DECODING CANCER-RELEVANT SPLICING NETWORKS IN *CD19* AND *MST1R*

Dissertation zur Erlangung des Grades

„Doktor der Naturwissenschaften“

am Fachbereich Biologie
der Johannes Gutenberg-Universität Mainz

Laura Schulz,
geboren am 04.01.1991 in Düren

Mainz, 2022

Tag der mündlichen Prüfung: 02.12.2022

Abstract

Alternative splicing is a highly complex cellular mechanism that enhances the protein-coding capacity of higher eukaryotic genomes. Like any complex process, it is prone to errors. These can lead to health-related issues. Nowadays, erroneous splicing is even considered one of the hallmarks of cancer.

In this work, we investigated two different cancer-related splicing events. In *MST1R* proto-oncogene, skipping of exon 11 results in a pathological splicing isoform that leads to progression and metastasis in cancer. In *CD19*, aberrant splicing of exon 2 has been associated with failure of the CAR-T cell therapy targeting CD19 (CART-19).

In both cases, we used a high-throughput minigene reporter assay to assess splicing changes upon thousands of different point mutations in the corresponding minigene regions. To decipher the splicing-effective mutations, linear regression-based *in silico* modeling was applied. As a result, the complete *cis*-regulatory landscape was revealed. In the case of *MST1R*, we also found that heterogeneous nuclear ribonucleoprotein H (HNRNPH) is an important *trans*-regulator. Using individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) as well as synergy analyses, we uncovered HNRNPH binding sites and showed that the protein regulates *MST1R* splicing in a switch-like cooperative manner.

Interestingly, for *CD19*, we found ~200 mutations that alter CD19 splicing and thus potentially predispose patients to relapse. In addition, we identified almost 100 novel cryptic splicing isoforms that most likely encode non-functional CD19 proteins. This, in turn, could lower CD19 protein levels and affect the long-term success of CART-19. In our analysis of *trans*-regulatory proteins, we found some RNA-binding proteins that significantly affect splicing isoforms relevant to CART-19 resistance.

We also analyzed a previously reported cryptic *CD19* splicing isoform in more detail. The isoform lacks classical splice sites and instead has repetitive sequence at both alleged splice sites. Using a splicing reporter and direct long-read RNA sequencing, we were able to demonstrate that the putative isoform is in fact an artifact caused by reverse transcription. Our results highlight the need for further validation of RNA junctions that do not exhibit the characteristics of classical splice sites.

Overall, this work not only demonstrates the importance of mutations that alter splicing in cancer, but also provides new insights into splicing regulatory networks and methodological challenges. In addition, we present potential prognostic markers that are important for assessing the risk of CART-19 resistance.

Zusammenfassung

Alternatives Spleißen ist ein hochkomplexer zellulärer Mechanismus, der die Proteinkodierungskapazität höherer eukaryotischer Genome vergrößert. Wie jeder komplexe Prozess ist er anfällig für Fehler. Diese können zu gesundheitlichen Problemen führen. Heutzutage werden Fehler im Spleißen als eines der Kennzeichen von Krebs angesehen.

In dieser Arbeit haben wir zwei verschiedene krebsbedingte Spleißereignisse untersucht. Zum einen haben wir uns das Protoonkogen *MST1R* angesehen, bei dem das „Skippen" von Exon 11 zu einer pathologischen Spleißisoform führt, die zur Krebsprogression und Metastasierung führt. Zum anderen haben wir das Spleißen von *CD19*-Exon-2 untersucht. Es wurde bereits gezeigt, dass das fehlerhafte Spleißen dieses Exons Grund für den Misserfolg der CD19-gerichteten CAR-T-Zell-Therapie (CART-19) sein kann.

In beiden Fällen haben wir einen Hochdurchsatz-Minigen-Reporter-Assay verwendet, um Spleißveränderungen von Tausenden verschiedenen Punktmutationen in den Minigenregionen zu bewerten. Um die wirkungsvollen Mutationen zu identifizieren, haben wir die Daten auf Basis linearer Regression modelliert. Als Ergebnis konnten wir die vollständige *cis*-regulatorische Landschaft darstellen. Im Fall von *MST1R* fanden wir außerdem heraus, dass das heterogene nukleare Ribonukleoprotein H ein wichtiger *Trans*-Regulator ist, der *MST1R* auf kooperative Weise reguliert.

Interessanterweise fanden wir für *CD19* etwa 200 Mutationen, die das Spleißen verändern und Patienten möglicherweise für einen Rückfall prädisponieren. Darüber hinaus identifizierten wir fast 100 neue kryptische Spleißisoformen, die höchstwahrscheinlich für nicht-funktionale CD19-Proteine kodieren. Dies wiederum könnte das CD19-Proteinlevel senken und den langfristigen Erfolg von CART-19 beeinträchtigen. Außerdem untersuchten wir auch hier die *Trans*-Regulation. Dabei fanden wir einige RNA-bindende Proteine, die das Spleißen von relevanten Isoformen erheblich beeinflussen.

Wir analysierten auch eine bereits beschriebene kryptische *CD19*-Spleißisoform genauer, da dieser Isoform klassische Spleißstellen fehlen und sie stattdessen eine repetitive Sequenz an beiden Stellen aufweist. Mithilfe eines Spleißreporters und direkter Long-Read-RNA-Sequenzierung konnten wir nachweisen, dass die vermeintliche Isoform tatsächlich ein Artefakt ist, das durch reverse Transkription verursacht wird. Diese Ergebnisse unterstreichen die Notwendigkeit einer weiteren Validierung von vermeintlichen Spleißstellen, die nicht die klassischen Merkmale aufweisen.

Insgesamt zeigt diese Arbeit nicht nur die Bedeutung von Mutationen, die das Spleißen bei Krebs verändern, sondern bietet auch neue Einblicke in regulatorische Spleißnetzwerke und methodische Herausforderungen. Wir stellen außerdem potenzielle prognostische Marker vor, die für die Bewertung des Risikos einer CART-19-Resistenz wichtig sind.

Table of Contents

1 Introduction

1.1 B cell acute lymphoblastic leukemia

B cell acute lymphoblastic leukemia (B-ALL) is the most common malignant disease in children and the most common acute leukemia in adults. The disease develops from lymphoid precursor cells, called lymphoblasts that acquired chromosomal abnormalities and other genetic alterations (Bertrand *et al.*, 2001). Usually lymphoblasts are able to differentiate into B cells, T cells, and natural killer cells. In case of B-ALL, the normal hematopoiesis is hindered. Immature, non-functional blasts grow too quickly and thereby crowd out all other healthy blood cells (Terwilliger and Abdul-Hay, 2017). B-ALL patients often suffer from anemia (low levels of red blood cells), leukopenia (low levels of white blood cells), and thrombocytopenia (low levels of platelets). For this reason, patients develop symptoms such as tiredness, shortness of breath, bruising, fever, and frequent infections. Apart from the blood, the malignant cells can also accumulate in other tissues, e.g. thymus, spleen or the central nervous system. In the United States alone, more than 6,500 new cases of ALL are diagnosed each year, of which approximately 60% occur in children and young adults below the age of 20 (American Cancer Society; Jabbour *et al.*, 2015).

In recent decades, the prognosis for children and young adults has improved dramatically. The 5-year survival rate is at ~90%. Nevertheless, relapsed and/or refractory ALL still represents one of the most important causes of cancer death in children (Paul *et al.*, 2016; Inaba and Mullighan, 2020).

Unfortunately, especially for elderly B-ALL patients, the prognosis is not nearly as encouraging as it is for children. The 5-year survival rate is only at 30-40% despite initial high remission rates (Narayanan and Shami, 2012; Jabbour *et al.*, 2015).

1.1.1 B-ALL etiology

Cancer and thereby also leukemia, is a disease caused by genetic mutations in the affected cells. The exact reasons for carcinogenesis are not well understood, though there are factors which may lead to an elevated risk of developing specific types of cancer. In the case of B-ALL these are: Down syndrome, Li-Fraumeni syndrome, Ataxia telangiectasia,

Fanconi anemia, Klinefelter syndrome, Wiskott-Aldrich syndrome, Bloom syndrome (Inaba and Mullighan, 2020). Also, environmental factors, such as radiation or harmful chemicals, or underlying infections (e.g. Epstein-Barr virus) can play a role in B cell malignant degeneration. However, in most cases the genetic alterations are acquired in otherwise healthy individuals with no genetic predisposition or unfavorable environmental exposition (Terwilliger and Abdul-Hay, 2017).

1.1.2 Subtypes

There are several subtypes of B-ALL. Usually, they are divided into groups based on the genomic alterations such as the amount of chromosomes or the kind of translocations present in the leukemic cells. The most common genetic aberration is a translocation of the tyrosine-protein kinase gene *ABL* from chromosome 9 to 22. This chromosomal aberration is called "Philadelphia chromosome" (Kurzrock *et al.*, 1988). The *ABL* gene is translocated to the breakpoint cluster region gene (*BCR*) on chromosome 22 creating a *BCR-ABL* fusion which produces an abnormal kinase that increases cell proliferation (Lugo *et al*, 1990). Around 30% of B-ALL cases carry the Philadelphia chromosome (Kurzrock *et al.*, 1988).

1.1.3 Treatment course

Depending on the age, overall health, and disease status, the treatment options vary. In many cases, different lines of treatment might be needed for long-term success in relapsed patients or patients with a refractory disease.

The first line of therapy is always induction therapy. This involves destroying as many leukemia cells as possible. The goal is to destroy, in fact, more than 99% of all leukemic cells, so that the patient goes into remission – (temporary) subsiding of disease symptoms – and normal hematopoiesis is restored. The backbone of this therapy is a multi-agent chemotherapy with vincristine, corticosteroids, and an anthracycline (Terwilliger and Abdul-Hay, 2017). Due to induction-related mortality, the therapy concentration must be much lower in elderly patients than in children and young adults (Aldoss *et al.*, 2019).

Consolidation therapy is the second line of therapy, without which most patients would relapse within the first two years after their initial remission. Consolidation therapy aims to kill any remaining, perhaps even undetectable, cancer cells that did not respond to the first line of therapy (Terwilliger and Abdul-Hay, 2017). With consolidation therapy, it is important to use only drugs that do not have cross-resistance with the drugs that were

used during the first round of therapy. In this way, it is less likely that a clone resistant to the therapy will emerge.

Maintenance therapy prevents B-ALL from coming back. It is rather lowly dosed and some treatments are only needed monthly or even every three months. It is administered for two to three years after initial therapy as any longer time frame has not been shown to have a beneficial effect (Rambaldi *et al.*, 2020).

1.1.4 Hematopoietic stem cell transplantation

One way to permanently cure leukemia and other blood disorders is allogeneic hematopoietic stem cell transplantation (allo-SCT). This option is only considered for high-risk patients and patients with relapsed and/or refractory disease with otherwise low survival rates. In order to perform an allo-SCT, a donor that matches the patient's human leukocyte antigen (HLA) markers must be found. Only around 30% of patients have HLA-matched siblings. If there is no matching family donor, a search can be made for an unrelated matching donor in the worldwide donor registries or cord blood banks (Singh and McGuirk, 2016).

After destruction of the patient's bone marrow by either high-dose chemotherapy or total body irradiation or both, the patient receives the hematopoietic stem cell product. The transferred graft not only replaces the patient's hematopoietic system, but also helps kill cancer cells, which is called the "graft-versus-leukemia effect" (Copelan, 2006). In high-risk adult patients undergoing allo-SCT, the 5-year survival rate is approximately 44% (Thomas *et al.*, 2004). In children, the overall survival rate is as high as 70% (Gassas *et al.*, 2015).

1.2 Recent developments in leukemia therapy

For decades, chemotherapeutic agents, ionizing radiation, or in many cases both, were the only way to treat any type of malignant tumor. Despite the great success of these treatments, there is still a large percentage of patients who relapse after initial remission. Some of the cancer cells have become resistant to the therapy, allowing them to multiply again. The next therapy is needed.

A very attractive option for cancer treatment is to use the body's own immune system to eliminate the remaining tumor cells before they can multiply again. A better understanding

of the role of the immune system in cancer has enabled modern immunotherapy: antibody-based therapy and even immuno-cellular therapy.

1.2.1 Antibody-based therapy

Antibodies are part of the natural humoral immune system. They can fight tumor cells by targeting surface antigens that are only present on tumor cells. Once they bind to the target cells, there are several mechanisms by which the tumor cell can be eliminated.

First, antibodies can spatially interfere with ligand receptors that are required for growth or survival pathways, cutting off the tumor cell from vital molecules. In addition, the fragment crystallizable region (Fc) portion of the antibody can be recognized by effector cells of the innate immune system to elicit an immune response against the tumor cell (Jiang *et al.*, 2011).

Antibody-based therapy is based on monoclonal antibodies. However, much has changed since antibodies were first used to treat cancer. Today, antibodies are modified in various ways to increase their natural effectiveness in eliciting an immune response. Antibodies can be coupled with drugs or toxins designed to act directly at the site where the antibody binds. In this way, the drug is delivered specifically to the tumor (Steiner *et al.*, 2011). In addition, there are bispecific antibody models with dual affinity. They bind to two different tumor antigens or, as a second target, to an antigen in the tumor microenvironment. Since antibodies do not automatically stimulate T cells, CD3, the activating T cell receptor, is often used as one of the target antigens. In this way, T cell activation is no longer restricted, making T cells more likely to participate in the immune response (Ruf and Lindhofer, 2001).

Another promising approach is "BiTEs", bispecific T cell engager molecules (Stieglmaier *et al.*, 2015). They have a different structure compared to normal antibodies. They consist of two single-chain variable tandem fragments linked by a non-immunogenic glycine-serine linker sequence (Loffler *et al.*, 2000). Both variable fragments have unique antigen specificity with one designed to bind specifically to a selected tumor antigen and the other to bind CD3 on T cells. BiTEs do not have a human Fc receptor like normal bispecific antibodies, which means that no other accessory immune cell can bind to the Fc fragment. The main advantage is the short and flexible linker between the single-chain variable fragments. The efficient bridging of the two antigens allows rapid activation of T cells and polyclonal expansion (Lutterbuese *et al.*, 2010; Stieglmaier *et al.*, 2015). The best known therapy of this type is "blinatumomab", manufactured by the company "Amgen" and

marketed under the brand name "Blincyto". It was approved in 2014 in the United States and a year later in Europe. It couples CD3 and CD19 affinity. Since CD19 is a cell surface molecule found only on B cells, this therapy was developed for B cell malignancies and is approved for use in children and adults (Martinelli *et al.*, 2017; Gökbuget *et al.*, 2018; Locatelli *et al.*, 2020).

1.2.2 The first cell-based immunotherapy: CAR-T cell therapy

In addition to the above therapies used in the fight against B-ALL, there is another, even newer type of immunotherapy on the market: CAR-T cell therapy. This is the first cell-based immunotherapy. T cells from the patient are isolated and modified, e.g. by viral transfection, to express a recombinant receptor, the "chimeric antigen receptor" (CAR) (Kershaw *et al.*, 2013). CARs combine a single-chain variable fragment with the signaling domains of the T cell receptor chain. In addition, they also carry co-stimulatory domains of receptors such as CD28, OX40, and CD137 (Huang *et al.*, 2020). In this way, CARs are able to overcome an important mechanism of cancer: evasion of the immune system by loss of major histocompatibility complex (MHC) molecules. Normally, T cells need to interact with MHC to be activated, but with the help of the co-stimulatory domains of CARs, MHC on tumor cells is no longer needed to activate T cells. Otherwise, the therapy relies on the body's normal immune reactions. Activated T cells (CAR-T cells) bind to their target on the cancer cells. After binding, they start to release cytolytic molecules, perforin and granzymes to kill the recognized tumor cell. They also attract other immune defense cells to the site by releasing cytokines such as IFN-γ and IL-2 (Andersen *et al.*, 2006). The first CAR-T cell therapy ever approved by the U.S. Food and Drug Administration (FDA) (2017) and the European Commission (2018) targets CD19. The drug is marketed by "Novartis Pharmaceuticals" and is known by the trade name "Kymriah". The whole therapy process is shown in figure 1.

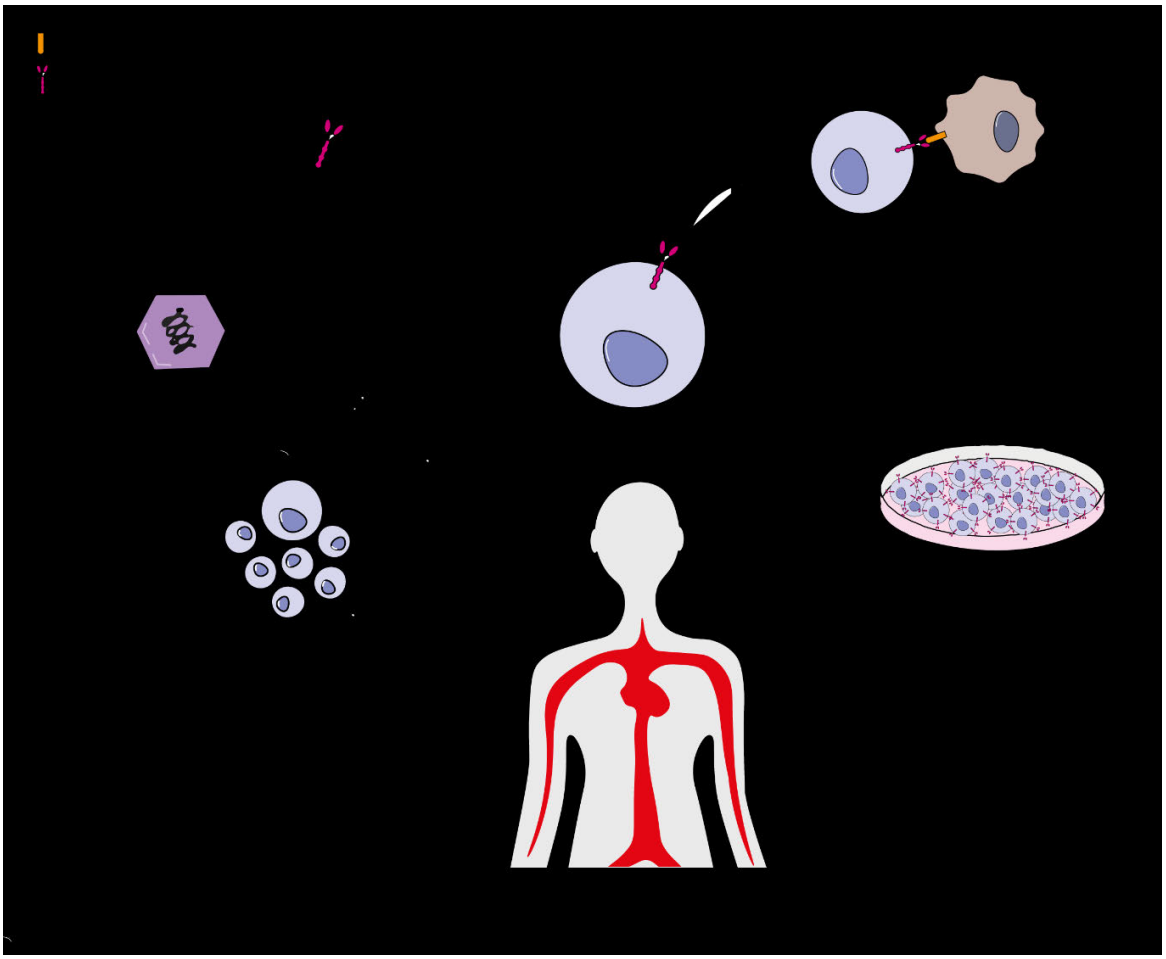


Figure 1| CAR-T cell therapy directed against the surface marker CD19. Autologous cells from the B-ALL patient are isolated and modified to express CARs. Once expanded, they can be administered back to the patient, where they eventually kill the CD19-presenting tumor cells.

1.2.3 CD19 – the target of CAR-T cells

CD19 is a transmembrane glycoprotein with a size of 95 kilodalton (kDa). It has three extracellular immunoglobulin-like domains (Nadler *et al.*, 1983). The gene consists of 15 exons that form a 556 amino acid protein with a cytoplasmic C-terminus and an extracellular N-terminus. It is expressed exclusively by the B cell lineage, making the therapy harmless to all other cells in the body (Nadler *et al.*, 1983). Moreover, unlike other potential targets such as CD20 or CD22, CD19 is strongly and widely expressed on the surface of B cells and leukemic B cells (Ning *et al.*, 2005; June and Sadelain *et al.*, 2018). It is a nearly ideal target for immunotherapy (Wang *et al.*, 2012). A disadvantage of the CAR-T cell therapy targeted against CD19 (CART-19) is the profound aplasia of B cells even after remission, as CAR-T cells remain in the body and target CD19-presenting cells. This side effect can be successfully treated with intravenous immunoglobulin as replacement

therapy, especially in children. In adults, long-lived plasma cells, which are usually CD19$^-$, can take over humoral immunity (Bhoj *et al.*, 2016).

1.2.4 Relapse under CAR-T cell therapy

Unfortunately, even successful therapies have their pitfalls, such that a certain number of patients cannot benefit from treatment in the long term due to acquired resistance. Relapses of CART-19 occur at a rate of approximately 35-50% (Gardner *et al.*, 2017; Maude *et al.*, 2018; Hay *et al.*, 2019). About half of the relapses are due to CAR-T cell failure, i.e., the transferred cells did not persist long-term in the patient and therefore failed to kill CD19$^+$ cells. The other half of the relapses are due to the loss of CD19 antigen in the cancerous B cells (Maude *et al.*, 2016). Indeed, loss of the CD19 epitope was also observed in patients treated with blinatumomab (Yannakou *et al.*, 2015), suggesting that tumor cells may escape CD19-targeted therapy due to selection pressure (Ruella *et al.*, 2016).

There are several mechanisms that can lead to loss of the CD19 epitope on B-ALL cells upon immunotherapy, most of which involve impaired expression of *CD19* mRNA: *de novo* frameshift/missense mutations, alternatively spliced *CD19* mRNA isoforms, and gene deletions spanning the *CD19* locus (Sotillo *et al.*, 2015; Orlando *et al.*, 2018; Asnani *et al.*, 2020). Importantly, alternatively spliced *CD19* mRNAs are in the focus of the scientific community as one of the most important mechanisms leading to relapses.

1.3 Regulation of pre-mRNA splicing

On average, a human protein-coding gene is 67 kbp long. It contains about 11 exons and 10 introns. An average human exon is about 309 and an intron 6,355 bp long. When exported to the cytosol after processing, an average mRNA consists of only 3,392 nucleotides, which is about half the size of an intron (Piovesan *et al.*, 2016). Thus, as little as ~5% of the pre-mRNA sequence is required for translation into a protein. This means that most of the original sequence is removed in a process called "pre-mRNA splicing" (Hastings and Krainer, 2001).

1.3.1 The major splicing machinery: the spliceosome

Several processing steps are required for eukaryotic genes to be expressed as proteins. Among these, splicing is a central step in the maturation of nascent RNA transcripts that

are transcribed from DNA. It involves the excision of pre-mRNA introns and the joining of the remaining exons. Once the introns are excised, the protein-coding genetic information of the exons can be decoded without interruption (Green, 1986).

Splicing is carried out by one of the largest ribonucleoprotein complexes in the cell, called the spliceosome. In eukaryotes, two unique spliceosomes coexist: the U2-dependent spliceosome and the U12-dependent spliceosome. The latter splices only a small subset of introns, the rare U12-type introns. Therefore, we will focus only on the U2-dependent spliceosome since U2-type introns are present in most human mRNAs (Hall and Padgett, 1996; Hastings and Krainer, 2001).

The major spliceosome is a dynamic multi-subunit complex. It consists of five small nuclear ribonucleoprotein complexes (snRNPs) U1, U2, U4, U5, and U6. In addition, a large number of non-snRNP proteins (~150) are involved, called *trans*-acting splicing factors (Zhou *et al.*, 2002). This assembly leads to a variety of RNA-RNA, RNA-protein, and protein-protein interactions that together enable the precise removal of introns and the joining of exons.

1.3.2 *Cis*-acting pre-mRNA splicing elements

RNA molecules contain short conserved sequence elements, called *cis*-elements, which are required for the definition of the intron. The core *cis*-elements consist of two splice sites, a branch point, and a polypyrimidine tract (Coolidge *et al.*, 1997; Wang and Burge, 2019). Introns are classically flanked by the 5' donor splice site, which begins with a "G-T", and the 3' acceptor splice site, which ends with an "A-G". The branch point is usually located 19-32 nucleotides upstream of the 3' splice site (Padgett *et al.*, 1984). It is defined by the nucleobase adenine. Downstream of the branch point, within 4-24 nucleotides, follows the polypyrimidine tract (Gao *et al.*, 2008), which consists of the nucleobases uridine and cytosine and is usually 15-20 nucleotides long (Lodish *et al.*, 2004). The consensus sequence around the branch point is not strictly defined in humans, so it is likely to be recognized in association with the polypyrimidine tract and/or with other *cis*-elements present (Gao *et al.*, 2008).

1.3.3 Stages of splicing

Splicing takes place in several steps. During the process, the spliceosome forms various complexes. Chemically, splicing involves two successive transesterification reactions (Moore *et al.*, 1993). First, the exon-intron boundaries are recognized. The 5' splice site is

recognized by U1 snRNP, while the 3' splice site is recognized by U2 and its auxiliary factor U2AF2, which binds to the polypyrimidine tract. The interaction of U1 and U2 brings the two ends of the intron into close proximity, and the spliceosome is initially constructed across the intron. However, when the intron is longer than 250 nucleotides - as is usually the case in metazoans - the upstream U2 snRNP interacts first with the downstream U1, which marks the exon-intron boundaries across the exon: the exon definition model (De Conti *et al.*, 2013).

