, M. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell *164*, 805–817.

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

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Journal of computational biology a journal of computational molecular cell biology *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., and Li, W.-X. (2010). Gene expression atlas for human embryogenesis. FASEB journal official publication of the Federation of American Societies for Experimental Biology *24*, 3341–3350.

Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N.M., and Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. Cell *152*, 453–466.

Zhang, X.H.-F., and 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., and 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., and 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*, 1351–136