Recognition of the two splice sites initiates the entire splicing process. Splicing factor 1 (SF1) binds to the branch point. Together, they form the early spliceosomal complex E. In a further step, the U2 snRNP displaces SF1 from the branch point and forms complex A (also called pre-spliceosome). U5, U4, and U6 assemble to form a snRNP trimer; together they form the pre-catalytic complex B. Through a series of major rearrangements within the RNA-RNA and RNA-protein interactions, complex B converts to its activated state, complex B*. U6 binds to the 5' splice site and thereby releases U1. When U6 forms a snRNA structure with U2, U4 is released as well. The interaction of U6 and U2 triggers the following catalytic reaction (Will and Lührmann, 2011; Matera and Wang, 2014).

Now the actual chemical reactions - two transesterifications - take place, which change the RNA composition. First, the 2'-OH group of the branch site adenosine attacks the phosphate of the conserved guanosine of the 5' splice site, cleaving the guanosine from the downstream exon and forming an intron lariat (complex C). The next catalytic step (complex C*) is triggered by the free 3'-OH group remaining at the upstream exon. To link both exons, the free 3'-OH group attacks the phosphate group of the upstream 5'-splice site, which is still bound to the intron lariat. As a result, both exons are covalently linked and the spliced intron is released (Will and Lührmann, 2011; Matera and Wang, 2014). The intron is degraded and all snRNPs are recycled (Matera and Wang, 2014). The two splicing reactions, including the intermediate product, are illustrated in figure 2.

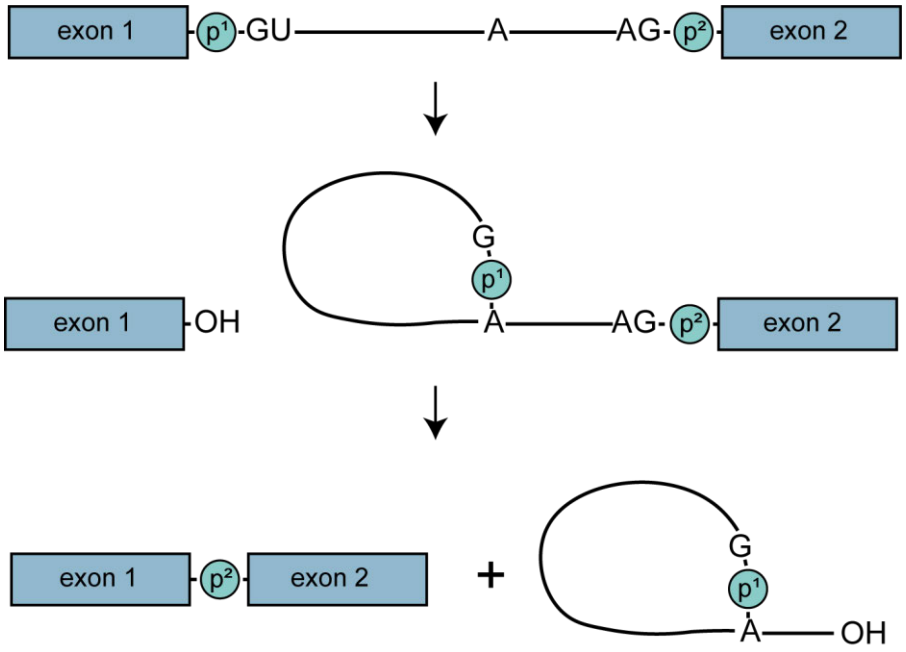


Figure 2| Splicing involves two successive transesterifications. After the first reaction, a lariat forms as an intermediate in which the 5'-splice site of the intron binds covalently to the branch site. In a second step, the two exon ends are joined by the second transesterification. The intron lariat is degraded. Adapted from Green, 1986.

1.4 Alternative splicing

A process called alternative splicing allows exons to join in different combinations to generate different mature transcript isoforms derived from a single gene. This process is one of the most important mechanisms for the generation of a variety of mRNA and protein isoforms, considering the relatively small number of human genes (~20,000) (Baralle and Giudice, 2017). This leads to increased transcriptome diversity with diverse functions that are particularly important for cell differentiation, lineage determination, tissue identity and maintenance, and organ development (Wang *et al.*, 2008; Baralle and Giudice, 2017). It is estimated that more than 95% of all human genes are alternatively spliced. Human genes generate on average two to three mRNA transcript isoforms, but there are also extreme cases, e.g., *NRXN3* comprises 1728 different RNA transcript isoforms due to the use of alternative splice sites (Stamm *et al.*, 2005).

Alternative splicing includes all changes in splicing compared to the main splicing isoform. This includes mRNAs that differ in both their untranslated regions and coding sequence. However, most alternative splicing events (~75%) occur in the translated regions of mRNAs and thus involve the protein-coding region (Okazaki *et al.*, 2002).

Possible transcript changes include exon skipping or choice between mutually exclusive exons. Also, various alternative splice sites can be used that alter the length of the included exons by either lengthening or shortening them. Another principle is intron retention, which results from the absence of splicing at a particular site. These mechanisms can lead to new or lost sequences within the mRNA, but can also alter the entire open reading frame of the transcript. All mentioned alternative splice options are shown in figure 3.

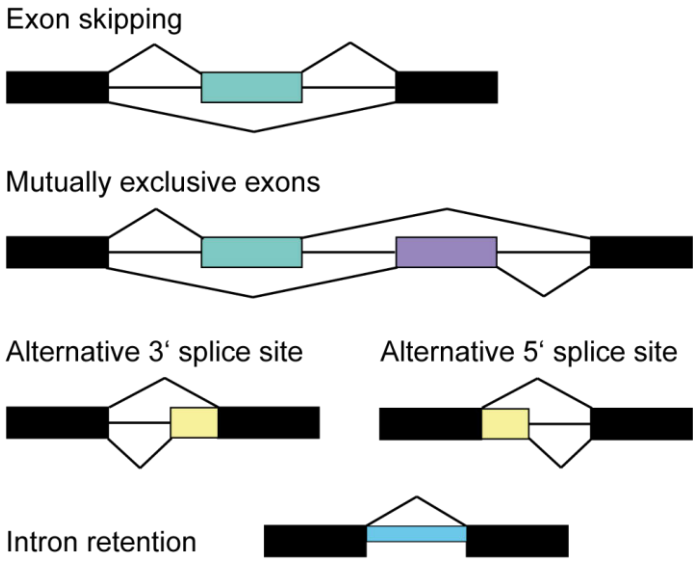


Figure 3| Different forms of alternative splicing. These are the main forms: exon skipping, mutually exclusive exons, usage of alternative 3' or 5' splice sites, and intron retention. Adapted from Stamm *et al.*, 2005.

The effects of alternative splicing on the fate of the protein can be diverse. It can alter not only transcript abundance but also protein properties such as intracellular localization, enzymatic activity, stability, binding properties, and posttranslational modifications (Stamm *et al.*, 2005). Thus, protein content may change due to splicing changes. Nonsense-mediated decay is an important mechanism in this context. It leads to mRNA degradation when a stop codon is present more than 50-55 nucleotides upstream of the last 3'-exon junction (Maquat, 2004). The appearance of new stop codons can occur especially when the reading frame has been altered. Furthermore, as described above, protein function can change too. The effects range from complete loss of function to the appearance of new functions that can alter important features of the cell. Sometimes the changes in protein isoforms are very subtle, and very sensitive assays are required to detect them at all (Stamm *et al.*, 2005).

1.4.1 The relevance of *trans*-acting splicing factors

In addition to the core splicing elements, there are further *cis*-regulatory sequence elements distributed throughout the nascent RNA transcript. Depending on where they are located, in the exon or intron, and whether they enhance or repress splicing, they are referred to as exonic or intronic splicing enhancers (ESE or ISE) or silencers (ESS or ISS) (Dvinge *et al.*, 2016). They often have the strongest effect when located near a splice site (Wang *et al.*, 2008). *Trans*-acting RNA-binding proteins (RBPs), also known as splicing factors, can recognize these elements, thereby directing the spliceosome and ultimately determining the splicing decision at each alternative exon individually. The combination of *cis*-regulatory elements and their interpretation by RBPs is commonly referred to as the "splicing code" (Fu and Ares, 2014). To ensure correct and context-dependent alternative splicing, splicing factors, similar to transcription factors, are tightly regulated and may also participate in signaling pathways (Oltean and Bates, 2014).

The most common splicing factors are serine- and arginine-rich proteins, SR proteins, and heterogeneous nuclear ribonucleoproteins, HNRNPs. SR proteins contain up to two RNA recognition motif domains at the amino terminus to provide RNA-binding specificity. The arginine/serine-rich domain at the carboxy end likely contributes to protein-protein interactions (Long and Caceres, 2009). HNRNPs contain two RNA-binding domains and an unstructured domain involved in protein-protein interactions (Krecic and Swanson, 1999). Their specific roles in different splicing scenarios are often context-specific and quite complex (Fu and Ares, 2014). Both groups of proteins are capable of promoting or suppressing splicing. SR proteins typically bind to exonic enhancers and activate splicing. They normally counteract the repressive effects of HNRNPs. However, SR proteins also bind to intronic repressor sequences downstream of the target splice site when they oppose a splicing event (Schaal *et al.*, 2005; Erkelenz *et al.*, 2013).

1.4.2 Alternative splicing in cancer

The complex process of alternative splicing can be a natural cause of disease when it becomes unbalanced. Disrupted gene expression can be carcinogenic. Some single point mutations within genes and coding sequence may be silent, so that neither the reading frame nor the amino acid sequence is altered. However, mutations can be located at positions that are important *cis*-elements for splicing and thus still affect gene expression.

This is the case when splice sites, branch points, or the polypyrimidine tract are affected. If the two highly conserved and major nucleotides of the splice sites are affected by a

mutation, this leads to the exclusion of an exon. In more than half of the cases, exon deletion leads to truncation of the encoded protein and thus to classical non-sense-mediated mRNA decay (Venables, 2004). If this happens in a tumor suppressor gene such as p53, as shown by Holmila *et al.* (2003) in many examples, it may well promote the development of cancer.

In addition to changes at the *cis*-element level, undesirable splicing changes may also be due to deregulated expression of *trans*-acting splicing factors. Thus, some splicing factors are considered tumor suppressors or even proto-oncogenes. By extension, they alter splicing of other, presumably important mRNAs involved in cancer-associated signaling pathways (Dvinge *et al.*, 2016).

Pancreatic ductal adenocarcinoma is often triggered by a mutation in p53 that leads to an increase in the splicing factor HNRNPK. The increased HNRNPK levels lead to a switch in splicing and thereby to the inclusion of cytosine-rich exons in GTPase-activating proteins (GAPs). The new GAP isoforms lead to increased *KRAS* activity, the most commonly mutated oncogene in human cancers (Escobar-Hoyos *et al.*, 2020; Uprety and Adjei, 2020).

Another example of a splicing factor that drives cancer progression is HNRNPH, which is frequently upregulated in gliomas. Its upregulation leads to exon 11 skipping in the tyrosine kinase receptor "Recepteur D'Origine Nantais" (RON), which is encoded by the *MST1R* gene. This exon skipping makes the receptor isoform (RONΔ165) ligand-independent, meaning that the receptor is continuously activated without ligand binding. The resulting continuous signal for proliferation and migration causes the cancer cell to become more aggressive and invasive (Lefave *et al.*, 2011).

1.4.3 Alternative *CD19* splicing during CART-19

Defective splicing also plays a role in cancer treatment. In acute lymphoblastic leukemia, for example, up to 50% of patients treated with innovative therapies such as CART-19 immunotherapy relapse (Gardner *et al.*, 2017; Hay *et al.*, 2019). Alternative splicing of *CD19* pre-mRNA is one of the determining factors.

CD19 exon 2 skipping (*CD19*Δex2) has been described as one of the isoforms significantly upregulated in CART-19 relapsed patients (Sotillo *et al.*, 2015). The reason for this seems to be the loss of the CD19 epitope partly encoded by exon 2 and the reduction of the presentation of this isoform on the cell surface. In a further study, the lower surface

presentation of *CD19*Δex2 was attributed to the fact that it is retained as a misfolded protein isoform in the endoplasmic reticulum (Bagashev *et al.*, 2018).

Another mRNA isoform implicated in treatment failure is retention of the second *CD19* intron (*CD19*I2R). In a retroviral cassette assay, Asnani *et al.* (2020) found *CD19*I2R in CD19$^-$ subpopulations in addition to *CD19*Δex2. *CD19*I2R is functionally equivalent to a missense mutation because intron 2 carries a stop codon in frame. To verify this, they examined translation of *CD19*I2R and found that the isoform associates mainly with the monosome and not the translating polysome, supporting the theory. With increasing *CD19*I2R transcripts in B-ALL cells, the total concentration of CD19 protein on the cell surface is also likely to decrease, which in turn would reduce the efficacy of therapy.

There is one final isoform that has been previously observed in the literature. It has an extra intron within exon 2, resulting in exon 2 being only partially incorporated into the mature mRNA. We refer to this isoform as "*CD19*ex2part" (Sotillo *et al.*, 2015; Fischer *et al.*, 2017). However, its occurrence has not yet been functionally linked to CAR-T cell therapy relapse. All four splicing isoforms of *CD19* mentioned above are shown in figure 4.

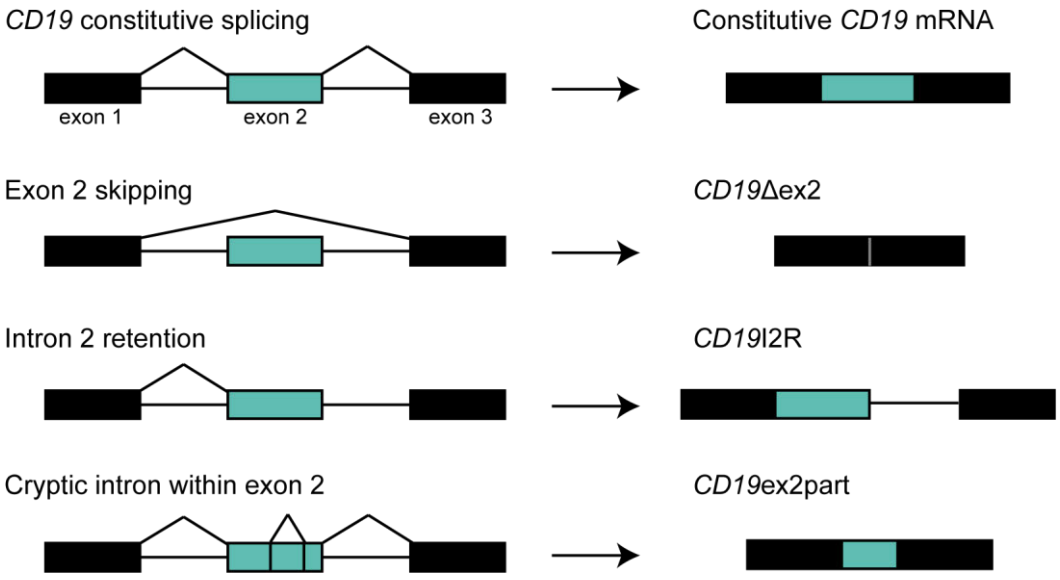


Figure 4| The known splicing isoforms of CD19. The constitutive one includes all three exons in the final mRNA. The aberrant isoforms are "*CD19*Δex2", "*CD19*I2R", and "*CD19*ex2part".

Trans-acting splicing factors also play a role in acquired CART-19 resistance. For example, SRSF3 was found to be significantly downregulated in B-ALL relapsed samples. This

reduction in splicing factor leads to increased production of the skipping isoform *CD19*Δex2 (Sotillo *et al.*, 2015).

1.4.4 How to study splicing regulation

Splicing is known for both its complexity and high specificity. However, the actual "splicing code", i.e., combinations of hundreds of RNA features to predict tissue-dependent changes in RNA splicing, is incompletely understood (Baralle and Giudice, 2017). To get closer to deciphering the code, splicing is being studied experimentally in the wet lab as well as with *in silico* strategies.

A rather old but still very useful tool to predict the strength of a splice site is based on the maximum entropy distribution. It calculates a score called "MaxEntScore" (Yeo and Burge, 2003). The higher the score, the stronger the splice site (i.e., the more similar the site is to the consensus sequence) and the more likely the site is to be used in splicing. This method is very convenient to get a first impression of the splice sites within the gene of interest, but does not take into account information about the more distant surrounding sequence.

Extensive bioinformatics studies have paved the way for advanced algorithms and deep neural networks that predict splicing much more accurately. One of these recent studies introduced a new tool called "SpliceAI" (Jaganathan *et al.*, 2019). This is an artificial intelligence tool that aims to predict the effects of clinical mutations on splicing outcomes. It can accurately determine splice site gains and losses, while also accounting for synonymous and intronic mutations that are often neglected. Unlike MaxEntScore, SpliceAI considers the entire surrounding nucleotide sequence (~10,000 nucleotides) rather than a small window around the splice sites.

Although these tools have gained functionality and precision over time, the only way to fully understand splicing changes is through experimental analysis, especially when assessing the effects of single-nucleotide changes (Soukarieh *et al.*, 2016). To test *in silico* predictions, experimental validation in the form of minigene reporters is often sought (Cooper, 2005). In these studies, minigenes are transiently expressed in cells to identify features that control splicing: relevant *cis*-elements as well as *trans*-acting factors that bind to the identified sequences.

The field began with time-consuming assays that could examine only one minigene variant at a time (Baralle *et al.*, 2003; Raponi *et al.*, 2011; Nasrin *et al.*, 2014). Nowadays, high-throughput minigene reporter studies allow scientists to examine hundreds or even

thousands of minigene variants simultaneously (figure 5). In this way, many possible mutations and, for example, every possible single-nucleotide mutation in a given region can be detected in a single experiment. From this, a complete mutational splicing landscape can be constructed (Rosenberg *et al.*, 2015; Julien *et al.*, 2016; Ke *et al.*, 2018; Souček *et al.*, 2019). However, some of these studies only consider exonic regions for mutations and neglect to analyze *trans*-acting factors involved. To study splicing in full detail, the entire minigene, including interspaced introns and relevant *trans*-acting factors, must be examined.

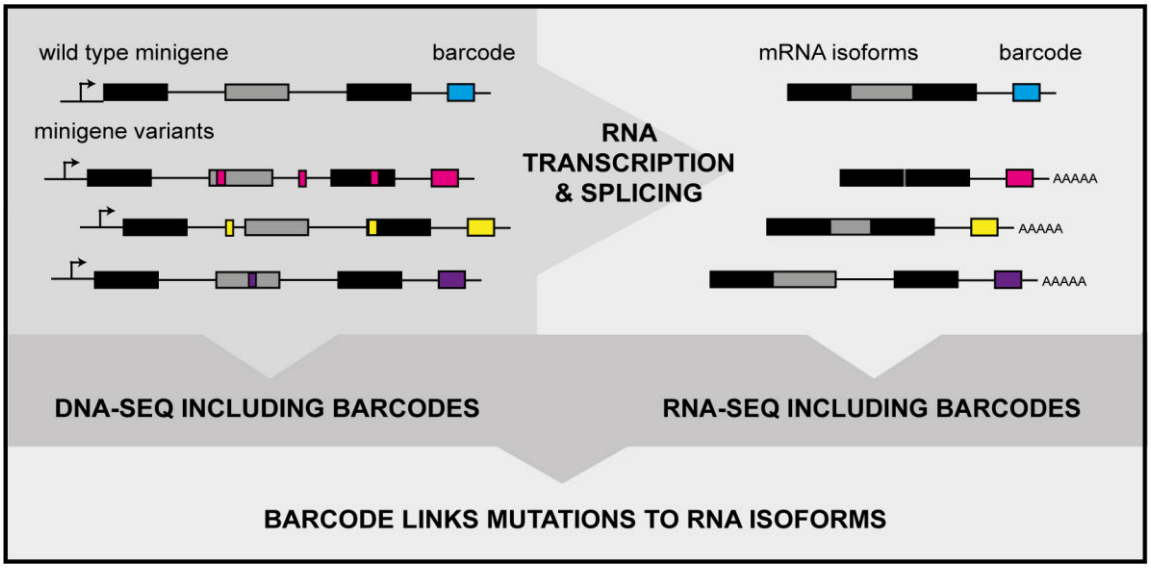


Figure 5| Overview of high-throughput mutagenesis screens with minigenes. Many minigenes are mutagenized. Wild type minigenes are always included as internal controls. The variant minigenes are sequenced by DNA-seq. When transfected into cells, they are transcribed into RNA and spliced. These mRNAs are also sequenced via RNA-seq. The barcodes at the end of the minigenes can be used to track which minigenes produced which splicing isoforms. This data can eventually be computationally processed.

1.4.5 Sequencing techniques to specifically capture mRNA isoforms

In recent years, RNA sequencing has become the most important tool for detecting gene expression and differential splicing of pre-mRNA. Many new sequencing protocols (such as the "Translatome", King *et al.*, 2016; or the "Structurome", Kwok *et al.*, 2015) have helped shed new light on RNA biology. However, the basic steps have remained the same: RNA extraction, mRNA enrichment, cDNA synthesis, amplification, followed by the actual sequencing library (Ozsolak *et al.*, 2011; Stark *et al.*, 2019). In addition, each RNA-seq experiment requires downstream computational analysis: alignment of reads, quantification, filtering, normalization, statistical modeling, or other additional analyses (Stark *et al.*, 2019).

1.4.5.1 Short-read sequencing

When writing about RNA sequencing, one usually understands short-read sequencing on platforms such as "Illumina". Illumina technology involves DNA fragments being attached to a flat-surface flow cell, amplified *in situ* into clusters, and then used as templates for a process called sequencing-by-synthesis. This artificial process uses the sequential incorporation of deoxyribonucleotides just as in natural DNA synthesis. The synthesis can be detected because it uses fluorescent reversible terminator deoxyribonucleotides (Bentley *et al.*, 2008). It is referred to as short-read sequencing because reads are usually fragmented to a length of 200-250 base pairs (Stark *et al.*, 2019). The sequencing depth for Illumina libraries starts at 4 million and goes up to 20 billion reads per sample (Kumar *et al.*, 2019). Short-read sequencing has low error rates and is relatively inexpensive. However, it also comes with some disadvantages, especially with regard to alternatively spliced isoforms.

Because the reads are quite short, many of them cannot be unambiguously assigned to a particular splice isoform of a gene. This is the case when the important splice junction is not present in a read, or even when there is a certain junction present in more than one isoform. For this reason, short-read sequencing cannot answer every biological question, especially those related to long and complex splice isoforms. Another problem with most sequencing methods like Illumina RNA sequencing is that RNA is not actually sequenced. It must first be transcribed into cDNA. Already in the past, there have been some publications where this was admitted to be a problem and was suggested to be a cause for artifacts (Qin *et al.*, 2016; Tardaguila *et al.*, 2018).

1.4.5.2 Oxford Nanopore technology

Recently, Oxford Nanopore has come on the scene with a new technology that enables long-read sequencing of direct RNA. This method represents a turning point in the RNA world because it solves two problems for RNA researchers at once. It is independent of transcription of RNA into cDNA, which makes it more reliable because it eliminates an unwanted RNA processing step (Branton *et al.*, 2010). Second, the reads can reach lengths of up to 50 kilobase pairs, making even the most complicated splicing isoforms visible.

The detection mechanism differs significantly from the Illumina platform and other similar platforms based on DNA synthesis. In nanopore sequencing, single-stranded RNA is detected as it passes through a protein nanopore stabilized by an electrically resistant polymer membrane (Branton *et al.*, 2010). Currently, the only drawbacks are the high error rate ranging from 5-20% and high cost primarily due to low throughput (500,000 –

1 million reads per run) (Rang *et al.*, 2018; Stark *et al.*, 2019). Even RNA base modifications such as adenine methylation can be detected using nanopore sequencing, making the technology incredibly attractive to all types of RNA scientists (Liu *et al.*, 2019).

1.5 Aim of the work

The aim of this PhD project is to understand the regulatory splicing landscape of two specific cancer-related splicing events. One occurs at exon 11 of the *MST1R* gene, which is important in cancer progression. The other takes place at exon 2 of the *CD19* gene, which plays an important role in CART-19 resistance.

To explore all *cis*-elements controlling splicing, we perform a high-throughput mutagenesis screen. To this end, a minigene is created from both regions, which is then used to create a library of thousands of mutant minigene variants. Each library is transfected into human cancer cell lines where the variants are transcribed and spliced. Both the minigene DNA library and the resulting mRNAs are sequenced using next-generation sequencing. The sequencing information will be used to understand the *cis*-mutation effects on splicing through computational modeling. In particular, for the *CD19* project, we aim to analyze the effects of mutations on different splicing isoforms relevant to CAR-T cell therapy and to discover novel splicing isoforms that may play a role in therapy resistance.

We also aim to understand the mechanism behind the previously described *CD19* cryptic isoform: "*CD19*ex2part" (chapter 1.4.3). It is of great interest because it has no conventional splice sites, but a repetitive motif of eight nucleotides at the putative splice sites. To study this cryptic splicing isoform, we will use a fluorescently labeled splicing reporter and direct long-read RNA as well as long-read cDNA sequencing.

Furthermore, we would like to understand the *trans*-regulation by splicing factors in both splicing events. Therefore, we perform bioinformatic screening for relevant *trans*-regulatory proteins and investigate them in knockdown experiments. Moreover, we compare our results to patients' clinical data.

Taken together, these comprehensive data sets can be used in the future to find prognostic markers for patients prone to cancer progression and metastasis, or in the case of *CD19*, to find markers predictive of CART-19 therapy relapse. In particular, the study of therapy resistance markers is of great importance because each individual CAR-T cell therapy is not only costly but also time-consuming, and time is one of the things that patients do not have in abundance. If the chances of success are known in advance, a hematologist can adjust the overall treatment plan which in turn can improve disease outcomes.

2 Publications

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

2.1.1 Abstract

Aberrant splicing caused by mutations is commonly associated with human diseases such as cancer. The *MST1R (RON)* proto-oncogene was used to develop a high-throughput splicing screen of randomly mutated minigenes. The *RON* minigene is comprised of exons 10 to 12, including both introns. Mathematical modeling based on linear regression was applied to decipher the effects of several thousand single-nucleotide mutations on *RON* splicing. As a result, the complete *cis*-regulatory landscape of the minigene was revealed. In particular, mutations that lead to skipping of *RON* exon 11, as this corresponds to the pathological splicing isoform variant *RON*Δ165, were of great interest. Notably, the results correlate with cancer patients carrying the same mutations leading to exon 11 skipping.

In addition, we found an important *trans*-regulator of *RON* splicing: heterogeneous nuclear ribonucleoprotein H (HNRNPH). iCLIP as well as synergy analyses reveal HNRNPH binding sites and demonstrate that HNRNPH regulates *RON* exon 11 splicing in a switch-like cooperative manner. This study not only demonstrates the importance of mutations that alter splicing in cancer, but also provides new insights into splicing regulation in general.

2.1.2 Zusammenfassung

Ein fehlerhaftes Spleißen, z.B. verursacht durch Mutationen, wird häufig mit menschlichen Krankheiten wie Krebs in Verbindung gebracht. Das Protoonkogen *MST1R (RON)* wurde zur Entwicklung eines Hochdurchsatz-Splicing-Screens zufällig mutierter Minigene verwendet. Das *RON*-Minigen umfasste die Exons 10 bis 12, einschließlich beider Introns. Die mathematische Modellierung auf der Grundlage linearer Regression wurde angewandt, um die Auswirkungen von einigen tausend Einzelnukleotidmutationen auf das *RON*-Spleißen zu entschlüsseln. Als Ergebnis wurde die komplette *cis*-regulatorische Landschaft des Minigenes aufgedeckt. Insbesondere Mutationen, die zum Skippen von

RON-Exon-11 führen, da dies der pathologischen Isoform RONΔ165 entspricht, waren von großem Interesse. Es ist außerdem zu erwähnen, dass die Ergebnisse mit denen von Krebspatienten korrelieren, die die gleichen Mutationen tragen, die zum Skippen von Exon 11 führen.

Darüber hinaus haben wir einen wichtigen Transregulator des *RON*-Spleißens gefunden: das heterogene nukleare Ribonukleoprotein H (HNRNPH). UV-Kreuzvernetzungs- und Immunpräzipitationsexperimente (engl. individual-nucleotide resolution UV crosslinking and immunoprecipitation = iCLIP) sowie Synergie-Analysen zeigen HNRNPH-Bindungsstellen auf und belegen, dass HNRNPH das Spleißen von *RON*-Exon-11 kooperativ reguliert. Diese Studie zeigt nicht nur die Bedeutung von Mutationen, die das Spleißen bei Krebs verändern, sondern liefert auch neue Erkenntnisse über die Regulation des Spleißens im Allgemeinen.

2.1.3 Statement of contribution

To validate the mathematical model created to decode the single-nucleotide effect from minigenes with several mutations at a time, we needed to show that a given point mutation indeed also leads to the same splicing changes when analyzed via RT-PCR. To this end, I cloned minigenes and transfected them into HEK293 cells. I then extracted their RNA and performed cDNA synthesis. Using RT-PCR, I was able to quantify the corresponding splicing changes of all isoforms compared to the isoform amounts of the *RON* wild type minigene. I prepared the corresponding figure and reviewed the manuscript.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-05748-7 OPEN

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

Simon Braun[1], Mihaela Enculescu[1], Samarth T. Setty[2], Mariela Cortés-López[1], Bernardo P. de Almeida[3,4], F.X. Reymond Sutandy[1], Laura Schulz[1], Anke Busch[1], Markus Seiler[2], Stefanie Ebersberger[1], Nuno L. Barbosa-Morais[3], Stefan Legewie[1], Julian König[1] & Kathi Zarnack[2]

Mutations causing aberrant splicing are frequently implicated in human diseases including cancer. Here, we establish a high-throughput screen of randomly mutated minigenes to decode the *cis*-regulatory landscape that determines alternative splicing of exon 11 in the proto-oncogene *MST1R* (*RON*). Mathematical modelling of splicing kinetics enables us to identify more than 1000 mutations affecting *RON* exon 11 skipping, which corresponds to the pathological isoform RONΔ165. Importantly, the effects correlate with *RON* alternative splicing in cancer patients bearing the same mutations. Moreover, we highlight heterogeneous nuclear ribonucleoprotein H (HNRNPH) as a key regulator of *RON* splicing in healthy tissues and cancer. Using iCLIP and synergy analysis, we pinpoint the functionally most relevant HNRNPH binding sites and demonstrate how cooperative HNRNPH binding facilitates a splicing switch of *RON* exon 11. Our results thereby offer insights into splicing regulation and the impact of mutations on alternative splicing in cancer.

[1] Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz, Germany. [2] Buchmann Institute for Molecular Life Sciences (BMLS), Goethe University Frankfurt, Max-von-Laue-Str. 15, 60438 Frankfurt, Germany. [3] Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina da Universidade de Lisboa, Av. Prof. Egas Moniz, 1649-028 Lisboa, Portugal. [4] Departamento de Ciências Biomédicas e Medicina, Universidade do Algarve, Campus Gambelas, 8005-139 Faro, Portugal. These authors contributed equally: Simon Braun, Mihaela Enculescu, Samarth T. Setty. Correspondence and requests for materials should be addressed to S.L. (email: s.legewie@imb-mainz.de) or to J.Kön. (email: j.koenig@imb-mainz.de) or to K.Z. (email: kathi.zarnack@bmls.de)

Alternative splicing constitutes a major step in eukaryotic gene expression. More than 90% of human genes undergo alternative splicing[1,2], which allows the production of distinct protein isoforms with different functionalities[3,4] and plays a critical role in development and tissue identity[5]. Strikingly, tumour suppressor genes and proto-oncogenes are particularly susceptible to splicing defects. Moreover, abnormally expressed splicing factors can have oncogenic properties[6], and changes in alternative splicing contribute to key processes in cancer initiation and progression[7–9]. A detailed characterisation of splicing mechanisms is therefore fundamental to our understanding of human biology and disease.

Splicing is an important step in the maturation of nascent transcripts that comprises excision of introns and joining of exons. During alternative splicing, certain exons can be either included or excluded, thus leading to different transcript isoforms. Splicing is catalysed by the spliceosome, a multi-subunit complex that recognises the 5′ and 3′ splice sites and flanking sequence elements in the pre-mRNA. The latter include the polypyrimidine tract (Py-tract) and the branch point upstream of each exon[10]. In addition to these core splice signals, multiple *cis*-regulatory elements reside in exons and flanking introns which can be primary RNA sequence elements as well as RNA secondary structures. The recognition of *cis*-regulatory elements by *trans*-acting RNA-binding proteins (RBPs) guides the spliceosome and ultimately determines the splicing decision at each alternative exon. Altogether, the information in the pre-mRNA sequence and how it is interpreted by RBPs is commonly referred to as the splicing code[11–13].

Despite many efforts to understand the molecular rules of splicing, our knowledge about *cis*-regulatory elements and *trans*-acting factors in most cases remains far from complete. Recent bioinformatic studies aimed to decipher the splicing code by predicting the impact of sequence variants on alternative splicing decisions[14,15]. Moreover, mutagenesis screens were employed to map sequence determinants of alternative splicing. However, these studies were limited to targeted mutagenesis of synthetic reporter constructs or short exonic regions[16–18].

Recepteur d'origine nantais (RON) is a receptor tyrosine kinase encoded by the proto-oncogene *MST1R* (also referred to as *RON*). Under normal conditions, the protein is cleaved to form a functional receptor. Skipping of *RON* alternative exon 11 results in the isoform RONΔ165, which remains as a single-chain protein. Spontaneous oligomerisation of RONΔ165 results in constitutive phosphorylation[19] that promotes epithelial-to-mesenchymal transition and contributes to tumour invasiveness[20–23]. Consistently, RONΔ165 is frequently upregulated in solid tumours, including ovarian, pancreatic, breast and colon cancers[21,24,25]. On the molecular level, previous studies identified a handful of mutations that influence *RON* exon 11 splicing[26,27]. Moreover, several RBPs were reported to regulate *RON* splicing[7,26,27]. For instance, heterogeneous nuclear ribonucleoprotein H (HNRNPH; collectively referring to HNRNPH1 and its close paralogue HNRNPH2 which are 96% identical at the amino acid level[28]) was found to repress *RON* exon 11 inclusion via binding within the alternative exon. While these studies suggested that *RON* splicing is heavily regulated, most *cis*-regulatory elements remain unknown.

Here, we establish a high-throughput mutagenesis approach to comprehensively characterise the regulatory landscape of *RON* exon 11 splicing. Starting from a library of almost 5800 randomly mutated minigenes, we employ a mathematical model of the splicing kinetics to detect more than 1000 point mutations that significantly affect *RON* alternative splicing. Importantly, the deduced single mutation effects correlate with the splicing levels in cancer patients bearing the same mutations. Moreover, we comprehensively characterise how HNRNPH acts as a key regulator of healthy and pathophysiological *RON* splicing by recognising multiple *cis*-regulatory elements in a cooperative fashion. Our mutagenesis screening approach promises insights into the splicing effects of mutations in humans and the mechanisms of alternative splicing regulation in general.

Results

Random mutagenesis introduces 18,000 mutations. To systematically study the *cis*-regulatory sequence elements that control *RON* alternative splicing, we designed an in vivo screening approach based on random mutagenesis of a splicing reporter minigene (Fig. 1a). The minigene harbours *RON* exon 11 together with the complete flanking introns and the constitutive exons 10 and 12 (Supplementary Fig. 1a). We confirmed that the minigene gives rise to the same transcript isoforms as the endogenous gene in human HEK293T cells (Supplementary Fig. 1b). Moreover, mutations in a known *cis*-regulatory element led to increased *RON* exon 11 skipping as reported previously[7] (Supplementary Fig. 1c). We next amplified the minigene with error-prone PCR to spread mutations randomly across all exons and introns. A 15-nt barcode sequence was introduced downstream of constitutive exon 12 via a randomised sequence in the reverse primer to uniquely identify each mutated minigene variant. Upon vector ligation and amplification, we pooled ~6000 clones into a minigene library (Supplementary Fig. 1d). As an internal reference, the library was supplemented with wild-type (wt) minigene variants that carry distinct barcode sequences but no mutations.

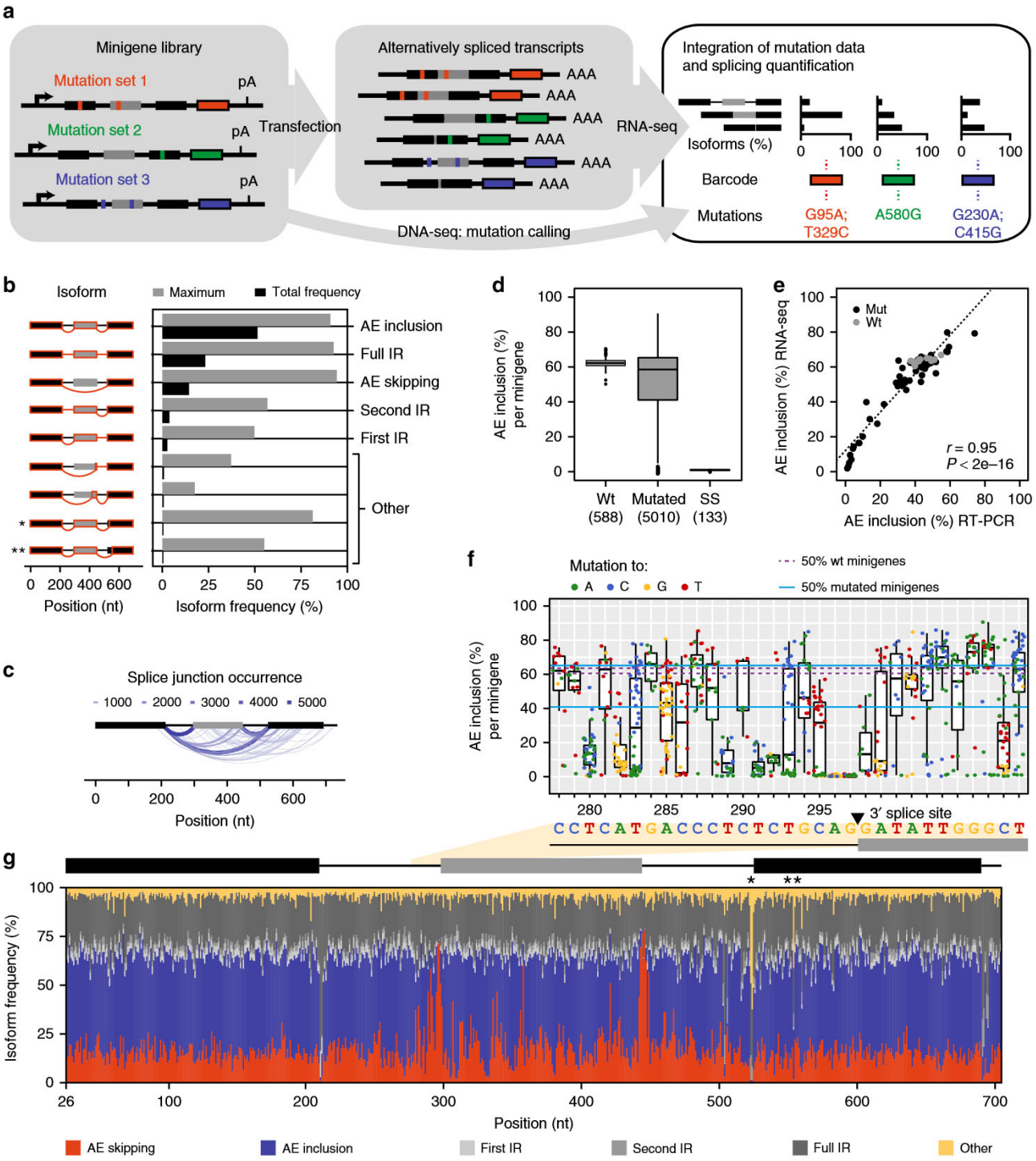
To map the introduced mutations, we sequenced the minigene library with 300-nt paired-end reads and five overlapping amplicons. The 15-nt barcode included in each read pair enabled us to assign and reconstruct the complete sequence of all minigene variants in the library (Supplementary Fig. 2a). Using a custom-tailored analysis pipeline (Supplementary Fig. 2b), we capture a total of 5791 unique minigene variants (see Methods), including 5200 with randomly introduced mutations as well as 591 with the wt sequence (Supplementary Data 1). Mutation calling identified 18,948 point mutations with an average frequency of 3.6 mutations per minigene variant. The mutations are randomly spread across the *RON* minigene, such that 97% of the positions are mutated at least ten times within the library (average 28 times per position; Supplementary Fig. 2c–e). We validated the accuracy of mutation calling with Sanger sequencing of 59 randomly selected minigene variants, confirming all 169 mutations without additional false positives.

Targeted RNA-seq quantifies alternative splicing outcome. To measure the splicing outcome, we transfected the library as a pool into human HEK293T cells where the minigenes are transcribed and spliced. We devised a targeted RNA-seq strategy based on 300-nt paired-end reads, which allows us to assemble the complete sequence of all splice products including the 15-nt barcode sequence that is present in all read pairs (Supplementary Fig. 2f, g). A total of 5598 (97%) minigene variants were captured in all three independent biological RNA-seq replicates (Supplementary Fig. 2h and Supplementary Data 1). From the RNA-seq data, we could reconstruct and quantify 163 distinct splice isoforms. The most abundant isoforms reflect the canonical splicing events, i.e., alternative exon (AE) inclusion and skipping, as well as partial and full intron retention (IR) (Fig. 1b, g and Supplementary Fig. 2g). In addition, we detected non-canonical splicing events at 82 and 71 cryptic 3′ and 5′ splice sites, respectively, which are collectively referred to as 'other' (Fig. 1c). For instance, mutations disrupting the 3′ splice site of the downstream constitutive exon 12 trigger activation of a cryptic AG (marked by one asterisk in

Fig. 1b, g). While the overall abundance of the cryptic isoforms in the RNA-seq libraries is low, they can dominate the splice products of individual minigene variants (Fig. 1b).

For the wt minigenes, i.e., in the absence of mutations, the frequency of the AE inclusion isoform (i.e., the ratio of AE inclusion over the sum of all measured isoforms) shows little variance, supporting the notion that confounding effects of the barcode sequences are negligible (Fig. 1d). In contrast, almost half of the mutated minigenes (2248, 45%) show more than 10% deviation in AE inclusion, suggesting that many introduced mutations strongly affect the splicing outcome (Fig. 1d). As expected, any mutation within the splice sites of *RON* exon 11 completely abolishes AE inclusion (Fig. 1d, f). We validated the accuracy of the RNA-seq quantification using individual RT-PCR measurements of the 59 Sanger-sequenced minigene variants (Fig. 1e and Supplementary Data 8). We conclude that the random mutagenesis approach enables precise high-throughput quantification of alternative splicing.

Linear regression modelling infers single mutation effects. Since each mutated minigene variant carries several mutations, the measured splicing changes are an overlay of multiple effects.



Consequently, a set of minigenes that share a given mutation displays a certain degree of variation in their splicing behaviour (Fig. 1f and Supplementary Data 2). To extract the impact of individual mutations, we made the simplifying assumption that mutations affect splicing independently and derived a linear regression-based mathematical modelling approach. In the linear regression model, the splicing change of each minigene relative to wt is described as the sum of single mutation effects (Fig. 2a). By fitting this model to the measured combined mutation effects, the underlying single mutation effects can be inferred.

To assess whether additivity of mutation effects can indeed be assumed, we analysed a reaction network representing splicing of the *RON* minigene using kinetic modelling (Supplementary Note 1 and Supplementary Fig. 3a). Model analysis shows that only when we consider splice isoform ratios (i.e., ratios of two measured isoform frequencies), mutation effects do not depend on the presence of other mutations in a minigene. Thus, for splice isoform ratios, mutation effects add up in log-space and a linear regression can be performed. In contrast, at the level of individual splice isoform frequencies (or related metrics such as percent spliced-in, PSI), mutation-induced fold changes depend on the mutational background and are thus not additive in log-space. We directly confirmed the additive behaviour of isoform ratios for mutations that are present as single mutation minigenes and simultaneously occur as combinations in double/triple mutation minigenes (Supplementary Fig. 4).

To integrate the full mutation information available in the data set, we formulated five separate regression models, each expressing the splicing outcome as a ratio of one splice isoform relative to the reference AE inclusion isoform. By simultaneously fitting the complete set of linear equations, each reflecting one minigene, to the experimental data, we were able to estimate 1800 single mutation effects. Based on the regression results, we could infer the frequency of five canonical splice isoforms for each of these single mutations, or combinations thereof (Supplementary Table 1 and Supplementary Data 3). The models fit the data with high accuracy, as judged by the excellent correlation between model fit and experimental data (Pearson correlation coefficient, $r = 0.99$, P value $< 2e{-}16$; Fig. 2b and Supplementary Fig. 5a, b). This supports our assumption that mutations affect splicing independently and can be described as a sum of single mutation effects (Supplementary Figs. 3b and 4a).

To test for ability of the model to infer novel combined mutations, we employed tenfold cross-validation, in which the model was fitted to 90% of the minigenes and used to predict the splicing outcome for the remaining 10%. The excellent cross-validation accuracy (Pearson correlation coefficients $r =$ 0.96–0.97, P value $< 2e{-}16$; Supplementary Fig. 6) outperformed alternative regression model variants that were fitted directly to the measured splice isoform frequencies (Supplementary Note 2 and Supplementary Fig. 7a, b). The inference power of the model for novel single mutation effects was assessed by separately leaving out one of >500 single-mutation minigene variants and fitting the model to the remaining data, or to subsets, in which further occurrences of the considered mutation were left out. This procedure revealed that the inference error for a single mutation effect decays with increasing occurrence of a mutation in our data set as $(E \sim 1/\sqrt{\text{occurrence}})$ (see Supplementary Note 2, Fig. 2c and Supplementary Fig. 5c). For mutations with occurrences >5 (i.e., present in more than five minigene variants), the estimated standard deviation of the inference error levels around 6%, suggesting that at sufficiently high occurrence the model inference accuracy is close to the experimental variation for wt minigene variants (3% standard deviation). We compared the cross-validation results to a simpler proxy in which single mutations effects are estimated from the median splice isoform frequencies over all minigenes containing a particular mutation. Even though the latter approach should average out the effect of accompanying mutations when enough minigenes are present, the regression model outperforms the median-based estimation across all occurrence levels (Fig. 2c and Supplementary Fig. 5c, d).

To independently validate the modelling results, we generated 26 minigene variants with individual mutations for which the model predictions substantially differed from the simpler median-based estimation of single mutations effects. Using RT-PCR to assess splicing outcomes, we find a strong correlation with the splice isoform frequencies inferred by the model (Fig. 2d, Supplementary Fig. 4b and Supplementary Data 8). The gain in accuracy by the model is particular apparent for mutations with a low frequency, i.e., appearing in only few minigenes. We conclude that the regression model offers a reliable method to quantify the impact of single mutations on *RON* alternative splicing.

Numerous positions contribute to *RON* alternative splicing. Using the model inference for HEK293T cells, we find a total of 778 mutations that significantly alter the frequency of at least one isoform (henceforth called splicing-effective mutations; >5% change in isoform frequency, 5% false discovery rate, FDR; Fig. 2e–g, Supplementary Fig. 8 and Supplementary Table 2). At the 5′ splice site of *RON* exon 11, we observe a good correlation between AE inclusion levels and in-silico-predicted splice-site

Fig. 1 High-throughput mutagenesis screen provides quantitative splicing information across the *RON* minigene. **a** High-throughput detection of splicing-effective mutations. Mutagenic PCR creates mutated minigene library (left) that gives rise to alternatively spliced transcripts (middle). Mutations and corresponding splicing products are characterised by DNA and RNA sequencing, respectively, and linked by unique 15-nt barcode sequence in each minigene (coloured boxes). Black and grey boxes depict constitutive and alternative exons, respectively. **b** Nine most frequent isoforms found in HEK293T cells. Bar diagram shows total frequency in RNA-seq library (black) and maximal frequency for any individual minigene variant (grey). Asterisks mark non-canonical isoforms from cryptic 3′ splice site usage upon mutations at positions marked in **g**. AE, alternative exon, IR, intron retention, other, non-canonical isoforms. **c** Occurrence of distinct splice junctions in HEK293T cells. Line thickness and colour represent number of minigene variants producing a given junction (only junctions accounting for ≥1% of all junctions for a given minigene). **d** Boxplot showing distribution of AE inclusion frequencies (as % of all isoforms) for all wild-type (wt) and mutated minigenes and a subset with mutations in splice sites (ss) of *RON* exon 11. Boxes represent quartiles, centre lines denote 50th percentile, and whiskers extend to most extreme values within 1.5× interquartile range (IQR). **e** Validation of AE inclusion frequencies for 59 randomly selected minigene variants. Scatterplot compares the RNA-seq quantification to semiquantitative RT-PCR for individual minigene variants in HEK293T cells. *r*, Pearson correlation coefficient and associated *P* value. **f** Mutational landscape around the 3′ splice site of *RON* exon 11. Boxplot of AE inclusion frequencies in HEK293T cells for all minigenes with mutation at indicated positions (*x*-axis). Box representation as in **d**. Colours illustrate inserted nucleobase (see legend). Blue and purple lines indicate IQR of AE inclusion frequencies for all mutated and wt minigenes, respectively. Sequence of wt *RON* minigene given below. **g** Isoform frequencies arising from mutations along *RON* minigene. Stacked bar chart shows median frequency of six isoform categories for all minigenes with mutation at a given position. Average of three biological replicates in HEK293T cells. Asterisks highlight positions where mutations lead to non-canonical isoforms depicted in **b**

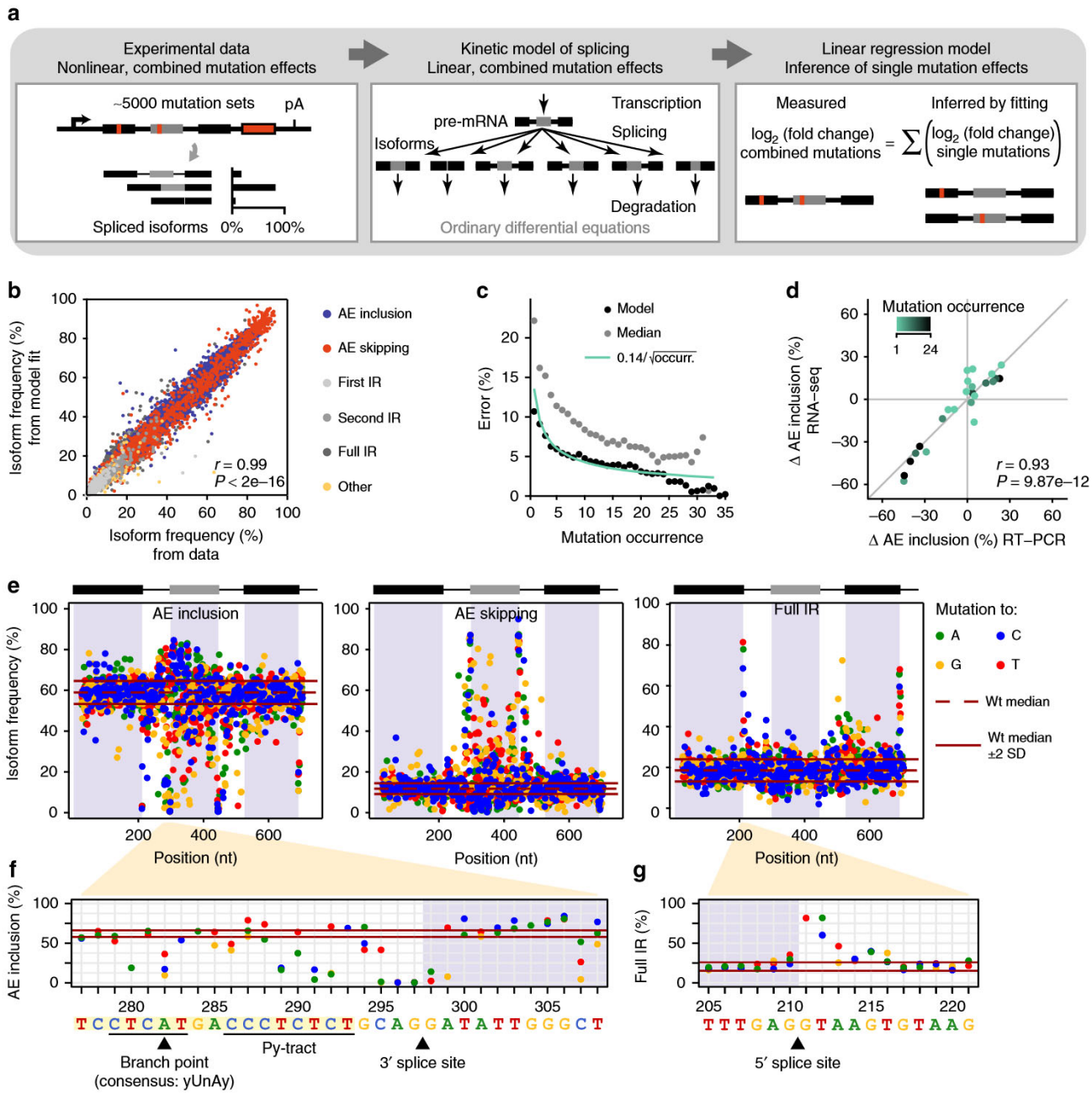


Fig. 2 A linear regression model determines more than 1900 single mutation effects. **a** Model-based inference of single mutation effects. Isoform quantifications from RNA-seq for 5598 unique minigene variants, each harbouring multiple mutations, are used as input. A kinetic model of splicing reactions reveals that splice isoform ratios show linear mutation effects, irrespective of other mutations. A linear regression model is used to infer single mutation effects in a system of 5010 linear equations, one per mutated minigene variant. **b** Regression model describes experimental measurements with high correlation (Pearson correlation coefficient $r = 0.99$, P value <2e−16). Scatterplot shows frequencies of distinct splice isoforms (see legend; separately shown in Supplementary Fig. 5a) for combined mutations calculated from fitted model against one biological replicate (see Supplementary Note 2). **c** The model more accurately infers frequently occurring single mutation effects. Cross-validation by separately excluding single-mutation minigenes (and permutations of other minigenes containing this mutation). Inference is expressed as standard deviation of inference error in AE inclusion (y-axis) and analysed for different permutations containing mutation at different frequencies (x-axis). Inference power of model (black dots) matches theoretical expectation (green line) and outperforms median-based estimation (grey dots; see Supplementary Note 2). **d** Experimental validation of model-inferred single mutation effects. Semiquantitative RT-PCR measurements of AE inclusion (other isoforms in Supplementary Fig. 4b) for targeted single-mutation minigenes that were not used for model fitting. Discrepancies between model and data appear if mutation infrequently occurs in the library (colour-coded). r, Pearson correlation coefficient and associated P value. **e** Model-inferred landscapes of 1747 single mutation effects on AE inclusion, AE skipping and full IR in HEK293T cells. Each mutation effect is indicated as a coloured dot (inserted nucleobase, see legend). Red lines indicate median (dashed) ± 2 standard deviations (SD; solid) for wt minigenes. **f**, **g** Zoom-in landscapes of single mutation effects on AE inclusion around 3′ splice site of *RON* exon 11 (**f**) and on full IR around 5′ splice site of constitutive exon 10 (**g**). Black lines and arrowheads mark splicing signals, including branch point, polypyrimindine tract (Py-tract) and splice sites. Visualisation as in **e**

strength upon mutation[29] (Spearman correlation coefficient $r = 0.89$, P value = 2.36e−08; Supplementary Fig. 9a). In contrast, predictions for 3′ splice site strength capture the effects of Py-tract composition, but fail to detect branch point mutations and other sequence contributions (Spearman correlation coefficient $r = 0.62$, P value = 4.02e−07; Supplementary Fig. 9b). As expected, transitions between pyrimidines within the Py-tract upstream of *RON* exon 11 act neutrally, whereas transversions into purines reduce inclusion, illustrating that the screen allows to discriminate base-specific effects (Fig. 2f). Consistent with the exon definition model of splicing, we find that disrupting the 5′ splice site of constitutive exon 12 (not spliced in the minigene context) also changes AE inclusion (Fig. 2e, Supplementary Table 2 and Supplementary Data 4), underlining that flanking constitutive exons can distally influence alternative splicing[11,30].

Notably, 91% of all positions within *RON* exon 11 (134/147 nt) harbour at least one splicing-effective mutation, revealing that the alternative exon is densely packed with *cis*-regulatory elements (Fig. 2e and Supplementary Fig. 8). Moreover, neighbouring positions or even different base substitutions in the same positions often affect different isoforms or change splicing in opposite directions (e.g., regions 404–429 nt or 565–567 nt, respectively, in Supplementary Data 4). The resulting patterns likely resemble footprints of the RNA sequence specificity of the interacting RBPs (see below) or RNA secondary structures. In addition to disrupting existing *cis*-regulatory elements, some mutations may also generate new elements, which further increases the complexity of the observed regulatory landscape.

The widespread occurrence of splicing-regulatory effects in *RON* exon 11 highlights that the majority of exonic positions mediate splicing regulation and thus harbour a second layer of information beyond their protein-coding function. As previously described[16], splicing-regulatory effects occur with similar frequency and effect sizes for synonymous and non-synonymous mutations (Supplementary Fig. 9e, f). Moreover, we detect substantial effects in the flanking introns and constitutive exons (50–82% splicing-effective positions per region; Supplementary Table 2). Albeit less frequent, the splicing-effective mutations within introns show comparable effect sizes to those in the alternative exon (Supplementary Fig. 9d). Globally, mutations in and around the alternative exon primarily affect the AE inclusion and skipping isoforms, whereas mutations in the downstream constitutive exon strike a balance between AE inclusion and full intron retention (Supplementary Fig. 8).

In line with a pathological relevance, we find that splicing-effective positions within introns are more conserved than non-effective positions, evidencing an evolutionary selection pressure towards maintaining the splicing-effective positions[31] (Supplementary Fig. 9c). In contrast, within exons both splicing-effective and non-effective positions show high conservation but no difference, likely reflecting constraints on amino acid composition that may overrule conservation of splicing signals. A total of 135 (25%) of splicing-effective mutations within the three exons are synonymous with respect to the encoded RON protein and would hence not be interpreted as potentially deleterious variants when considering protein sequence only. Importantly, our results clearly indicate that albeit preserving the protein sequence, such synonymous mutations may contribute to disease by changing alternative splicing patterns[32,33].

Splicing-effective positions are mutated in human cancers. Since altered *RON* splicing is involved in cancer progression[21,25], we repeated the splicing measurements in the human breast cancer cell line MCF7. Compared to HEK293T cells, the wt minigene shows lower AE inclusion in MCF7 cells, supporting a shift towards the pathophysiological state (Supplementary Fig. 1b). Nevertheless, the measured mutation effects are highly consistent between both cell lines, underlining the robustness of our screening approach (Pearson correlation coefficient $r = 0.96$, P value = 2.2e−16; Fig. 3a).

In order to address the physiological relevance of the mutations, we compared our data to the Catalogue of Somatic Mutations in Cancer (COSMIC). Out of 33 COSMIC entries within the region of the *RON* minigene, 20 coincide with splicing-effective mutations and seven of these are synonymous with respect to the encoded RON protein (Fig. 3b). It is thus conceivable that their splicing-regulatory function rather than their protein-coding role is involved in cancer progression. Prompted by this observation, we analysed patient data from The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/) to investigate *RON* splicing in human cancers. We identified 153 patients, from 19 different cohorts (representing different cancer types), that carry mutations in the *RON* minigene region specifically in their tumour samples, but not in their matched normal samples (Supplementary Data 5). We next quantified the difference in *RON* exon 11 splicing (in PSI), per cohort, between tumour samples of mutation-bearing and non-bearing patients. Strikingly, we observe a good correlation between *RON* splicing changes in mutated TCGA tumour samples and the single mutation effects determined by our approach (Pearson correlation coefficient $r = 0.62$, P value = 4.8e−05; Fig. 3c). Strongest *RON* exon 11 skipping associates with a splice site mutation (G297A; identified in a patient with thyroid carcinoma, THCA). Of note, the second largest effect is found for mutation G370T (head–neck squamous cell carcinoma, HNSC), which introduces a missense mutation at the level of the encoded protein (Fig. 3d, see Discussion). The correlation between our screen and the TCGA data is reduced if these two strongest sites are removed from the analysis (Pearson correlation coefficient $r = 0.27$, P value = 0.12; Fig. 3c), most likely because the remaining effects are weaker and compromised by experimental variation. In conclusion, our high-throughput screen recapitulates strong in vivo splicing changes in human cancer patients.

***cis*-Regulatory elements in *RON* are targeted by multiple RBPs.** In MCF7 cells, a total of 1022 mutations across the minigene affect *RON* alternative splicing, pointing towards the presence of multiple *cis*-regulatory elements (Supplementary Data 6). We used the ATtRACT database[34] to identify putative RBP binding sites, thereby predicting RBPs that recognise these *cis*-regulatory elements. In order to focus on sites that are actively involved in splicing regulation, we retained only RBP motifs if at least 60% of the positions therein showed a mutation effect on at least one splice isoform (referred to as splice-regulatory binding sites, SRBS). The analysis recovers two previously reported *cis*-regulatory elements in the alternative and the downstream constitutive exon that are targeted by HNRNPH[26] and SRSF1[21], respectively. In total, we identify 76 potential RBP regulators (Fig. 3e and Supplementary Fig. 10), suggesting that *RON* splicing is extensively controlled by multiple RBPs. To prioritise among them, we overlaid our data with a large-scale knockdown (KD) screen which tested the KD effect of 31 RBPs from our list on *RON* exon 11 splicing in HeLa cells[35]. Notably, 17 of these RBPs showed a substantial impact on *RON* splicing, with HNRNPH and SRSF2 being the strongest repressor and activator, respectively (Fig. 3e).

In a complementary approach, we investigated the expression of 190 RBPs which were identified as putative regulators of *RON* splicing by our ATtRACT analyses and/or by the published RBP KD screen[35] using matched RNA-seq data sets for 4514 TCGA

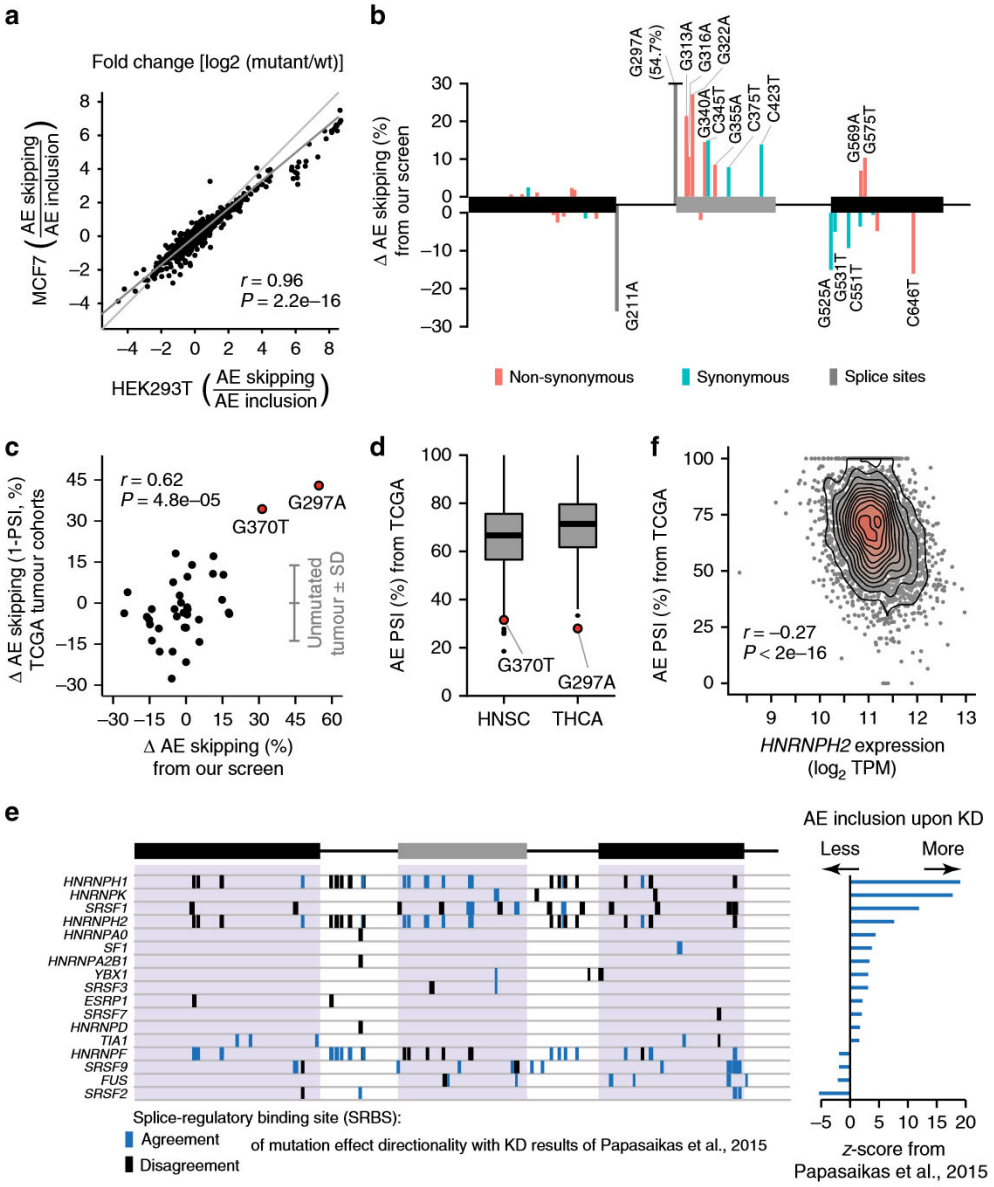


Fig. 3 Mutation effects are recapitulated in human cancer patients and point to regulatory RNA-binding proteins. **a** Model-inferred single mutation effects are consistent between HEK293T and MCF7 cells. Scatterplot compares changes in splice isoform ratios for mutations along *RON* minigene. Light and dark grey lines correspond to diagonal and linear regression, respectively. *r*, Pearson correlation coefficient and associated *P* value. **b** Somatic mutations in cancer can cause significant splicing effects. Bar diagram displays changes in AE inclusion for 33 mutations from COSMIC database. Mutations with significant effect on AE skipping are labelled ($n = 16$). Orange, blue and grey indicate non-synonymous, synonymous and splice site mutations, respectively. Mutation G297A extends beyond visualised range. **c** Mutation effects from our screen recapitulate altered splicing in cancer patients. Scatterplot compares *RON* AE skipping between mutated and non-mutated TCGA tumour samples (percent spliced-in, PSI) from 117 cancer patients (with 36 different mutations in 14 cohorts) to mutation-induced change in AE skipping in MCF7 cells. Grey lines indicate mean and standard deviation of unmutated tumour samples. *r*, Pearson correlation coefficient and associated *P* value. $r = 0.27$, *P* value = 0.12 without two mutations with strongest impact (G370T and G297A). **d** Tumour samples from mutation-bearing patients show strong *RON* exon 11 skipping. Boxplot summarises *RON* exon 11 inclusion (PSI) in head-neck squamous cell carcinoma (HNSC) and thyroid carcinoma (THCA) cohort. Box represents quartiles, centre line denotes 50th percentile and whiskers extend to most extreme data points within 1.5× interquartile range. Tumour samples with mutations G370T and G297A are labelled. **e** In silico predictions for RNA-binding proteins (RBPs) identify splice-regulatory binding sites (SRBS; predicted binding sites that show substantial mutation effects, see Methods). Boxes indicate SRBS for 17 putative RBP regulators that overlap with published data on *RON* exon 11 splicing upon RBP KD[35]. Colour code indicates whether majority of mutation effects within SRBS agree with direction of published RBP KD effect (*z*-scores; right panel). **f** *HNRNPH2* shows strongest correlation of expression levels with *RON* exon 11 splicing across 27 tumour cohorts. Density scatterplot 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

cancer patient samples from 27 different cancer types. We detect 140 RBPs whose transcript levels significantly correlated with *RON* exon 11 inclusion (FDR for Spearman correlation <5%; Supplementary Fig. 11a–c and Supplementary Data 7). Compared to all annotated RBPs or all protein-coding genes, the 190 pre-selected RBPs significantly enriched among the most highly correlated (gene set enrichment analysis, *P* value = 0.04 or 0.003, respectively).

Strikingly, the strongest association in the TCGA data set is observed for *HNRNPH2*, whose expression shows a significant negative correlation with *RON* exon 11 inclusion (Spearman

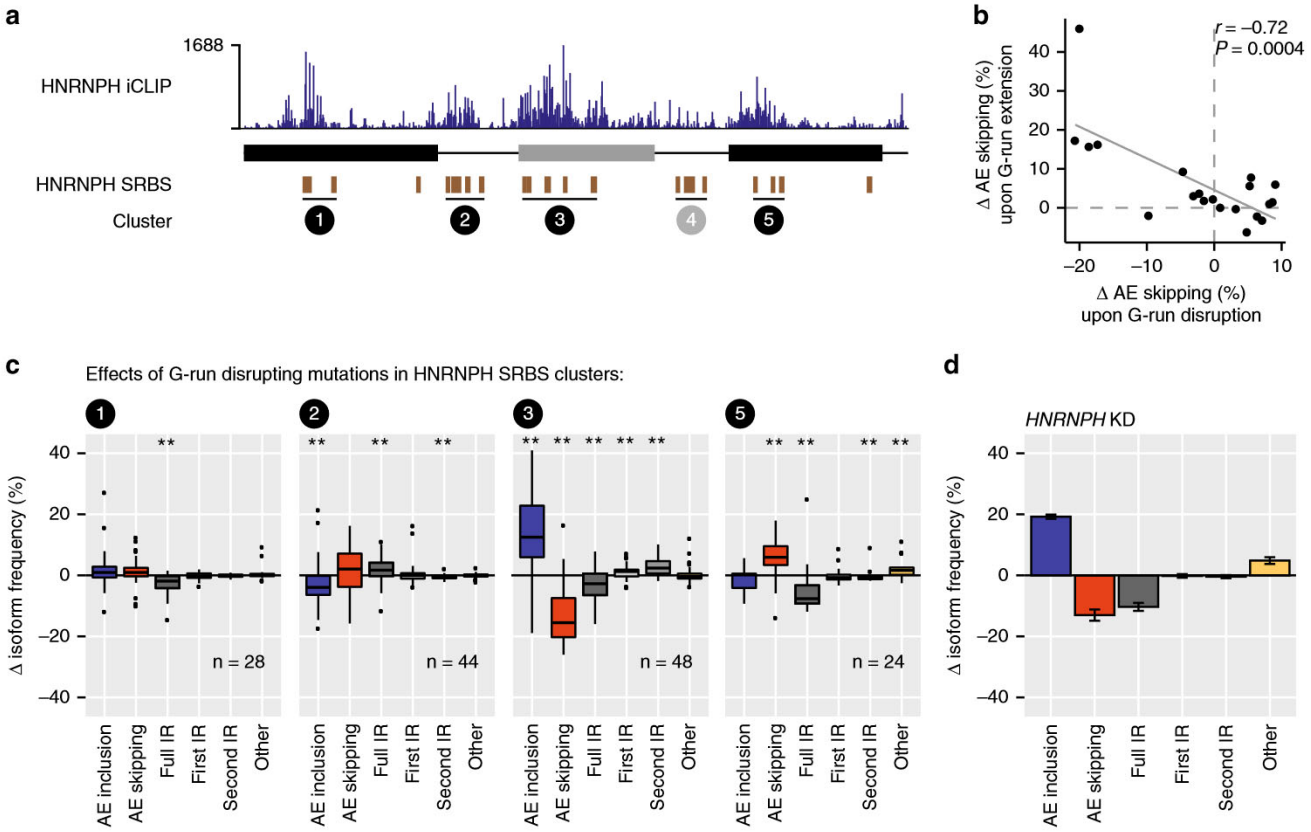


Fig. 4 HNRNPH controls *RON* exon 11 splicing via multiple intronic and exonic binding sites. **a** HNRNPH iCLIP validates HNRNPH binding to predicted splice-regulatory binding sites (SRBS). Bar diagram shows the number of HNRNPH crosslink events from HEK293T cells on each position along the wt *RON* minigene. HNRNPH SRBS (brown boxes) were assigned to five SRBS clusters (circled numbers). iCLIP data show HNRNPH binding to four out of the five SRBS clusters. **b** Extending or disrupting G-runs results in opposite splicing effects. Scatterplot shows inverse correlation of median changes in AE inclusion (average of three biological replicates) of all mutations per SRBS that extend or disrupt the G-run (see Methods) with linear regression line. *r*, Spearman correlation coefficient and associated *P* value. **c** Exonic and intronic HNRNPH binding exerts distinct effects on *RON* exon 11 splicing. Boxplots summarise the change in frequencies of each isoform in MCF7 cells (mean, $n = 3$) for all G-run-disrupting mutations within HNRNPH SRBS of different clusters (circled numbers). Box represents quartiles, centre line denotes 50th percentile and whiskers extend to most extreme data points within 1.5× interquartile range. Number of considered mutations for each cluster given below. **P* value < 0.05, ***P* value < 0.01, one-sample Wilcoxon test against population mean of zero. **d** Bar diagram shows the changes in isoform frequency of wt *RON* minigenes upon *HNRNPH* KD in MCF7 cells. Error bars indicate standard error of the mean from three biological replicates

correlation coefficient $r = -0.27$, *P* value < 2e−16; Fig. 3f). This behaviour is consistent with the previously described function of HNRNPH as a repressor of *RON* exon 11 inclusion[26]. Differentiating into the 27 TCGA cohorts, we observe a significant correlation in 11 individual cancer types (FDR for Spearman correlation < 5%; Supplementary Table 3), all negative, suggesting that HNRNPH2-mediated repression of *RON* exon 11 commonly occurs in human cancers. Notably, we find a similar association in RNA-seq data of 24 healthy human tissues from the Genotype-Tissue Expression (GTEx) Project[36] (Spearman correlation coefficient $r = -0.12$, *P* value = 5.7e−11; Supplementary Fig. 11d). Comparing GTEx and TCGA samples, we observe consistently lower *RON* exon 11 inclusion levels in the tumour samples (mean PSI 76% vs. 67%, *P* value < 2.2e−16, Mann–Whitney–Wilcoxon test), supporting an increased expression of the constitutively active RONΔ165. Accordingly, *HNRNPH2* expression is increased in cancer (mean transcripts per million [TPM] 57.88 vs. 46.29, *P* value < 2.2e−16; Mann–Whitney–Wilcoxon test). Together, these observations suggest that HNRNPH is a major determinant of *RON* alternative splicing in healthy human tissues and cancer.

HNRNPH binding can both activate and repress *RON* splicing. Within the *RON* minigene, we predict 22 SRBS for HNRNPH (Fig. 4a), all of which harbour the G-rich sequences (G-runs) recognised by HNRNPH[37]. The HNRNPH SRBS occur across all transcript regions and arrange into five clusters, each containing at least three SRBS (clusters 1–5, Fig. 4a). Individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP; Supplementary Fig. 12a) in HEK293T cells confirms that endogenous HNRNPH significantly binds at the predicted HNRNPH SRBS clusters (Fig. 4a and Supplementary Fig. 12b), with the exception of cluster 4. The strongest iCLIP signal locates in the alternative exon (Supplementary Fig. 12b).

Consistent with HNRNPH's sequence preference towards G-runs, mutations within the binding sites show opposing impact when either disrupting or generating G-runs in the RNA sequence (Fig. 4b). The direction and the most susceptible isoform depend on the position of the HNRNPH SRBS. Most prominently, mutations within SRBS cluster 3 in the alternative exon promote inclusion (Fig. 4c). A similar splicing pattern is observed for the wt *RON* minigene upon *HNRNPH* KD (Fig. 4d),

indicating that cluster 3 plays an important role in HNRNPH-mediated repression of *RON* exon 11. Mutations in the intronic clusters 2 reduce AE inclusion, whereas mutating cluster 5 in the downstream constitutive exon 12 leads to decreased intron retention, accompanied by increased AE skipping. These observations cumulate into a complex regulatory scenario, in which HNRNPH acts via multiple binding sites that have activating or repressing effects on *RON* splicing.

Synergy analysis identifies predominant HNRNPH sites. In order to identify which sites are most relevant for HNRNPH-dependent regulation, we tested the splicing response of the minigene library upon *HNRNPH* KD. We hypothesised that mutations that either weaken or reinforce an HNRNPH binding site would display positive or negative synergy with the *HNRNPH* KD. For instance, a reduced KD response compared to the wt minigene would be expected if an important HNRNPH binding site is compromised by a mutation (negative synergy).

In order to test this idea, we performed siRNA-mediated *HNRNPH* KD in MCF7 cells expressing the minigene library and used targeted RNA-seq to measure the splicing outcome. As previously reported[26,35], *HNRNPH* KD results in a strong increase in *RON* exon 11 inclusion for both wt and mutated minigene variants in the library. In line with synergy, a subset of minigene variants reproducibly show a weaker or stronger KD response compared to the remainder of the library. For instance, minigenes harbouring mutations G305A or G310A within cluster 3 consistently show elevated control AE inclusion levels, but a reduced KD response, suggesting that HNRNPH regulation is partially abolished due to these mutations (Fig. 5a).

To comprehensively identify synergistic interactions, we again turned to linear regression modelling and inferred the single mutation effects in control and KD conditions (Fig. 5b). We then calculated a *z*-score, in which the difference in the KD effect between wt and mutant is normalised by the experimental variation of the wt minigenes. Using our model based on isoform ratios (Supplementary Note 3 and Supplementary Fig. 13), we estimate that the *HNRNPH* KD on average has a 2.4-fold effect on the AE skipping-to-inclusion isoform ratio. Importantly, this effect is largely independent of the mutational background and hence the AE inclusion frequency which a minigene exhibits under control conditions (Fig. 5b, right). The exception are very strong mutations that on their own completely abolish splicing, i.e., prevent the KD from having additional measurable effects (Supplementary Fig. 7f, g). In contrast, at the level of individual splice isoform frequencies, the KD effect of all mutations strongly depends on the starting isoform level and thereby introduces systematic biases (Fig. 5b, left, and Supplementary Fig. 7e). Since such biases can be minimised by modelling splice isoform ratios, our approach allows to reliably estimate synergy.

By modelling the splice isoform ratios, we derive landscapes of synergistic interactions between *HNRNPH* KD and distinct mutations in the *RON* minigene sequence (Fig. 5c, Supplementary Fig. 12c and Supplementary Data 6). For the vast majority of point mutations, no significant synergistic interaction is observed (1428 out of 1786, 80%; Supplementary Table 2). Importantly, 354 mutations (20%) in 278 positions show significant synergy for at least one splice isoform ($|z\text{-score}| > 2$, adjusted P value < 0.001, Stouffer's test). These are significantly enriched in the SRBS in cluster 3 in the alternative exon, in which 64% of mutations (93% of positions) display synergy with *HNRNPH* KD (Fig. 5d and Supplementary Fig. 12c, d). This observation suggests that the SRBS in the alternative exon are most relevant for HNRNPH-dependent regulation. This is further supported by the fact that 42% of the strongest synergistic interactions that affect AE skipping ($|z\text{-score}| > 5$) fall into SRBS cluster 3 (Fig. 5c). Consistent with the known sequence preference of HNRNPH, we observe that the disruption of G-runs in cluster 3 leads to a weaker KD response (negative synergy; Fig. 5d and Supplementary Fig. 14). Instead, synergistic interactions at clusters 1 and 5 in the constitutive exons frequently reinforce HNRNPH-dependent regulation by creating new or extending existing G-runs, leading to a stronger-than-average KD response (positive synergy; Supplementary Fig. 14). Hence, while the HNRNPH SRBS clusters outside *RON* exon 11 do not prevail under the tested conditions, they can become more important when HNRNPH binding for these sites is increased.

In order to validate the functional relevance of SRBS cluster 3, we generated ten minigene variants with individual point mutations disrupting G-runs within HNRNPH SRBS from the five clusters (Supplementary Data 8), and tested their splicing under *HNRNPH* KD conditions using semiquantitative RT-PCR. Indeed, single mutations in cluster 3, for which the model had inferred the strongest synergistic interactions, almost completely cancel out the KD response (Fig. 5e). In contrast, minigenes with mutations in other clusters still respond to the *HNRNPH* KD, in agreement with their less pronounced synergistic interaction with HNRNPH. In summary, the synergy analysis allows to link an RBP to its functionally most relevant *cis*-regulatory elements.

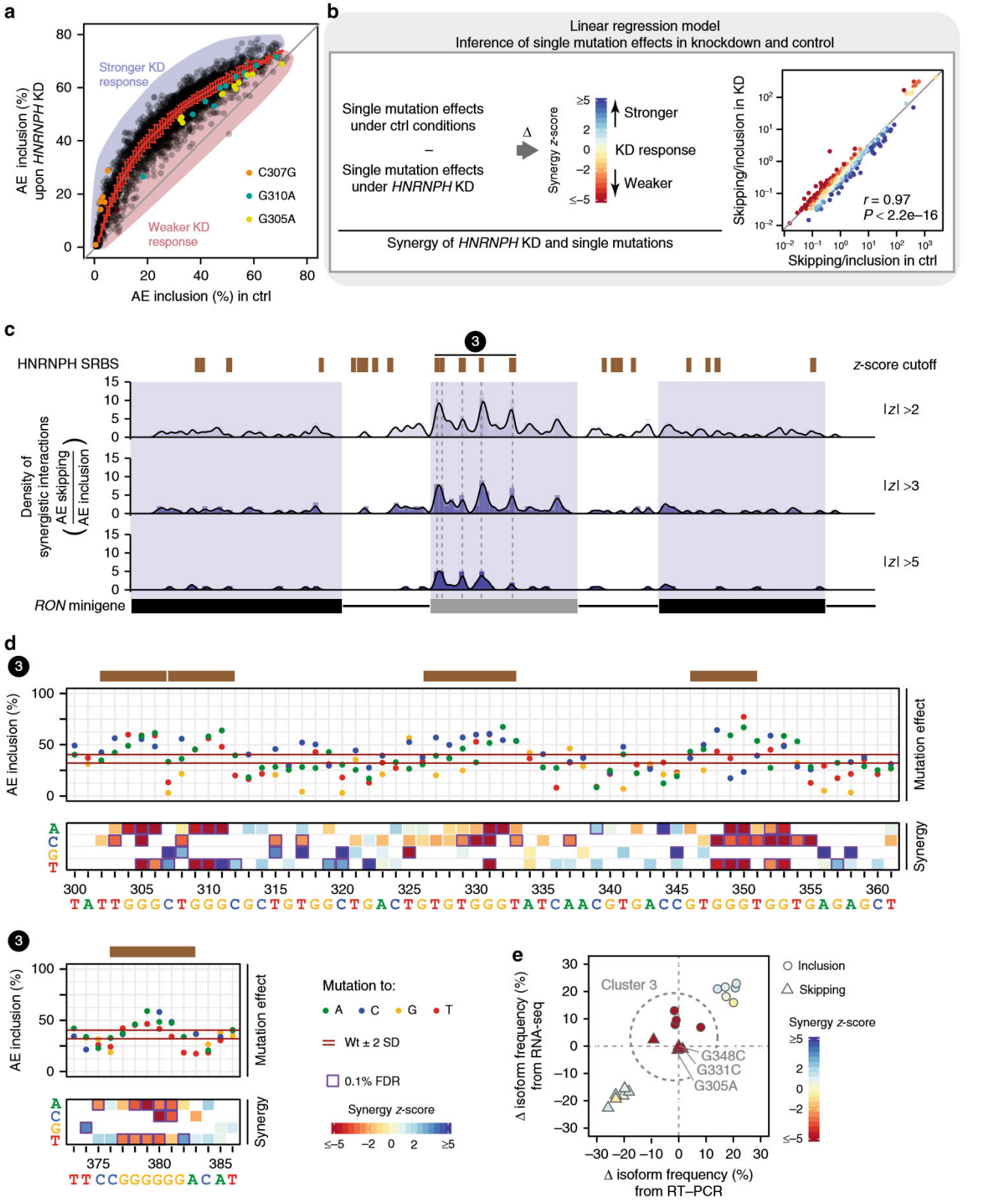
Cooperative HNRNPH binding establishes a splicing switch. We find that individual G-run-disrupting point mutations within the alternative exon (e.g., G305A, G331C and G348C) are sufficient to almost completely abolish the response to *HNRNPH* KD (Fig. 5e). This suggests that the corresponding SRBS cooperatively recruit HNRNPH. In line with this notion, our linear regression model which does not consider cooperativity, on average provides a worse fit to minigenes containing simultaneous mutations in two HNRNPH SRBS in the alternative exon (Supplementary Fig. 15e). In order to test for interdependent HNRNPH binding, we repeated the HNRNPH iCLIP experiments in the context of mutated *RON* minigenes harbouring point mutations within three different SRBS of cluster 3. In line with cooperative binding, the resulting drop in HNRNPH crosslinking is not limited to the site of the point mutation, but spreads to several further SRBS in *RON* exon 11 (Fig. 6a).

Cooperative HNRNPH binding to multiple SRBS would imply that HNRNPH regulates splicing with a steep, sigmoidal dose–response curve. To test this, we performed gradual *HNRNPH* KD and *HNRNPH1* overexpression experiments, in which we transfected MCF7 cells with increasing amounts of HNRNPH-specific siRNA and *HNRNPH1* overexpression construct, respectively. Notably, we observe a switch-like splicing response of *RON* exon 11 from the minigene as well as the endogenous *RON* gene. Indicative of strong cooperativity, we find that the dose–response curves can be described by high Hill coefficients for the endogenous *RON* gene (nH = 17.4, confidence interval (CI) [10.8,35.2]) as well as the transfected wt *RON* minigene (nH = 13.8, CI [10.4,17.7]; Fig. 6b and Supplementary Figs. 15a–d, 16). Consistently, we observe that *HNRNPH2* shows the steepest regression slope among the 190 RBPs tested for expression correlation in the TCGA data (Figs. 3f and 6c). Even though *HNRNPH2* expression in the TCGA data is not variable enough to reach plateaus in splicing, the steep slope further supports a switch-like behaviour of *RON* exon 11 inclusion.

Based on these observations, we conclude that *RON* exon 11 splicing is extensively regulated via multiple interdependent HNRNPH binding sites that exert strong cooperativity. This enables switch-like splicing with small changes in HNRNPH concentration causing large changes in splicing (Fig. 6d), potentially explaining why *HNRNPH* expression is a strong predictor of *RON* exon 11 splicing in cancer cells.

Discussion

Systems approaches combined with mathematical modelling are required to fully comprehend the complex regulation of alternative splicing. Our work builds on previous approaches to measure the effect of mutations in defined regions of splicing-reporter minigenes[16–18,38]. Central to our analytical framework is the mathematical splicing model which allows us to predict the



effects of individual mutations based on measurements of combined mutation effects. We employ linear regression modelling to disentangle these effects, and validate the predictive power of this approach using cross-validation, targeted single mutations and by relating *RON* mutations to splicing outcomes in cancer patients.

To formulate the linear regression model, we investigated how mutation effects cumulate in minigenes exhibiting several mutations. We termed a mutation effect linear if a mutation induces the same fold change in splicing irrespective of the mutational background (i.e., other mutations being present). If linearity holds true, the mutation effects add up in log-space and a linear regression can be performed to infer single mutation effects from the measured combined mutations. Using a kinetic model reflecting *RON* alternative splicing, we found that splice isoform ratios show the desired linear behaviour (Supplementary Note 1 and Supplementary Fig. 7e). Accordingly, the isoform ratio-based regression model fitted the complete minigene library with high accuracy. In line with our approach, Rosenberg et al. quantitatively modelled the contribution of randomised *k*-mer sequences in 25-nt regions of a synthetic minigene using an additive model that was based on the AE inclusion-to-skipping ratio[18]. In contrast, direct linear regression using the splice isoform frequencies decreases the accuracy and inference power of the model (Supplementary Note 2 and Supplementary Fig. 7b), possibly indicating nonlinear interactions between mutations at this level. Therefore, care needs to be taken when interpreting the interplay of mutations and/or other perturbations directly based on the abundance of certain splice isoforms (e.g., percent spliced-in/PSI, or equivalent metrics), as each perturbation shifts the operating point of the system. As a global trend, we observe that minigenes showing inclusion frequencies around 50% are most sensitive to perturbations such as *HNRNPH* knockdown (Fig. 5a). However, after transformation to isoform ratios, *HNRNPH* knockdown elicits linear, context-independent changes (Fig. 5b). Thus, isoform ratios are superior when analysing the interplay of multiple treatments or mutations, while isoform frequencies are essential for judging the physiological impact of splicing changes.

Our conclusion that combined mutations can be accurately described as a linear combination of single point mutations implies that synergistic interactions between mutations have only a minor impact on *RON* splicing outcomes. Intriguingly, our data suggest that even simultaneous mutations in two *cis*-regulatory elements mostly elicit linear, independent effects: across more than 100 minigenes containing two or more simultaneous mutations in any HNRNPH SRBS, 93% of the splice isoform frequencies can be explained within 5% deviation from the measured value using the linear regression approach (Supplementary Fig. 15e). Thus, the goodness of fit of this subset is comparable to the complete minigene population, suggesting that *cis*-regulatory elements in many cases act independently on *RON* alternative splicing.

Despite most mutations acting independently, we observe cooperative effects for adjacent HNRNPH binding sites in the alternative exon (see below). Such nonlinear effects are not captured by our model, but are in line with previous work showing that splicing-relevant mutations can amplify each other in combination, thereby showing cooperative interactions[11,16]. That such nonlinear effects are more prevalent in a previous screen by Julien et al.[16] may result from the fact that their study systematically screened double mutations in close vicinity. However, when relating the goodness of fit of our linear regression model to the nearest distance between two splicing-effective mutations, we found no clear effect of the mutation proximity on the fitting error (Supplementary Fig. 7c, d). This suggests that also nearby mutations typically act independently of each other and agrees well with results from a recent saturation mutagenesis study of a 51-nt region in the alternatively spliced exon of the *WT1* gene[17]. Since our data set does not exhibit enough coverage to comprehensively detect cooperative interactions of nearby mutations, it remains possible that adjacent sites mutually influence each other, whereas distal *cis*-regulatory elements act independently. Such a scenario would be consistent with a local assembly of ribonucleoprotein complexes that act as independent regulatory units.

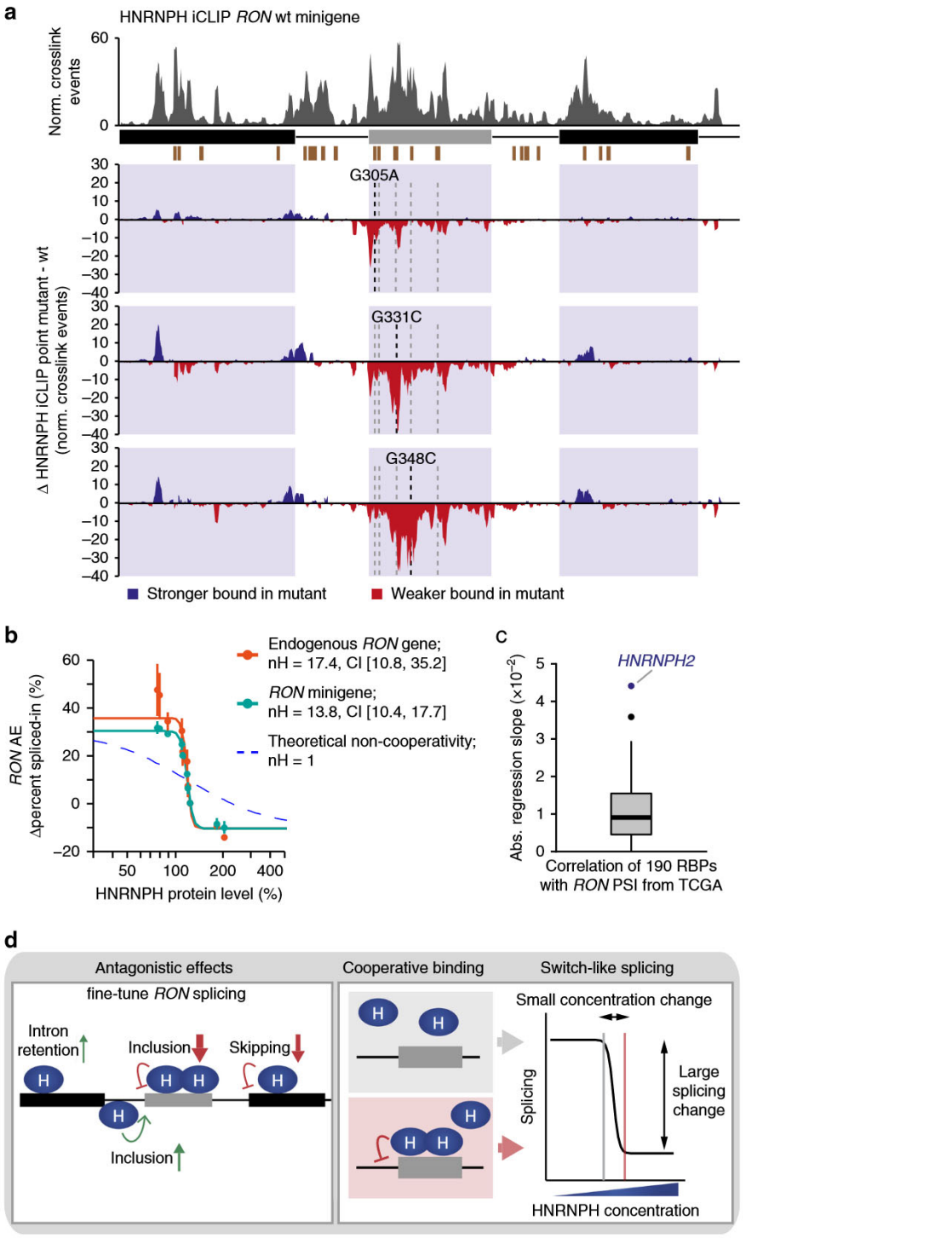
Our high-throughput mutagenesis screen uncovers a highly complex *cis*-regulatory landscape, with >80% of all positions affecting *RON* alternative splicing. Within this set, we recover mutations in all previously identified *cis*-regulatory elements[7,21,26,27,39]. Within the alternative *RON* exon 11, we find that 91% of all positions show a significant impact on *RON* splicing. Conceptually, these splicing-effective mutations either disrupt existing *cis*-regulatory elements at the RNA sequence or structure level or generate novel elements, thereby further increasing the complexity of *RON* splicing regulation. Even though the newly generated *cis*-regulatory elements do not occur under normal conditions, they may be relevant in cancer when mutations accumulate. A similar density of effective mutations was also reported for *FAS* exon 6[16]. Our study demonstrates that other than previously suggested, such a densely packed

Fig. 5 Synergistic interactions highlight the functionally most relevant HNRNPH binding sites. **a** Minigene variants are differentially spliced upon *HNRNPH* knockdown (KD). Scatterplot compares AE inclusion under control (ctrl) and *HNRNPH* KD conditions for all minigene variants. Average behaviour illustrated by running mean and standard deviation (red). Shadings schematically highlight stronger/weaker-than-average KD response. Minigenes with mutations C307G, G310A and G305A in HNRNPH splice-regulatory binding sites (SRBS) cluster 3 are highlighted. **b** Quantification of synergistic interactions by linear regression modelling. Single mutation effects are determined separately for ctrl and *HNRNPH* KD using linear regression and subtracted to estimate KD responses compared to wt (*z*-score based on standard deviation of wt minigenes, colour-coded; see Supplementary Note 3). Right graph shows model-inferred AE skipping-to-inclusion isoform ratios of single mutations in ctrl vs. KD. Regression line indicates average KD effect. Consideration of isoform ratios, as compared to isoform frequencies (**a**), leads to linearisation of KD response in line with predictions of kinetic splicing model (Supplementary Fig. 3a). **c** Model-inferred synergistic interactions accumulate in *RON* exon 11. 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. **d** Mutations in HNRNPH SRBS cluster 3 lead to increased AE inclusion and reduced *HNRNPH* KD response. 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. Heat maps (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). **e** Consistent with strong synergistic interactions (colour-coded), mutations in cluster 3 almost completely abolish the *HNRNPH* KD response in MCF7 cells. Scatterplot compares model-inferred estimates with semiquantitative RT-PCR measurements of AE inclusion (circles) and AE skipping (triangles) upon *HNRNPH* KD (mean, $n = 3$) for ten targeted mutations in HNRNPH SRBS (Supplementary Data 8)

arrangement is not exclusive to short exons such as *FAS* exon 6[16] (63 nt) or *SMN1* exon 7[38] (54 nt), as *RON* exon 11 (147 nt) is around the average length of human cassette exons[40].

A major advance of our study is that we detect *cis*-regulatory elements along the entire minigene, including introns and flanking constitutive exons. In the constitutive exons, the effect sizes are generally smaller, possibly due to an accumulation of partially redundant exonic splicing enhancers (ESEs) that ensure constitutive splicing[13]. Notably, we find that mutations not necessarily trigger intron retention, but can also specifically swap between AE inclusion and skipping (such as T581G or G686C; Supplementary Data 6). These distal effects agree with previous observations from positional splicing maps (RNA-maps), showing that RBP binding at flanking constitutive exons can regulate the inclusion of neighbouring alternative exons[41].

Our ATtRACT analysis together with a previous study[35] suggest that splicing of *RON* exon 11 is controlled by a multilayered network of at least 17 *trans*-acting RBPs. Many of these RBPs are linked to multiple binding sites in different regions, further increasing the regulatory complexity. Importantly, significant correlations in our TCGA analyses suggest that many of the regulatory relationships are functional in humans in vivo. A multilayered regulation of alternative splicing events has also been suggested by a recent high-throughput screen for RBP KD



effects[42]. We extend beyond this view by directly identifying synergistic interactions between sequence mutations and RBP KD.

Our study highlights HNRNPH as a key regulator of *RON* exon 11 splicing. Using mutational analysis and iCLIP, we demonstrate that it acts via five clusters of intronic and exonic SRBS that antagonistically affect *RON* splicing (Fig. 6d). In line with previous global splicing maps[37,43,44], we observe that HNRNPH binding in the alternative exon represses AE inclusion, whereas binding in the flanking introns increases AE inclusion. Synergy analysis pinpoints the SRBS in the alternative exon as the functionally most relevant sites. We speculate that this interwoven arrangement of antagonistic SRBS may allow to fine-tune *RON* splicing. More generally, tightly regulated exons might benefit from modulating not just one but several competing splicing reactions in order to achieve an optimal adjustment of alternative splicing under changing physiological conditions.

Other than previously suggested for intronic HNRNPH sites[44], we find strong indications for cooperative binding of HNRNPH to multiple SRBS within the alternative exon. One possible mechanism for the observed cooperativity would be oligomerisation of HNRNPH via its glycine/tyrosine (GY)-rich domain. It was recently shown that other hnRNP proteins form multimeric assemblies via GY-rich domains to regulate splicing[45]. Moreover, HNRNPH binding sites might fold into G-quadruplex structures[46–49], which could contribute to the observed cooperativity through sequestering and simultaneously releasing G-runs. The cooperative HNRNPH binding renders *RON* splicing sensitive to individual mutations in HNRNPH SRBS or to small changes in HNRNPH protein expression.

Extensive changes in alternative splicing are characteristic for many human cancers[50], and it has been estimated that about half of all synonymous driver mutations change splicing[51]. The skipping of *RON* exon 11 results in a ligand-independent, constitutively active variant of the encoded RON receptor tyrosine kinase, RONΔ165[20]. Contrary to initial reports[52], we and others detect RONΔ165 expression also in healthy human tissues[21], suggesting that the encoded protein could play a role under physiological conditions. However, overexpression of RONΔ165 was shown to trigger increased cell motility and invasive tumourigenesis[20]. Consistent with this oncologic potential, abnormal RONΔ165 accumulation has been described in breast and colon cancers, among others[25].

With the help of our mutagenesis screen, we identify many mutations that trigger a strong skipping of *RON* exon 11. Importantly, the mutation effects in our screen are reflected in cancer patients bearing the same mutations. Several of the mutations are synonymous with respect to the encoded protein, suggesting that mutation-induced splicing changes can have deleterious impact in cancer[51,53–55]. In addition, we also identify numerous non-synonymous mutations that have a strong, and in some cases a surprising, impact on splicing regulation: for instance, the nonsense mutation G370T found in a head–neck squamous cell carcinoma (HNSC) patient also triggers *RON* exon 11 skipping (Fig. 3c, d). Intriguingly, this splicing change inverts the physiological consequence of the mutation, as the majority of mature *RON* transcripts will exclude the mutated exon and thereby translate into a constitutively active rather than a prematurely truncated RON protein.

Due to their prevalence in cancers, altered *RON* isoforms represent a promising target for intervention[56]. For instance, clinical trials assessed the therapeutic potential of monoclonal antibodies targeting RON to block the binding of its ligand MSP (MST1)[57] (ClinicalTrials.gov Identifier: NCT01119456; antibody RON8, Narnatumab, ImClone; phase-I discontinued). However, tumours expressing the constitutively active isoform RONΔ165 can specifically escape this kind of therapies, as this protein no longer requires ligand-dependent activation[24]. A detailed knowledge of mutations that promote this isoform might therefore allow a personalised therapy in the future.

Methods

Cloning of the *RON* wt minigene. To generate the *RON* wt plasmid, a segment of the *MST1R* gene was amplified by polymerase chain reaction using Phusion DNA polymerase (NEB) with the forward primer 5′- CCCAAGCTTTGTGAGAGGCA GCTTCCAGA-3′ and the reverse primer 5′- CAGTCTAGANNNNNNNNNNN NNNGGATCCGCCATTGGTTGGGGGTAGGGGCTGATTAAAGGTAGG-3′ at 65 °C annealing temperature with human genomic DNA (Promega) as a template (Supplementary Table 4). The 779 bp DNA product was gel-purified with the QIAquick Gel Extraction Kit (QIAGEN) and then digested using *Hin*dIII and *Xba*I 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 harbouring point mutations were generated using the Q5 Site-Directed Mutagenesis Kit (NEB) according to the manufacturer's instructions.

Mutagenic PCR and library construction. For the mutagenesis of the *RON* minigene, we used the GeneMorph II Random Mutagenesis Kit (Agilent) according to the manufacturer's instructions. Aiming for an average mutation rate of 3.5 mutations/minigene, three libraries were independently generated and finally fused. To this end, 8 and 4 µg of the unmutated *RON* wt plasmid were amplified with 30 cycles, and 0.8 µg of the unmutated *RON* wt plasmid were amplified with 20 cycles. The primers used to amplify the mutagenic fragments were 5′-CCCAA GCTTTGTGAGAGGCAGCTTCCAGA-3′ (forward primer) and 5′-CAGTCTAG ANNNNNNNNNNNNNNNGGATCCGCCATTGGTTGGGGGTAGGGGCTGA TTAAAGGTAGG-3′ (reverse primer) (Supplementary Fig. 1a and Supplementary Table 4). The PCR products were purified using the QIAquick Gel Extraction Kit (QIAGEN). Purified DNA was cut with *Hin*dIII and *Xba*I (NEB) restriction endonucleases for 45 min at 37 °C and subsequently purified using a PCR

Fig. 6 Cooperative HNRNPH binding establishes a splicing switch of *RON* exon 11. **a** A single point mutation in an HNRNPH splice-regulatory binding site (SRBS) results in reduced HNRNPH binding in HEK293T cells also at neighbouring SRBS in *RON* exon 11. Bar diagrams show the number of HNRNPH iCLIP crosslink events on the wt *RON* minigene (top) and the difference in normalised crosslink events on wt and mutated *RON* minigenes (mutations G305A, G331C and G348C in different SRBS within cluster 3, marked by dashed lines; bottom) in a sliding 5-nt window along the wt *RON* minigene. HNRNPH SRBS (brown boxes) indicated below. **b** Splicing response to gradual *HNRNPH* KD and overexpression suggests cooperative regulation of *RON* exon 11 by HNRNPH. Scatterplot shows semiquantitative RT-PCR quantifications of *RON* exon 11 inclusion (in percent spliced-in/PSI, Supplementary Fig. 15a, b) from endogenous *RON* gene (orange) and wt *RON* minigene (blue) against corresponding HNRNPH protein levels (Supplementary Fig. 15c, d). Degree of cooperativity is quantified by fitting Hill equation (solid lines) and compared to theoretical fit for non-cooperativity (dashed line). Error bars denote standard deviation of three biological replicates. **c** Steep regression slope for *HNRNPH2* supports cooperative HNRNPH regulation and switch-like splicing of *RON* exon 11. Boxplot shows distribution of regression slopes for expression correlations of 190 RBPs with *RON* exon 11 inclusion in TCGA samples (Supplementary Data 7). Box represents quartiles, centre line denotes 50th percentile and whiskers extend to most extreme data points within 1.5× interquartile range. *HNRNPH2* is highlighted. **d** HNRNPH acts as key regulator of *RON* splicing by recognising multiple *cis*-regulatory elements in a cooperative fashion. Schematic model summarises position-dependent effects of HNRNPH on *RON* exon 11, indicating most strongly effected isoform for each site (left panel). Multiple interdependent HNRNPH binding sites within *RON* exon 11 exert strong cooperative control on the alternative exon, resulting in a splicing switch upon small changes in HNRNPH abundance (right panel)

purification kit (QIAGEN). The digested plasmid DNA and mutagenic fragments were ligated for 5 min at room temperature in a volume of 21 µl containing 50 ng of plasmid and 21 ng of insert (3:1 ratio of insert to plasmid DNA), 10 µl of 2× Quick Ligation Reaction Buffer and 1 µl Quick T4 DNA ligase (NEB). Transformations were carried out via $CaCl_2$ transformation of *Escherichia coli* DH5-alpha strain with 2 µl of the ligated DNA. Bacteria were plated in low density to allow the formation of similar-sized colonies and determination of the number of transformants by counting of the colonies. Sixteen hours after the transformation, ~2000 colonies per transformation were washed off the plates into lysogeny broth (LB) medium and plasmids were extracted using the Plasmid Plus Midi Kit (QIAGEN). In addition, 200 wt plasmids were generated to be used as a spike-in to the above-mentioned libraries by using the same primers and template wt plasmid but non-mutagenic 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. Note that the remainder of wt minigenes in the library represent the proportion of error-free minigenes within the product pool of the mutagenic PCR. Purification, digestion and transformation were performed as described above. Mutagenesis and wt libraries were pooled together to yield a library of ~6000 plasmids. To obtain single plasmids of the library for benchmarking via Sanger Sequencing and validation via RT-PCR, a re-transformation of the library was carried out and plasmids of resulting colonies were extracted using QIAprep Spin Miniprep Kit (QIAGEN).

Semiquantitative RT-PCR. Semiquantitative RT-PCR was used to quantify 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), 1 µl dNTPs (10 mM) 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 anneal to the upstream constitutive exon and a region located downstream of the random barcode but upstream of the polyadenylation site: 5′- TGCCAACCTAGTTCCAC TGA-3′ (forward primer) and 5′- GCAACTAGAAGGCACAGTCG-3′ (reverse primer). The primers to amplify endogenously derived isoforms were 5′-CCTGA ATATGTGGTCCGAGACCCCCAG-3′ (forward primer) and 5′-CTAGCTGCT TCCTCCGCCACCAGTA-3′ (reverse primer; Supplementary Table 4). The TapeStation 2200 capillary gel electrophoresis instrument (Agilent) was used for isoform quantification of the PCR products.

Cell culture and transfection of plasmids and siRNAs. HEK293T and MCF7 cells were grown in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% foetal bovine serum at 37 °C with 5% CO_2. Standard *HNRNPH* KD was carried out using single small interfering RNA (siRNA) against *HNRNPH*[58] (5′-GGAGCUGGCUUUGAGAGGA[dT][dT]-3′, Sigma-Aldrich) or non-targeting control siRNA (5′-UGGUUUACAUGUCGACUAA[dT][dT]-3′, Sigma-Aldrich) at a final concentration of 20 nM. KD efficiencies were assessed by western blot analyses. For gradual *HNRNPH* KD, the siRNA concentration was varied between 0.05 nM and 10 nM. One day prior to transfection, 2×10^5 HEK293T cells were seeded in a 6-well plate to result with ~20% confluence at the day of transfection. MCF7 cells were seeded 3 days prior to transfection with 0.5×10^5 cells per well of a 6-well plate. The transfection mix was prepared by incubating 3 µl RNAiMax (Invitrogen) with 2 µl siRNA (20 µM) in 200 µl OPTI-MEM (Invitrogen) for 20 min, and the mix was added in a dropwise manner to the cells. For transfection of plasmids 24 h later, a mixture of 2 µg minigene plasmid DNA and 20 µg polyethylenimine MW ~2500 transfection reagent (Polysciences, Inc.) in 100 µl OPTI-MEM (Invitrogen) was prepared and incubated for 20 min before it was added to the cells. Cultures were harvested another 24 h later. For the HNRNPH1 overexpression, 4×10^5 MCF7 cells were seeded in a 6-well plate 1 day prior to transfection. Transfection was carried out using Lipofectamine 2000 (Invitrogen) and 1 or 2.5 µg pcDNA 3.1 (+)-HNRNPH1 overexpression construct or pcDNA 3.1 (+) empty vector control. The minigene plasmid was transfected 24 h later as described above and cells were harvested another 24 h later. RNA was extracted using the RNeasy Plus Mini Kit (QIAGEN) according to the manufacturer's protocol. For semiquantitative RT-PCR analysis of splicing isoforms without KD conditions, 7×10^5 HEK293T cells were seeded and transfected the next day with plasmid DNA under the above-mentioned conditions.

No cell line used in this paper is listed in the database of commonly misidentified cell lines maintained by ICLAC. HEK293T (CRL-3216) and MCF7 cells (HTB-22) were purchased from ATCC (Manassas, VA) without further authentication. Cell lines were tested for mycoplasma contamination on a monthly basis.

Library preparation and high-throughput sequencing. For preparation of high-throughput RNA sequencing (RNA-seq) libraries, the total RNA obtained from transfected HEK293T cells or MCF7 cells was enriched for mRNA by performing polyA selection of 20 µg of total RNA using Dynabeads® Oligo $(dT)_{25}$ beads (Invitrogen) according to the manufacturer's protocol. Reverse transcription was carried out using 500 ng of enriched mRNA under the above-mentioned conditions. To prevent the formation of chimeric amplicons, the libraries were amplified using emulsion PCR[59], with Phusion DNA Polymerase (NEB) and either cDNA derived from polyA-selected RNA in the case of RNA-seq or plasmid DNA of the minigene library in the case of high-throughput DNA sequencing (DNA-seq). To amplify fragments for RNA-seq, the following primers containing Illumina sequencing adaptors were used (Supplementary Fig. 2g): 5′-CAAGCAGAA-GACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAA CCGCTCTTCCGATCTNNNNNNNNNNGTTCCACTGAAGCCTGAG-3′ (forward primer) and 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTC CCTACACGACGCTCTTCCGATCTNNNNNNNNNNATAGAATAGGGCCCT CTAGA-3′ (reverse primer) under the following PCR conditions: 98 °C for 30 s, 15 cycles of [98 °C for 10 s, 56 °C for 20 s, 72 °C for 60 s] and final extension at 72 °C for 5 min. For the DNA-seq library amplification, the same PCR conditions and 18 cycles with different primer combinations were used (Supplementary Fig. 2a and Supplementary Table 4). Following amplification, the DNA-seq PCR products were cleaned using the GeneRead Size Selection Kit (QIAGEN) according to the manufacturer's instructions. Products intended for RNA-seq were purified using Agencourt AMPure XP beads (Beckman Coulter). Purified products were first analysed with the TapeStation 2200 capillary gel electrophoresis instrument (Agilent) and then fluorimetrically quantified using a Qubit fluorimeter (Thermo Scientific). RNA-seq and DNA-seq were carried out on the Illumina MiSeq platform using paired-end reads of 300 nt length and a 10% PhiX spike-in to increase sequence complexity.

Western blot. Cell lysates were prepared with modified RIPA buffer (50 mM Tris HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% NP-40, 0.1% sodium deoxycholate, protease inhibitor cocktail; Roche). The following antibodies were used for western blot analyses: rabbit polyclonal anti-HNRNPH, 1:10,000 dilution (AB10374, Abcam) and mouse monoclonal anti-HNRNPA1, 1:10,000 dilution (R4528, Sigma-Aldrich).

iCLIP experiment and data processing. We used iCLIP to capture the binding pattern of HNRNPH on the *MST1R* transcript. iCLIP was performed according to a previously published protocol[60]. In brief, 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, we used 7.5 µg of a polyclonal rabbit anti-HNRNPH antibody from Abcam (AB10374). RNase digestion was performed by adding 10 µl of 1/100 diluted RNase I (Ambion) to the sample of the wt minigene experiment shown in Supplementary Fig. 12a or 1/300 diluted RNase I (Ambion) to each sample of the experiment shown in Fig. 6a (comparison of the iCLIP landscape of the *RON* wt minigene with point mutation minigenes). Reverse transcription was carried out with RT primers listed in Supplementary Table 4. We performed the sequencing on an Illumina HiSeq 2500 for the *RON* wt minigene (51-nt single-end reads) and the *RON* wt/point mutant minigene comparison was sequenced on either Illumina MiSeq or NextSeq 500 with 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[61] resulting in ~49 million (HiSeq 2500), ~10 million (MiSeq) or ~121 million (NextSeq 500) uniquely mapping reads. In order to quantitatively compare HNRNPH iCLIP data for the *RON* wt and point mutation minigenes (Fig. 6a), crosslink events were normalised to the total number of crosslink events within the minigene region excluding *RON* exon 11. Normalised counts were averaged between replicates, counted into 5-nt sliding windows and then subtracted between conditions to determine differences in HNRNPH crosslinking.

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 analysed with a custom Python pipeline (version 2.7.9: Anaconda 2.2.0, 64-bit; Supplementary Fig. 2b). In detail, we used FastQC (fastqc_v0.11.3; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for quality control and Trimmomatic[62] (version 0.33; parameters HEADCROP:20 SLIDINGWINDOW: 7:10 MINLEN:0) to remove excess sequence and trim low-quality bases (average Phred score < 10 in 7-nt window). After trimming, we 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, we used matchLRPatterns() from the R/Bioconductor package 'Biostrings' to search for the flanking restriction sites (Lpattern = TCTAGA, Rpattern = GGATCC, allowing one mismatch). We only retained read pairs with 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[63] (version 0.4.12). 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)[64] with -dt NONE. We recounted

overlapping reads 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 and are provided in addition to the point mutations in Supplementary Data 3. The final library contained 5791 minigene variants, including 591 wt and 5200 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.

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 (Supplementary Table 1), and analysed with a custom Python pipeline similar to DNA-seq (see above; Supplementary Fig. 2f). Briefly, we removed low-quality sequences (average Phred score < 20 in 6-nt window) and extracted the 15-nt barcode (read #1) as described above. Only reads originating from the 5791 minigene variants that were recovered from the DNA-seq library were considered for further analyses. Read pairs for each minigene variant were aligned to the *RON* wt minigene sequence using the splice-aware alignment algorithm STAR[61] (version 2.5.1b), allowing up to ten 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 5697, 5645 and 5623 minigene variants detected in RNA-seq replicates 1, 2 and 3 from HEK293T cells, respectively (Supplementary Fig. 2h and Supplementary Table 1).

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 <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 collected in the isoform category 'other'.

Dynamic model of splicing. We modelled the splicing dynamics using a set of ordinary differential equations, in which concentrations of RNA intermediates are determined by production and degradation terms (Supplementary Fig. 3a). 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 normalised by the total RNA concentration. These fractions can be calculated analytically from the steady state of the system (see Supplementary Note 1). As a result, we find that 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 the sum of the unspliced pre-mRNA (x_0) and full intron retention (x_3) isoforms, which cannot be discriminated experimentally. Thus, due to normalisation, 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 is 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.

Calculation of single mutation effects by linear regression. For the estimation of single mutation effects in HEK293T cells, we assumed that the combined log fold changes of multiple mutations on a splice isoform ratio can be written as the sum of individual log fold changes (see Supplementary Note 2). One such equation was formulated for each minigene, resulting in a system of 5621–5697 equations for each splice isoform ratio, depending on the amount of minigene variants that were detected in the RNA-seq replicates (Supplementary Table 1).

To support our assumption of additive mutation effects, we analysed how single mutation effects interact in minigenes containing several mutations. To this end, we analysed 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 (Supplementary Fig. 4a and Supplementary Table 1). For the majority of these mutations, we observed that the combined mutational effects on the splicing rates K_i are multiplicative, e.g., $K_i(m_1,m_2)/K_i(\mathrm{WT}) = K_i(m_1)/K_i(\mathrm{WT}) * K_i(m_2)/K_i(\mathrm{WT})$, where $K_i(\mathrm{WT})$, $K_i(m_1)$, $K_i(m_2)$ and $K_i(m_1,m_2)$ 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, we calculate the mutational effects $K_i(m_1,\ldots m_n)/K_i(\mathrm{WT})$ 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 (Supplementary Fig. 3a). 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 (see Supplementary Note 2 for details).

As an alternative approach to estimate the single mutation effects, we calculated the median isoform frequency across all minigene variants that harbour a given mutation, and compared these numbers to the estimation of the regression model (Supplementary Fig. 4b). 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.

Estimation of the inference accuracy of the model. The training data set contained about ~600 mutations that were measured also as single mutations in individual minigenes (Supplementary Table 1). We used these single mutation minigenes to estimate the inference accuracy of the model, and to assess the dependency of the inference accuracy on the occurrence of a mutation in the data set. For each such mutation, the following cross-validation procedure was repeated: The single mutation minigene was removed from the data set 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 data set successively and each time the effect of the mutation was assessed by regression and the estimation compared to the single mutation minigene value. In this way, we obtained estimates for the prediction error based on 1 up to $n-1$ minigenes containing a particular mutation, where n is the total occurrence of the mutation in the data set. In some cases, estimation of mutational effects was not possible from a reduced data set, e.g., the prediction error for a particular mutation was estimated only for occurrences between m and $n-1$, with $1 < m \le n-1$. Finally, the standard deviation of the inference errors for all mutations was estimated for each measured frequency (Fig. 2c).

Significant 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) (see Supplementary Note 2 for details). To quantify the effects of each individual mutation on each isoform frequency, we calculated a z-score value from the model-derived single mutation effects, using the mean and standard deviation of the 591 wt minigene variants: $\frac{\left(p_i^{\mathrm{mutation}} - \mathrm{mean}\left(p_i^{\mathrm{wt}}\right)\right)}{\mathrm{Standard\ deviation}\left(p_i^{\mathrm{wt}}\right)}$. The z-scores were independently calculated per replicate and later averaged. Only mutations present in all three replicates were kept for further analyses.

In order to combine the evidence from the three replicate experiments, we applied Stouffer's test to combine the z-scores[65]. The resulting standard-normally distributed metric was converted into a P value and subjected to multiple testing correction (Benjamini–Hochberg). We considered a mutation as significant for a given isoform if it displays (i) ≥5% change in isoform frequency compared to the mean of the 591 minigene variants (ΔIF ≥ 5%), and (ii) less than 5% false discovery rate (FDR, adjusted P value <0.05). Combining all six isoform categories, this approach identified 778 and 1022 splicing-effective mutations in HEK293T and MCF7 cells, respectively (Supplementary Table 2). 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, we divided the log-transformed fold change in isoform ratios (KD over control condition) by the wt variation (standard deviation; see Supplementary Note 3). z-scores were calculated by replicates and then averaged, removing mutations that were not present in the three replicates under KD conditions. We then used Stouffer's test and multiple testing correction as above. Since the uncertainty of the synergy z-score (measured as the standard deviation between replicates) increases near 0% AE inclusion due to boundary effects (Supplementary Fig. 7g), we excluded the splice sites (positions 209–210, 298–299, 443–444, 523–524 and 689–690) mutations that on their own completely abolish splicing (<1% isoform frequency under control conditions). To identify significant synergistic interactions, we applied a cutoff at 0.1% FDR (adjusted P value <0.001). Additionally, we required a consistent directionality of the synergistic effects in all three replicates. Combining the five different isoform ratios, this approach identified 354 significant synergistic

interactions (|*z*-score| > 2) on 278 positions between mutations and *HNRNPH* knockdown in MCF7 cells (Supplementary Table 2). Applying more stringent cutoffs at |*z*-score| > 3 or >5 identified 222 or 66 significant synergistic interactions, respectively (Supplementary Table 2).

Characterisation of splicing-effective positions. Splice site strengths were predicted using the sequence analysis software MaxEntScan[29] for all mutations in the positions considered by MaxEntScan (278–300 nt and 442–450 nt for the 3′ and 5′ splice site, respectively; Supplementary Fig. 9a, b). PhyloP scores[66] 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; Supplementary Fig. 9c).

Annotation of splice-regulatory RBP binding sites (SRBS). We used the Scan Sequence tool of the ATtRACT database[34] 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. We retained only those binding sites for which ≥60% of positions were identified as splicing effective in our 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 Supplementary Fig. 12b, 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 (used in Fig. 4c), while the two possible H-to-G mutations in immediately neighbouring positions were counted as G-run-extending. Figure 4b compares the median splicing effect (average of three biological replicates) of all G-run disrupting versus extending mutations for the 22 predicted HNRNPH SRBS.

Analysis of TCGA and GTEx data. Normalised gene expression data for 11,688 post mortem samples from 30 human tissues, collected from 714 non-diseased human donors, were retrieved from the GTEx project[36] (v7). Normalised gene expression data from TCGA tumour samples (https://cancergenome.nih.gov/) were retrieved from Firebrowse (http://firebrowse.org/). Alternative splicing for both data sets was quantified using *psichomics* (version 1.2.1, https://github.com/nuno-agostinho/psichomics), using the default minimum coverage to calculate *RON* exon 11 PSI values. We quantified both gene expression and *RON* exon 11 PSIs for 2743 normal samples, from 24 healthy human tissues, and 4514 tumour samples, from 27 cancer types. The comparison of *HNRNPH2* expression between tumours from TCGA (9807 samples) and healthy tissues from GTEx (7851 samples) was done using TPM values calculated at Toil[67], which are already normalised for comparison.

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/). We identified a total of 153 patients bearing 55 different mutations within the region of our *RON* minigene (Supplementary Data 5). 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 our screen for each mutation. Since we observed that the correlation was affected by the minimum read coverage used to calculate PSIs, we restricted the correlation analysis 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; Fig. 3c). 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 cohorts considered in Fig. 3c and with more than 24 reads mapping to the involved splice junctions.

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[35]. The study used *z*-scores calculated from the 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 cutoff 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, we searched for RBPs 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[35] 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 (Supplementary Data 7 and Supplementary Table 3). 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[8] and of all protein-coding genes using Gene Set Enrichment Analysis (GSEA) tool[68,69]. 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, we performed linear regressions 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.

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 (Supplementary Fig. 15a–d, 16) were fitted using the Hill function

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

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

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

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

1. for the endogenous *RON* gene: $y_{\min} = -0.11$, $y_{\max} = 0.36$, EC50 = 0.93, $n_H = 17.4$
2. 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

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

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

The 95% confidence intervals found for the endogenous *RON* gene were:

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

and for the wt *RON* minigene:

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

Code availability. Code that was used to generate the presented data is available from the corresponding authors upon request.

Data availability. The sequencing data generated in this study are available from ArrayExpress under the accession numbers E-MTAB-6216 and E-MTAB-6217 (RNA-seq), E-MTAB-6219 (DNA-seq), E-MTAB-6220 and E-MTAB-6221 (iCLIP). All other data supporting the findings of this study are available from the corresponding authors on reasonable request.

Received: 8 January 2018 Accepted: 19 July 2018
Published online: 17 August 2018

References

1. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).

2. Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
3. Ellis, J. D. et al. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol. Cell* **46**, 884–892 (2012).
4. Yang, X. et al. Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* **164**, 805–817 (2016).
5. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18**, 437–451 (2017).
6. Sterne-Weiler, T. & Sanford, J. R. Exon identity crisis: disease-causing mutations that disrupt the splicing code. *Genome Biol.* **15**, 201 (2014).
7. Bonomi, S. et al. HnRNP A1 controls a splicing regulatory circuit promoting mesenchymal-to-epithelial transition. *Nucleic Acids Res.* **41**, 8665–8679 (2013).
8. Sebestyen, E. et al. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res.* **26**, 732–744 (2016).
9. Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R. A. & Skotheim, R. I. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene* **35**, 2413–2427 (2016).
10. Wahl, M. C., Will, C. L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701–718 (2009).
11. Barash, Y. et al. Deciphering the splicing code. *Nature* **465**, 53–59 (2010).
12. Fu, X. D. & Ares, M. Jr. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.* **15**, 689–701 (2014).
13. Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802–813 (2008).
14. Barash, Y. et al. AVISPA: a web tool for the prediction and analysis of alternative splicing. *Genome Biol.* **14**, R114 (2013).
15. Xiong, H. Y. et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
16. Julien, P., Minana, B., Baeza-Centurion, P., Valcárcel, J. & Lehner, B. The complete local genotype-phenotype landscape for the alternative splicing of a human exon. *Nat. Commun.* **7**, 11558 (2016).
17. Ke, S. et al. Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Res.* **28**, 11–24 (2018).
18. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711 (2015).
19. Zhang, K., Zhou, Y. Q., Yao, H. P. & Wang, M. H. Alterations in a defined extracellular region of the RON receptor tyrosine kinase promote RON-mediated motile and invasive phenotypes in epithelial cells. *Int. J. Oncol.* **36**, 255–264 (2010).
20. Collesi, C., Santoro, M. M., Gaudino, G. & Comoglio, P. M. A splicing variant of the RON transcript induces constitutive tyrosine kinase activity and an invasive phenotype. *Mol. Cell Biol.* **16**, 5518–5526 (1996).
21. Ghigna, C. et al. Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. *Mol. Cell* **20**, 881–890 (2005).
22. Wang, D., Shen, Q., Chen, Y. Q. & Wang, M. H. 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 (2004).
23. Zhou, Y. Q., He, C., Chen, Y. Q., Wang, D. & Wang, M. H. 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 (2003).
24. Chakedis, J. et al. Characterization of RON protein isoforms in pancreatic cancer: implications for biology and therapeutics. *Oncotarget* **7**, 45959–45975 (2016).
25. Mayer, S. et al. RON alternative splicing regulation in primary ovarian cancer. *Oncol. Rep.* **34**, 423–430 (2015).
26. Lefave, C. V. et al. Splicing factor hnRNPH drives an oncogenic splicing switch in gliomas. *EMBO J.* **30**, 4084–4097 (2011).
27. Moon, H. et al. A 2-nt RNA enhancer on exon 11 promotes exon 11 inclusion of the Ron proto-oncogene. *Oncol. Rep.* **31**, 450–455 (2014).
28. Nazim, M. et al. Competitive regulation of alternative splicing and alternative polyadenylation by hnRNP H and CstF64 determines acetylcholinesterase isoforms. *Nucleic Acids Res.* **45**, 1455–1468 (2017).
29. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
30. Llorian, M. et al. Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nat. Struct. Mol. Biol.* **17**, 1114–1123 (2010).
31. Xing, Y. & Lee, C. Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.* **7**, 499–509 (2006).
32. Shabalina, S. A., Spiridonov, N. A. & Kashina, A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Res.* **41**, 2073–2094 (2013).
33. Xing, Y. & Lee, C. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc. Natl Acad. Sci. USA* **102**, 13526–13531 (2005).
34. Giudice, G., Sanchez-Cabo, F., Torroja, C. & Lara-Pezzi, E. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database* baw035 (2016).
35. Papasaikas, P., Tejedor, J. R., Vigevani, L. & Valcárcel, J. Functional splicing network reveals extensive regulatory potential of the core spliceosomal machinery. *Mol. Cell* **57**, 7–22 (2015).
36. GTEx Consortium. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
37. Uren, P. J. et al. High-throughput analyses of hnRNP H1 dissects its multi-functional aspect. *RNA Biol.* **13**, 400–411 (2016).
38. Mueller, W. F., Larsen, L. S., Garibaldi, A., Hatfield, G. W. & Hertel, K. J. The silent sway of splicing by synonymous substitutions. *J. Biol. Chem.* **290**, 27700–27711 (2015).
39. Moon, H. et al. SRSF2 promotes splicing and transcription of exon 11 included isoform in Ron proto-oncogene. *Biochim. Biophys. Acta* **1839**, 1132–1140 (2014).
40. Savisaar, R. & Hurst, L. D. Estimating the prevalence of functional exonic splice regulatory information. *Hum. Genet.* **136**, 1059–1078 (2017).
41. Witten, J. T. & Ule, J. Understanding splicing regulation through RNA splicing maps. *Trends Genet.* **27**, 89–97 (2011).
42. Han, H. et al. Multilayered control of alternative splicing regulatory networks by transcription factors. *Mol. Cell* **65**, 539–553 (2017). e537.
43. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
44. Xiao, X. et al. Splice site strength-dependent activity and genetic buffering by poly-G runs. *Nat. Struct. Mol. Biol.* **16**, 1094–1100 (2009).
45. Gueroussov, S. et al. Regulatory expansion in mammals of multivalent hnRNP assemblies that globally control alternative splicing. *Cell* **170**, 324–339 (2017). e323.
46. Conlon, E. G. et al. The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains. *eLife* **5**, e17820 (2016).
47. Dardenne, E. et al. RNA helicases DDX5 and DDX17 dynamically orchestrate transcription, miRNA, and splicing programs in cell differentiation. *Cell Rep.* **7**, 1900–1913 (2014).
48. Decorsiere, A., Cayrel, A., Vagner, S. & Millevoi, S. 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 Dev.* **25**, 220–225 (2011).
49. Fisette, J. F., Montagna, D. R., Mihailescu, M. R. & Wolfe, M. S. A G-rich element forms a G-quadruplex and regulates BACE1 mRNA alternative splicing. *J. Neurochem.* **121**, 763–773 (2012).
50. Singh, B. & Eyras, E. The role of alternative splicing in cancer. *Transcription* **8**, 91–98 (2017).
51. Supek, F., Minana, B., Valcárcel, J., Gabaldon, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**, 1324–1335 (2014).
52. Lu, Y., Yao, H. P. & Wang, M. H. Multiple variants of the RON receptor tyrosine kinase: biochemical properties, tumorigenic activities, and potential drug targets. *Cancer Lett.* **257**, 157–164 (2007).
53. Gartner, J. J. et al. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proc. Natl Acad. Sci. USA* **110**, 13481–13486 (2013).
54. Gotea, V., Gartner, J. J., Qutob, N., Elnitski, L. & Samuels, Y. The functional relevance of somatic synonymous mutations in melanoma and other cancers. *Pigment. Cell Melanoma Res.* **28**, 673–684 (2015).
55. Jung, H. et al. Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* **47**, 1242–1248 (2015).
56. Yao, H. P., Zhou, Y. Q., Zhang, R. & Wang, M. H. MSP-RON signalling in cancer: pathogenesis and therapeutic potential. *Nat. Rev. Cancer* **13**, 466–481 (2013).
57. O'Toole, J. M. et al. Therapeutic implications of a human neutralizing antibody to the macrophage-stimulating protein receptor tyrosine kinase (RON), a c-MET family member. *Cancer Res.* **66**, 9162–9170 (2006).
58. Rauch, J. et al. c-Myc regulates RNA splicing of the A-Raf kinase and its activation of the ERK pathway. *Cancer Res.* **71**, 4664–4674 (2011).
59. Williams, R. et al. Amplification of complex gene libraries by emulsion PCR. *Nat. Methods* **3**, 545-550 (2006).
60. Sutandy, F. X. R., Hildebrandt, A. & König, J. Profiling the binding sites of RNA-binding proteins with nucleotide resolution using iCLIP. *Methods Mol. Biol.* **1358**, 175–195 (2016).
61. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

62. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
63. Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790–2791 (2013).
64. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11–33 (2013). 11 10.
65. Whitlock, M. C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **18**, 1368–1373 (2005).
66. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
67. Vivian, J. et al. Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* **35**, 314–316 (2017).
68. Mootha, V. K. et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
69. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).

Acknowledgements

The authors would like to thank the members of all participating labs for their support and discussion. We gratefully acknowledge the Institute of Molecular Biology Core Facilities for their support, especially the Genomics and the Bioinformatics Core Facilities, and the use of the Illumina NextSeq 500 instrument (INST 47/870-1 FUGG) as well as Teresa Maia (NMorais Lab, iMM) for assistance with TCGA data retrieval and analyses and Tina Han for help and technical support. We would like to thank Guiseppe Biamonti and Heiner Schaal for advice on the *RON* minigene. The results published here are in part based upon data generated by TCGA managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov. This work was funded by a joint DFG grant (ZA 881/2-1 to K.Z., KO 4566/4-1 to J.K. and LE 3473/2-1 to S.L.). K.Z. was also supported by the LOEWE program Ubiquitin Networks (Ub-Net) of the State of Hesse (Germany) and the Deutsche Forschungsgemeinschaft (SFB902 B13). N. Barbosa-Morais' laboratory is supported by EMBO (Installation Grant 3057) and Fundação para a Ciência e a Tecnologia, Portugal (FCT Investigator Starting Grant IF/00595/2014). S.L. acknowledges support by the German Federal Ministry of Research (BMBF; e:bio junior group program, FKZ: 0316196). The Institute of Molecular Biology (IMB) gGmbH is funded by the Boehringer Ingelheim Foundation.

Author contributions

S.B. established the high-throughput screening approach and performed most experiments. S.T.S. performed most bioinformatics analyses. M.E. and S.L. designed the mathematical modelling approach and performed the analyses. M.C.-L. annotated putative RBP binding sites and analysed mutation effects and synergistic interactions. M. S. contributed to the RNA sequence annotation. B.P.d.A. and N.L.B.-M. performed the analyses of the TCGA and GTEx data sets. F.X.R.S. performed iCLIP experiments, and L. S. did validation experiments. A.B. performed iCLIP and RNA-seq data processing as well as splice isoform quantification. S.E. and K.Z. supervised the bioinformatics analyses. J.K. conceived the project and supervised the experimental work. S.B., S.L., J.K. and K.Z. wrote the manuscript with help and comments from all co-authors.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-05748-7.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

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

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

SUPPLEMENTARY INFORMATION

Braun et al.

Content:

Supplementary Notes

In these Supplementary Notes, we describe how we used mathematical modelling to infer the effect of single mutations on the splicing outcome. We employed a two-step modelling strategy in which we first calculate changes in splicing reactions from the isoform frequency using a dynamical model (Supplementary Note 1). In the second step, we describe the splice change in each minigene variant harbouring multiple mutations as a linear combination of single mutation effects, and estimate these single effects using a regression approach (Supplementary Note 2). Finally, we compare single mutation effects for control and *HNRNPH* knockdown conditions to identify synergistic interactions between these two types of perturbations (Supplementary Note 3).

Supplementary Note 1: Dynamic model of splicing reactions

We modelled the dynamics of splicing using a set of ordinary differential equations, in which concentrations of transcript intermediates are determined by production and degradation terms. The precursor mRNA (pre-mRNA) x_0 is produced at a constant rate *c* and spliced into different splice products with linear kinetics and rates r_i, leading to

$$\frac{dx_0}{dt} = \mathrm{c} - (r_1 + r_2 + r_3 + r_4 + r_5 + r_6)x_0. \qquad (1)$$

Additional differential equations describe the dynamics of the spliced isoforms:

$$\frac{dx_i}{dt} = r_i x_0 - d_i x_i, i = 1,\ldots,6. \qquad (2)$$

where $x_1 \ldots, x_5$ are the number of transcripts representing the alternative exon (AE) inclusion, AE skipping, full intron retention (IR), first IR and second IR isoforms. The additional non-canonical isoforms that were also measured are integrated in the model together by the species x_6, collectively referred to as 'other'. Furthermore, d_i are the degradation rates of the different isoforms.

The steady state found by setting dx_i/dt to zero in Supplementary Equations 2 reads $x_i = (r_i/d_i)x_0$. The measured isoform frequencies p_i correspond in the model to the fractions of the transcripts x_i within the total mRNA:

$$p_i = \frac{x_i}{x_0+x_1+x_2+x_3+x_4+x_5+x_6}, i = 1,2,4,5,6. \qquad (3)$$

For the frequency p_3 of mRNA transcripts that exhibit the complete sequence, we sum up the number of mRNA transcripts with full intron retention x_3 and the number of unspliced pre-mRNA transcripts x_0, since these two species were experimentally not differentiated. Thus, we get

$$p_3 = \frac{x_0+x_3}{x_0+x_1+x_2+x_3+x_4+x_5+x_6}. \qquad (4)$$

At steady state, we obtain from Supplementary Equations 1-4

$$p_i = \frac{K_i}{K_1+K_2+K_3+K_4+K_5+K_6}, i = 1,\ldots,6. \qquad (5)$$

where we introduced the parameters $K_i = r_i/d_i$, $i = 1,2,4,5,6$ for the isoforms involving splicing and $K_3 = 1 + r_3/d_3$ for the unspliced full IR isoform.

We remark that due to the normalisation condition $\sum_{i=1}^{6} p_i = 1$, not all model parameters K_i can be determined from the experimental data, but only ratios of K_i with respect to a reference isoform. If we normalise all K_i by the AE inclusion rate, we can determine the ratios K_i/K_1, $i = 2,\ldots,6$ from the measured isoform frequencies via $K_i/K_1 = p_i/p_1$.

1.1 *In silico* simulation of competing splicing reactions

Supplementary Equation 5 reflects the non-linear nature of the splicing system: For example, a perturbation affecting the splicing parameter K_2 will affect all transcript isoform frequencies $p_i$if $K_2$is large compared to other parameters, but not otherwise. In contrast, the splice isoform ratios respond in the same way to a perturbation affecting the splicing parameter K_2, irrespective of the other parameter values.

We confirmed that perturbation-induced fold-changes in the isoform frequencies, but not isoform ratios, depend on the mutational background by numerically simulating the steady state of the splicing system (Supplementary Equations 1 and 2). Random mutagenesis (i.e., the varying mutational background) was mimicked by uniformly sampling parameters $\mathrm{c}, r_i, d_i, i = 1,\ldots,6$ in logarithmic space within the range [0.1,10], and calculating the steady state for 5,000 different realisations. Subsequently, each parameter set was additionally perturbed by decreasing the parameter r_2 at 20% of the sampled value (representing an additional mutation or knockdown), and the new steady state was calculated. As expected from the inspection of the steady states given in Supplementary Equation 5, the effect of the perturbation on splice isoforms frequencies is nonlinear and strongly depends on the specific parameter values, i.e., the mutational background (**Supplementary Fig. 7e**).

In contrast, the perturbation of r_2 has a linear effect on the splice isoform ratios in the sense that the fold-change between perturbed and unperturbed steady states is the same for all parameter sets (**Supplementary Fig. 7e**). Therefore, a mutation (or knockdown) affecting splicing kinetics induces the same fold change of an isoform ratio, irrespective of the presence of other mutations in the minigene. Thus, perturbation effects on splice isoform ratios show additive behaviour in log-space and are therefore more suitable for the regression approach described below.

Supplementary Note 2: Inference of single mutation effects

2.1 Calculation of single mutation effects by linear regression

By analysing the cumulative mutational effects in minigenes containing two or three mutations that are also present as single mutations in other minigenes, we found that the effects of single mutations on the above defined splicing rates are in general multiplicative (**Supplementary Fig. 4a**). Thus, we assume that the splicing parameter K_i for a minigene exhibiting a combination of several mutations is given by

$$K_i^{\text{mutated}} = K_i^{\text{wild type}} m_i^1 m_i^2 \dots m_i^n, \qquad (6)$$

where n is the number of the mutations in the minigene and m_i^k the effect of the k-th mutation on K_i.

Using the same normalisation to the AE inclusion isoform as in Supplementary Note 1, and taking the logarithm of Supplementary Equation 6 leads to

$$\sum_{k=1}^{n} \log \frac{m_i^k}{m_1^k} = \log \frac{p_i}{p_1} - \log \frac{p_i^{\text{wt}}}{p_1^{\text{wt}}}, i = 2, \dots, 6, \qquad (7)$$

where p_i and $p_i^{\text{wt}}, i = 1, \dots, 6$ are the isoform frequencies of the mutated and wild type (wt) *RON* minigenes, respectively. The isoform frequencies for the wt *RON* minigene were calculated as the median of the measured values across the minigenes exhibiting the wt sequence (586 minigene variants present in all RNA-seq replicates).

By considering all minigene variants together, we get a system of linear equations for the mutational effects $x_i(k) = \log\left(m_i^k / m_1^k\right), i = 2, \dots, 6, k = 1, \dots, N$, where N is the total number of mutations present in the dataset. For each of the five splice isoform ratios K_i/K_1, we get a separate system of linear equations which can be written in the matrix form:

$$A\mathbf{x}_i = \mathbf{b}_i, i = 2, \dots, 6. \qquad (8)$$

The entries of the matrix are $A(j,k) = 1/0$ if mutation k is present/absent in minigene variant j, respectively. The vectors $\mathbf{b}_i$ contain the experimental observations which are given by

$$b_i(j) = \log \frac{p_i^j}{p_1^j} - \log \frac{p_i^{\text{wt}}}{p_1^{\text{wt}}}, i = 2, \dots, 6, j = 1, \dots, m, \qquad (9)$$

with m being the number of unique combinations of mutations included in the calculation (between 4,467 and 4,771, depending on the cell line and replicate, see **Supplementary Table 1**).

Since any minigene contains only a few of the total unique 2,042 mutations present in the whole dataset (up to 18 mutations, with a mean of 3.7 mutations/minigene including insertions and deletions), the systems to be solved are sparse. To get the single mutational effects $x_i(k)$, we solved the systems in Supplementary Equations 8 in least square sense using Matlab subroutine lscov.

From the estimated mutational effects $x_i(k)$, a model prediction for the isoform frequencies p_i^k in a minigene containing the single mutation k can be made: For the single-mutation minigene, we would have

$$\frac{p_i^k}{p_1^k} = \frac{p_i^{\text{wt}}}{p_1^{\text{wt}}} \mathrm{e}^{x_i(k)}, i = 2, \dots, 6. \qquad (10)$$

By summing up Supplementary Equations 10 and using the normalisation condition $\sum_{i=1}^{6} p_i^k = 1$, we therefore get

$$\frac{1-p_1^k}{p_1^k} = \sum_{i=2}^{6} \frac{p_i^{\mathrm{wt}}}{p_1^{\mathrm{wt}}} \mathrm{e}^{x_i(k)}, \qquad (11)$$

which can be solved to find the AE inclusion isoform frequency p_1^k as a function of the single mutation effects:

$$p_1^k = \frac{1}{1+\sum_{i=2}^{6} \frac{p_i^{\mathrm{wt}}}{p_1^{\mathrm{wt}}} \mathrm{e}^{x_i(k)}} = \frac{p_1^{\mathrm{wt}}}{p_1^{\mathrm{wt}} + \sum_{i=2}^{6} p_i^{\mathrm{wt}} \mathrm{e}^{x_i(k)}}. \qquad (12)$$

Finally, the remaining isoform frequencies can be estimated via:

$$p_i^k = p_1^k \frac{p_i^{\mathrm{wt}}}{p_1^{\mathrm{wt}}} \mathrm{e}^{x_i(k)} = \frac{p_i^{\mathrm{wt}} \mathrm{e}^{x_i(k)}}{p_1^{\mathrm{wt}} + \sum_{i=2}^{6} p_i^{\mathrm{wt}} \mathrm{e}^{x_i(k)}}, i = 2, \ldots, 6. \quad (13)$$

Supplementary Equations 8 and 9 were used to infer the effects of single mutations from the data. Different replicates were treated separately, since both wt and mutated minigene variants showed systematic shifts in the measured frequencies between replicates. Thus, we always calculate mutational effects by comparing isoform frequencies of mutated and wt minigenes within the same replicate.

We note that the library also contains some minigenes with different barcodes but the same combination of mutations. We have included such combinations of mutations only once and attributed to them the median of the measured isoforms frequencies over the different minigenes with the same combination of mutations. Thus, the number of unique combinations of mutations is smaller than the number of mutated minigene variants (i.e. unique barcodes) in the dataset (**Supplementary Table 1**). Furthermore, we have excluded barcodes containing ambiguous mutations from the calculation.

The predictive power of our modelling approach was confirmed using cross-validation (also see Methods; **Supplementary Fig. 6**), and by comparing the inferred splicing outcome in response to single mutations (according to Supplementary Equations 11 and 12) to RT-PCR measurements of previously untested minigenes containing only these single mutations (see main manuscript; **Fig. 2d**).

It should be noted that certain minigenes had to be excluded from the linear regression procedure because they deviated from linear behaviour: (i) Minigenes simultaneously harbouring two splice site mutations: these minigenes show a very similar distribution of inclusion frequencies as minigenes containing only one of these mutations (**Supplementary Fig. 7f**). The median inclusion frequency of both, one- and two-splice-site-mutation minigenes, was non-zero (0.7%). The apparent lack-of-effect of secondary splice site mutations at non-zero inclusion frequencies contradicts their strong effect as isolated splice site mutations, and introduces strong inconsistencies and biases in linear regression. In our opinion, this observation hints to a constant background signal, e.g., due to leaky sequencing reads originating from other minigenes where inclusion is the predominant isoform. Therefore, we excluded minigenes exhibiting any two mutations at positions proximal to splice sites (positions 210-212, 295-297, 443-446, 522-524, 689-691). (ii) Minigenes with strong activation of cryptic splice sites: The activation of cryptic splice sites by mutations leads to the generation of a plethora of new splicing products ('other') which behave heterogeneously and cannot be considered in our model. Therefore, we performed first the regression on the complete dataset and subsequently excluded the minigenes containing

mutations that were predicted to exhibit an increased 'other' isoform frequency $p_6 > 4p_6^{\mathrm{wt}}$ in this first run. The threshold used for the exclusion of minigenes from the regression dataset was four times the median p_6^{wt} of the 'other' isoform frequency for the wt minigenes, and thus cell line and replicate-specific. The final calculation of mutational effects was performed on this reduced dataset (**Supplementary Table 1**). As an alternative approach to estimate the mutation effects of the excluded mutations, we calculated the median of isoform frequencies for all minigene variants harbouring the given mutation (**Supplementary Data 3**).

Depending on the replicate and cell line, between 3-9% of the unique combinations of mutations were excluded from the calculations based on the above criteria (**Supplementary Table 1**). Still, the effects of 94-97% of the mutations present in the library could be assessed by regression that covered almost the entire length of the minigene (all but 3-4 out of all 679 nucleotides in the minigene).

2.2 Comparative analysis of linear regression approaches

As described above, kinetic modelling suggested that fitting to splice isoform ratios is most suitable for linear regression. To support this claim, we tested two alternative regression approaches for the inference of single mutation effects, both of which were based on direct fitting to splice isoform frequencies. Reassuringly, our isoform ratio-based approach outperformed these alternative methods.

First, we assumed that the mutation effects add up at the level of splice isoform frequencies (not at the level of ratios). Thus, we used

$$\sum_{k=1}^{n} \log m_i^k = \log p_i - \log p_i^{\mathrm{wt}}, i = 1,\ldots,6 \qquad (14)$$

instead of Supplementary Equation 7 for the computation of the single mutation effects m_i^k. The corresponding isoform frequencies for the single mutation minigene k then read

$$p_i^k = p_i^{\mathrm{wt}} m_i^k, i = 1,\ldots,6. \qquad (15)$$

Supplementary Equation 14 was solved in least square sense using the Matlab subroutine fmincon with the constraint that all isoform frequencies in a single mutation background are bounded to unity, i.e., $\sum_{i=1}^{6} p_i^k = 1, k = 1,\ldots,N$. During cross-validation, predictions for new combinations of mutations were given by

$$p_i = p_i^{\mathrm{wt}} m_i^1 m_i^2 \ldots m_i^n, i = 1,\ldots,6, \qquad (16)$$

where $m_i^1,\ldots,m_i^n$ are the infered single mutation effects, and $1,\ldots,n$ the mutations present in the new combined minigene. We have compared the prediction performance of this method to the isoform ratio-based regression in 10-fold cross-validation and found that the use of isoform frequencies instead of ratios is inferior in terms of the prediction-data correlation. The corresponding Pearson correlation coefficients between model-predicted isoform frequencies and measured values for each predicted subset not used in fitting are visualised in **Supplementary Fig. 7b**. The predictions of the frequency-based model were in many cases also qualitatively wrong, as isoform frequencies of minigenes were not bounded to 1, thus leading to mispredictions $p_i > 1$, especially for the AE skipping isoform. In contrast, the calculation of isoform frequencies by renormalisation of the ratio-based regression results (Supplementary Equations 12 and 13) inherently prevents such biologically unreasonable mispredictions.

As a second alternative approach, we used multinomial logistic regression to infer the isoform frequencies in single-mutation minigenes. In this case, the dataset was categorised by introducing

six copies of each minigene and assuming as splicing output a different isoform for each of the copies. The data was weighted by the measured isoform frequencies, so each of the six samples corresponding to one minigene got as weight the measured frequency of its output isoform. We used the Python package scikit-learn with cross entropy loss and L2 regularisation to infer the probabilities for each splicing isoform for single-mutation minigenes and minigenes with new combinations of mutations. The prediction performance of this method in 10-fold cross-validation was also inferior to the isoform ratios-based regression, as shown in **Supplementary Fig. 7a**.

2.3 Estimation of the prediction error of the model

The prediction accuracy for a single mutation effect depends on the occurrence of the mutation in the minigene library. To quantitatively benchmark the accuracy of our model, we focused on ~600 mutations whose effects have been measured directly in our dataset as minigenes containing single mutations.

Benchmarking was done by eliminating the corresponding single-mutation minigenes from the dataset (separately for each of these mutations) and repeating the linear regression for the remaining data, or after removing further minigenes containing this mutation. This procedure allowed us to estimate how the prediction error depends on the occurrence of a mutation in the minigene library.

After calculating the single mutation effects, the isoform frequencies were estimated (Supplementary Equations 12 and 13) and the values for the mutations of interest were compared to the measured isoform frequencies of the single-mutation minigene. We find that the standard deviation of the prediction error (over all mutations and permutations) decreases with the occurrence of the mutation in the subset used in linear regression by $1/\sqrt{\text{occurrence}}$ (see main manuscript; **Fig. 2c**).

This relationship can also be proven analytically by exploiting the profile likelihood which characterises the measurement-compliant range for each parameter value in the model (Raue et al., 2009). The agreement of the experimental data $\mathbf{b}_i$ with the model simulations $\mathbf{x}_i$ is measured by the sum of squared residuals:

$$\chi_i^2(\mathbf{x}_i) = \|A\mathbf{x}_i - \mathbf{b}_i\|^2 = \sum_{j=1}^{m}[\sum_{k=1}^{N} A\,(j,k)x_i(k) - b_i(j)]^2. \qquad (17)$$

The optimal values $\widehat{\mathbf{x}_i}$ of the model parameters estimated by linear regression minimise the objective functions χ_i^2, thus we have

$$\boldsymbol{\nabla}\chi_i^2(\widehat{\mathbf{x}_i}) = 2(A\mathbf{x}_i - \mathbf{b}_i)^{\mathrm{T}}A = 0. \qquad (18)$$

The confidence interval for a certain parameter $x_i(k)$ can be derived from the curvature of the objective functions, for example by calculating the Hessian matrices $H_i = \boldsymbol{\nabla}^{\mathrm{T}}\boldsymbol{\nabla}\chi_i^2(\widehat{\mathbf{x}_i})$. We find

$$H_i = 2A^{\mathrm{T}}A. \qquad (19)$$

The matrix A indicates the presence/absence of a particular mutation in a particular minigene variant, i.e. $A(j,k) = 1$ if mutation k is found in minigene variant j and $A(j,k) = 0$ otherwise. We therefore get for the diagonal elements of $A^{\mathrm{T}}A$

$$(A^{\mathrm{T}}A)_{kk} = \sum_{j=1}^{m} A\,(j,k)A(j,k) = \text{occurrence}(k). \quad (20)$$

which is equal to the number of minigene variants that exhibit the mutation k. For the non-diagonal elements of $A^{\mathrm{T}}A$, we get

$(A^T A)_{kl} = \sum_{j=1}^{m} A\,(j,k)A(j,l) = \text{occurrence}(k,l), k \neq l,$ (21)

which is equal to the number of minigenes that simultaneously exhibit the mutations k and l.

Therefore, the Taylor expansion of the objective function χ_i^2 around the minimum $\chi_i^2(\hat{\mathbf{x}}_i)$ is up to the second order given by

$\chi_i^2(\mathbf{x}_i) = \chi_i^2(\hat{\mathbf{x}}_i) + \sum_{k=1}^{N} \text{occurrence}\,(k)[x_i(k) - \hat{x}_i(k)]^2 +$
$\sum_{k=1}^{N}\sum_{l=1,l\neq k}^{N} \text{occurrence}\,(k,l)[x_i(k) - \hat{x}_i(k)][x_i(l) - \hat{x}_i(l)].$
(22)

Supplementary Equation 22 can be used to find the confidence intervals for the model parameters calculated by regression. For a given value of the parameter $x_i(k_0) = \hat{x}_i(k_0) + \delta_0$, the remaining parameters $x_i(k \neq k_0)$ can be refitted. Introducing $x_i(k) = \hat{x}_i(k) + \delta_k, k \neq k_0$ and using Supplementary Equation 18 for $k \neq k_0$ leads to a reduced system of equations for $\delta_{k\neq k_0}$, that can be written in matrix form as

$$C_{k_0}\left(\delta_1,\ldots,\delta_{k_0-1},\delta_{k_0+1},\ldots,\delta_N\right)^{\mathrm{T}} = -\mathbf{c}_{k_0}^{\mathrm{T}}\delta_0. \qquad (23)$$

Thereby, the symmetric matrix C_{k_0} is found by deleting the k_0th row and column from $A^{\mathrm{T}}A$, thus

$$C_{k_0} = (A^{\mathrm{T}}A)(k,l), k \neq k_0, l \neq k_0. \qquad (24)$$

Furthermore, the vector $\mathbf{c}_{k_0}$ contains the nondiagonal elements of the k_0th row of $A^{\mathrm{T}}A$:

$\mathbf{c}_{k_0} = [\text{occurrence}(k_0,1),\ldots.,\text{occurence}(k_0,k_0-1),\text{occurrence}(k_0,k_0+$
$1),\ldots.,\text{occurrence}(k_0,N)]^T.$ (25)

Solving Supplementary Equation 23 leads to the optimal values for the parameters $\delta_{k\neq k_0}$:

$$\left(\delta_1,\ldots,\delta_{k_0-1},\delta_{k_0+1},\ldots,\delta_N\right)^{\mathrm{T}} = -C_{k_0}^{-1}\mathbf{c}_{k_0}\delta_0. \qquad (26)$$

Introducing these solutions in Supplementary Equation 22 and regrouping the terms gives us

$\chi_i^2(\delta_0) = \chi_i^2(\hat{\mathbf{x}}_i) + \text{occurrence}(k_0)\delta_0^2 + \sum_{k=1,k\neq k_0}^{N} \text{occurrence}\,(k_0,k)\delta_0\delta_k +$
$\sum_{k=1,k\neq k_0}^{N}\sum_{l=1,l\neq k_0}^{N} \text{occurrence}\,(k,l)\delta_l\delta_k.$ (27)

By using the above notations in Supplementary Equations 23 and 25 as well as Supplementary Equation 26, we find

$$\sum_{k=1,k\neq k_0}^{N} \text{occurrence}\,(k_0,k)\delta_0\delta_k = \mathbf{c}_{k_0}^{\mathrm{T}}\delta_0\left[-C_{k_0}^{-1}\mathbf{c}_{k_0}\delta_0\right] = -\mathbf{c}_{k_0}^{\mathrm{T}}C_{k_0}^{-1}\mathbf{c}_{k_0}\delta_0^2 \qquad (28)$$

and

$\sum_{k=1,k\neq k_0}^{N}\sum_{l=1,l\neq k_0}^{N} \text{occurrence}\,(k,l)\delta_l\delta_k = \left(\delta_{k\neq k_0}\right)^{\mathrm{T}}C_{k_0}\left(\delta_{k\neq k_0}\right) = \left[C_{k_0}^{-1}\mathbf{c}_{k_0}\right]^T C_{k_0}C_{k_0}^{-1}\mathbf{c}_{k_0}\delta_0^2 =$
$\mathbf{c}_{k_0}^{\mathrm{T}}C_{k_0}^{-1}\mathbf{c}_{k_0}\delta_0^2.$ (29)

where we used the symmetry $C_{k_0}^{\mathrm{T}} = C_{k_0}$. Introducing Supplementary Equations 28 and 29 in Supplementary Equation 27 finally gives us the variation of the objective function with δ_0:

$$\chi_i^2(\delta_0) = \chi_i^2(\hat{\mathbf{x}}_i) + \text{occurrence}(k_0)\delta_0^2. \qquad (30)$$

Supplementary Equation 30 defines a parable with the minimal value $\chi^2(\hat{\mathbf{x}}_i)$ having the curvature $2\text{occurrence}(k_0)$. Thus, the more frequent the mutation k_0 is in the dataset, the steeper is the

parable and more constrained is the model parameter corresponding to this mutation. Setting a confidence threshold th for the objective function, e.g. imposing $\chi_i^2(\mathbf{x}_i) < \chi_i^2(\hat{\mathbf{x}}_i) + \mathrm{th}$, defines a confidence interval with respect to variation of the parameter $x_i(k_0)$ given by

$$|x_i(k_0) - \hat{x}_i(k_0)| < \sqrt{\frac{\mathrm{th}}{\mathrm{occurrence}(k_0)}}, \qquad (31)$$

which confirms the result obtained numerically by validation with the single-mutation minigenes (see main manuscript; **Fig. 2c**).

Supplementary Note 3: Model analysis of *HNRNPH* knockdown effects

We compared the effect of *HNRNPH* knockdown (KD) on wt and mutant minigene variants to identify synergistic interactions between both types of perturbations that may hint to the strengthening or weakening of HNRNPH binding sites by mutations (**Fig. 5a**). Using linear regression, we sought to trace back these synergistic interactions between mutations and *HNRNPH* KD to the single mutation level.

We initially checked the validity of our splice rate model (**Supplementary Fig. 3a**; see Supplementary Note 1) for the *HNRNPH* KD data: In the primary data, the fold-change in each isoform frequency upon *HNRNPH* KD is not stable and depends on the baseline value of the mutated minigene variant under non-targeting control conditions (**Supplementary Fig. 13a**). This can be understood from Supplementary Equation 5, in which a KD affecting a splice rate K_i has a strong (linear) effect or a weak (less than linear) effect depending on how K_i relates to the other competing splice rates $K_{j\neq i}$. To correct for this effect and to facilitate linear regression modelling, we therefore employed ratios of splice isoform frequencies, which show a similar effect (fold-change) of the *HNRNPH* KD for the majority of minigenes (**Fig. 5b**, right, and **Supplementary Fig. 13b**). This can be explained as follows: If for all minigenes, the splice parameters in the *HNRNPH* KD $\bar{K}_i$ relate to the control splice parameters K_i by the same, isoform and KD-specific factors α_i

$$\bar{K}_i = \alpha_i K_i, i = 1,\ldots,6, \qquad (32)$$

then the isoform ratios $\bar{p}_i/\bar{p}_1$ and p_i/p_1 in *HNRNPH* KD and control conditions will also be related by the factors α_i/α_1, independent of splice-rate competition effects. This suggests that the splice model is able to correct for nonlinearities in the data, thereby facilitating the identification of true synergistic interactions.

Large discrepancies from the linear behaviour in Supplementary Equation 32 imply that a particular minigene variant reacts differently than the majority of the library to the *HNRNPH* KD, pointing to a change in a binding site of HNRNPH itself or other means that enhance or repress its function (positive or negative synergy). We used modelling to identify such synergistic interactions of sequence mutations and *HNRNPH* KD at single-nucleotide resolution. Instead of calculating KD-induced fold-changes per minigene, we employed linear regression modelling to infer single mutation effects before comparing KD effects on wt minigenes and individual mutations.

By the linear regression setup (see Supplementary Note 2), we can determine the mutational effects of single mutations in control $x_i(k)$ and KD $\bar{x}_i(k)$ conditions. According to our model, we have

$$\frac{\bar{K}_i^k}{\bar{K}_1^k} = \frac{\bar{K}_i^{\mathrm{wt}}}{\bar{K}_1^{\mathrm{wt}}} \mathrm{e}^{\bar{x}_i(k)}, \frac{K_i^k}{K_1^k} = \frac{K_i^{\mathrm{wt}}}{K_1^{\mathrm{wt}}} \mathrm{e}^{x_i(k)}, i = 2,\ldots 6, k = 1,\ldots,N. \qquad (33)$$

Using Supplementary Equation and assuming the same KD factors α_i on both mutated and wt minigenes, we get

$$\frac{\bar{K}_i^k}{\bar{K}_1^k} = \frac{\alpha_i}{\alpha_1}\frac{K_i^k}{K_1^k}, \frac{\bar{K}_i^{\mathrm{wt}}}{\bar{K}_1^{\mathrm{wt}}} = \frac{\alpha_i}{\alpha_1}\frac{K_i^{\mathrm{wt}}}{K_1^{\mathrm{wt}}}, i = 2,\ldots 6, k = 1,\ldots,N. \quad (34)$$

From Supplementary Equations 33 and 34, we find

$$\frac{\bar{K}_i^k}{\bar{K}_1^k} = \frac{\bar{K}_i^{\mathrm{wt}}}{\bar{K}_1^{\mathrm{wt}}} \mathrm{e}^{\bar{x}_i(k)} = \frac{\alpha_i}{\alpha_1}\frac{K_i^{\mathrm{wt}}}{K_1^{\mathrm{wt}}} \mathrm{e}^{\bar{x}_i(k)}, \qquad (35)$$

and

$$\frac{\bar{K}_i^k}{\bar{K}_1^k} = \frac{\alpha_i}{\alpha_1}\frac{K_i^k}{K_1^k} = \frac{\alpha_i}{\alpha_1}\frac{K_i^{\mathrm{wt}}}{K_1^{\mathrm{wt}}}\mathrm{e}^{x_i(k)}. \qquad (36)$$

By comparing Supplementary Equations 35 and 36 we conclude that the mutation effects $x_i(k)$ should not change significantly between control and KD conditions, e.g.

$$\bar{x_i}(k) = x_i(k) \qquad (37)$$

should be valid for all mutations present in minigenes that react to the *HNRNPH* KD similarly to the wt minigenes. By contrast, above-average deviations from Supplementary Equation 37 are expected for mutations present in minigenes that react non-linearly to the KD.

We used *z*-scores to quantify to what extent a mutation shows different effects under control and *HNRNPH* KD conditions:

$$z_i^{\mathrm{kd}}(k) = \frac{x_i(k) - \bar{x_i}(k)}{\delta_i^{\mathrm{wt}}}, i = 2,\ldots,6, k = 1,\ldots,N. \qquad (38)$$

Due to the additivity of perturbation effects, this *z*-score can be interpreted to reflect differential *HNRNPH* KD effects in wt vs. single mutant backgrounds, allowing us to formulate positive and negative synergy as stronger or weaker KD responses in mutants compared to wt (see **Fig. 5b** and main text). In these *z*-scores, the difference between *HNRNPH* KD and control behaviour is normalised by the variation of KD effects in the wt minigenes to correct for experimental noise: Based on the wt minigenes present in both control and KD datasets, the standard deviation for the wt difference between control and KD conditions can be calculated by

$$\delta_i^{\mathrm{wt}} = \mathrm{STD}\left\{\log\frac{p_i^{\mathrm{wt}}}{p_1^{\mathrm{wt}}} - \log\frac{\bar{p}_i^{\mathrm{wt}}}{\bar{p}_1^{\mathrm{wt}}}\right\}, i = 2,\ldots 6. \qquad (39)$$

When calculating synergies between mutations and knockdowns using *z*-scores, the results may become unstable if one of the two perturbations already induces a close-to-maximal effect on the splice isoform frequencies. In fact, when analysing the variation of *z*-scores over the three replicates, we find that mutations that shift the inclusion frequency close to 0% increase the error in synergy *z*-score calculations and are thus potentially problematic. We show this effect in **Supplementary Fig. 7g**, in which we plot the uncertainty of the synergy *z*-score (standard deviation over the three replicates) against the (inferred) inclusion frequency in a single-mutation minigene.

Supplementary Tables 1 - 4

Supplementary Table 1: Information on the input and output data of the mathematical model on the different RNA-seq replicates.

	HEK293T			MCF7 – control			MCF7 – *HNRNPH* KD		
	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
General information									
Internal ID	imb_koenig_2015_13			imb_koenig_2016_07			imb_koenig_2016_08		
Initial reads	17,261,922	19,501,750	18,166,077	19,103,473	17,132,590	22,075,639	17,956,862	19,551,048	21,930,173
Minigenes	5,697	5,645	5,623	5,680	5,680	5,684	5,686	5,700	5,683
Wt minigenes	586	586	586	586	586	586	586	586	586
Unique mutation comb.	4,938	4,886	4,865	4,923	4,923	4,927	4,929	4,942	4,926
Model input									
Comb. used by model	4,571	4,467	4,472	4,672	4,678	4,650	4,763	4,771	4,739
Excluded comb.	367 (7%)	419 (9%)	393 (8%)	251 (5%)	245 (5%)	277 (6%)	166 (3%)	171 (3%)	187 (4%)
Singlets	606	603	603	612	608	609	613	613	613
Doublets	1,009	1,000	1,001	1,023	1,025	1,021	1,034	1,032	1,030
Triplets	869	859	858	891	888	886	910	909	905
Model output									
Mutations in dataset	2,042	2,033	2,032	2,038	2,040	2,041	2,039	2,042	2,040
Estimated mutation effects	1,942 (95%)	1,915 (94%)	1,915 (94%)	1,957 (96%)	1,956 (96%)	1,957 (96%)	1,972 (97%)	1,974 (97%)	1,974 (97%)
Positions in dataset	680	679	680	680	680	680	680	680	680
Estimated position effects	676 (99.4%)	675 (99.6%)	676 (99.4%)	677 (99.6%)	677 (99.6%)	677 (99.6%)	677 (99.6%)	677 (99.6%)	677 (99.6%)

For each RNA-seq replicate (Rep), the internal library identifier is given together with information on the number of total and wild type (wt) minigene variants detected in each dataset, the number of unique mutation combinations (differentiated into those used or excluded from the model analysis; see Supplementary Note 2) as well as the used single-/double-/triple-mutation combinations (singlets/doublets/triplets, respectively). Output information summarises the mutation and position effects that can be estimated by the model in relation to all mutations and mutated positions represented in each dataset.

Supplementary Table 2. Summary of splicing-effective mutations and synergistic interactions with *HNRNPH* knockdown per region in HEK293T and MCF7 cells.

		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%)

(a) 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%). Mutation effects are additionally given for each individual isoform. AE - alternative exon; IR - intron retention. Related to Fig. 2e and Supplementary Fig. 8.

Supplementary Table 2 (continued). Summary of splicing-effective mutations and synergistic interactions with *HNRNPH* knockdown per region in HEK293T and MCF7 cells.

		Exon 10	Intron 10	Exon 11	Intron 11	Exon 12	Intron 12	Total
MCF7	**Mutations**	555	261	441	240	498	42	2037
	Measured	501 (90.3%)	229 (87.7%)	386 (87.5%)	196 (81.7%)	440 (88.4%)	35 (83.3%)	1787 (87.7%)
	Any isoform >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%)

(b) Splicing effective mutations (top) and positions (bottom) in MCF7 cells. Format as in (a).

Supplementary Table 2 (continued). Summary of splicing-effective mutations and synergistic interactions with *HNRNPH* knockdown per region in HEK293T and MCF7 cells.

		Exon 10	Intron 10	Exon 11	Intron 11	Exon 12	Intron 12	Total
MCF7 – synergistic interactions with *HNRNPH* knockdown	**Mutations**	555	261	441	240	498	42	2037
	Measured	501 (90.3%)	229 (87.7%)	385 (87.3%)	196 (81.7%)	440 (88.4%)	35 (83.3%)	1786 (87.7%)
	Any isoform $\|z\| > 2$	**70** (14%)	**35** (15.3%)	**135** (35.1%)	**51** (26%)	**58** (13.2%)	**5** (14.3%)	**354** (19.8%)
	AE skipping	37	21	100	17	39	1	215
	First IR	8	3	5	4	8	1	29
	Second IR	10	6	6	4	7	0	33
	Full IR	30	20	47	14	18	3	132
	Other	21	17	54	30	9	1	132
	Any isoform $\|z\| > 3$	**44** (8.8%)	**25** (10.9%)	**89** (23.1%)	**31** (15.8%)	**31** (7%)	**2** (5.7%)	**222** (12.4%)
	Any isoform $\|z\| > 5$	**10** (2%)	**3** (1.3%)	**35** (9.1%)	**7** (3.6%)	**11** (2.5%)	**0** (0%)	**66** (3.7%)
	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 $\|z\| > 2$	**61** (33%)	**28** (32.2%)	**93** (63.3%)	**38** (48.7%)	**54** (32.5%)	**4** (28.6%)	**278** (41.1%)
	Any isoform $\|z\| > 3$	**42** (22.7%)	**23** (26.4%)	**61** (41.5%)	**25** (32.1%)	**31** (18.7%)	**2** (14.3%)	**184** (27.2%)
	Any isoform $\|z\| > 5$	**10** (5.4%)	**3** (3.4%)	**27** (18.4%)	**7** (9%)	**11** (6.6%)	**0** (0%)	**58** (8.6%)

(c) Synergistic interactions between mutations (top) or positions (bottom) and *HNRNPH* knockdown in MCF7 cells. Same format as in (a). Interactions for any isoform are reported at different absolute *z*-score cutoffs ($|z|$ >2, >3 and >5). Note that synergistic interactions are calculated from ratios of a given isoform over AE inclusion, so no synergistic interactions are given for AE inclusion. Related to Fig. 5c and Supplementary Fig. 12c.

Supplementary Table 3: Association of *HNRNPH2* expression with *RON* exon 11 inclusion levels in different TCGA cohorts. Related to Fig. 3f.

TCGA cohort	# samples	Spearman correlation	*HNRNPH2* variance	*P*-value	FDR	Significance (FDR < 0.05)
BRCA	778	-0.28	0.24	1.5e-15	3.9e-14	TRUE
LUAD	485	-0.25	0.15	3.5e-08	4.6e-07	TRUE
COAD	323	-0.27	0.16	8.0e-07	6.9e-06	TRUE
READ	103	-0.41	0.16	2.0e-05	1.3e-04	TRUE
ESCA	181	-0.29	0.13	7.3e-05	3.8e-04	TRUE
PAAD	163	-0.29	0.08	2.2e-04	9.5e-04	TRUE
LUSC	315	-0.2	0.12	3.8e-04	1.4e-03	TRUE
CESC	248	-0.2	0.16	1.3e-03	4.2e-03	TRUE
STAD	414	-0.14	0.13	3.2e-03	9.2e-03	TRUE
HNSC	455	-0.13	0.16	4.3e-03	1.1e-02	TRUE
THYM	51	-0.37	0.10	7.0e-03	1.7e-02	TRUE
OV	178	-0.18	0.14	1.5e-02	3.3e-02	TRUE
KIRC	11	-0.56	0.25	7.0e-02	1.4e-01	FALSE
PRAD	10	-0.52	0.03	1.3e-01	2.4e-01	FALSE
TGCT	17	0.36	0.10	1.5e-01	2.6e-01	FALSE
BLCA	251	-0.086	0.17	1.7e-01	2.8e-01	FALSE
THCA	282	-0.077	0.04	1.9e-01	2.9e-01	FALSE
KIRP	47	-0.19	0.10	2.1e-01	3.0e-01	FALSE
CHOL	28	-0.2	0.09	3.1e-01	4.2e-01	FALSE
LIHC	24	-0.21	0.15	3.2e-01	4.2e-01	FALSE
SKCM	59	-0.11	0.14	4.0e-01	5.0e-01	FALSE
KICH	4	-0.6	0.18	4.2e-01	5.0e-01	FALSE
UCEC	61	-0.09	0.19	4.9e-01	5.5e-01	FALSE
SARC	19	0.076	0.37	7.6e-01	8.2e-01	FALSE
DLBC	3	0.5	0.04	1	1	FALSE
LAML	3	0.5	0.08	1	1	FALSE
GBM	1	NA	NA	NA	NA	NA

Cancer types: BLCA, Bladder Urothelial Carcinoma; BRCA, Breast Invasive Carcinoma; CESC, Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma; CHOL, Cholangiocarcinoma; COAD, Colon Adenocarcinoma; DLBC, Lymphoid Neoplasm Diffuse Large B-cell Lymphoma; ESCA, Esophageal Carcinoma; GBM, Glioblastoma Multiforme; HNSC, Head-Neck Squamous Cell Carcinoma; KICH, Kidney Chromophobe; KIRC, Kidney Renal Clear Cell Carcinoma; KIRP, Kidney Renal Papillary Cell Carcinoma; LAML, Acute Myeloid Leukemia; LIHC, Liver Hepatocellular Carcinoma; LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Cell Carcinoma; OV, Ovarian Serous Cystadenocarcinoma; PAAD, Pancreatic Adenocarcinoma; PRAD, Prostate Adenocarcinoma; READ, Rectum Adenocarcinoma; SARC, Sarcoma; SKCM, Skin Cutaneous Melanoma; STAD, Stomach Adenocarcinoma; TGCT, Testicular Germ Cell Tumours; THCA, Thyroid Carcinoma; THYM, Thymoma; UCEC, Uterine Corpus Endometrial Carcinoma.

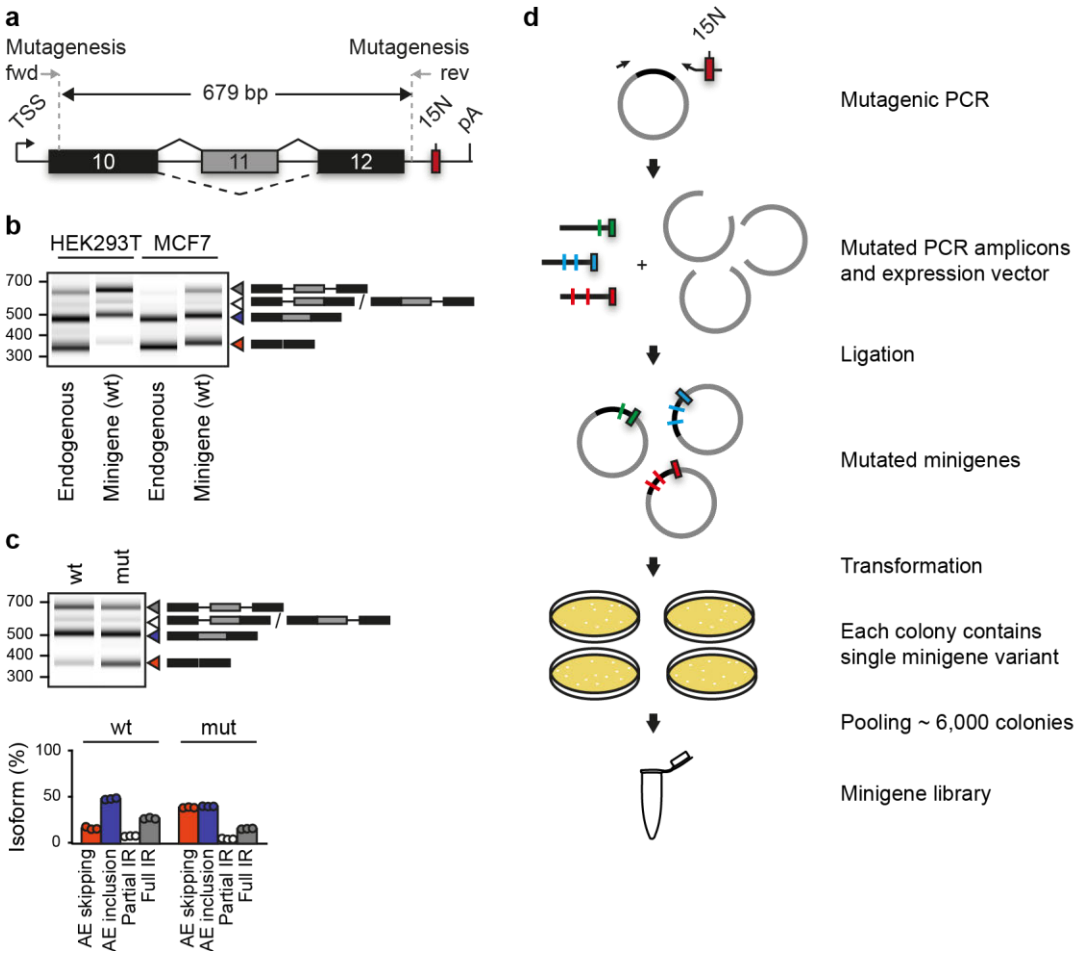
Supplementary Table 4: Oligonucleotides used in this study.

Name	Sequence (5'-3')	Purpose
minigene_cloning_fwd	CCCAAGCTTTGTGAGAGGCAGCTTCCAGA	Cloning of wt *RON* minigene
minigene_cloning_rev	CAGTCTAGANNNNNNNNNNNNNNNGGATCCGCCATTGGTTGGGGGTAGG-GGCTGATTAAAGGTAGG	Cloning of wt *RON* minigene
BamHI_HNRNPH1_fwd	catGGATCCaccatgatgttgggcacggaagg	Cloning of HNRNPH1 overexpression construct
XbaI_HNRNPH1_rev	cattctagactatgcaatgtttgattgaaaatc	Cloning of HNRNPH1 overexpression construct
RT-PCR_minigene_fwd	TGCCAACCTAGTTCCACTGA	RT-PCR for *RON* minigene
RT-PCR_minigene_rev	GCAACTAGAAGGCACAGTCG	RT-PCR for *RON* minigene
RT-PCR_endo_fwd	CCTGAATATGTGGTCCGAGACCCCCAG	RT-PCR for endogenous *RON* gene
RT-PCR_endo_rev	CTAGCTGCTTCCTCCGCCACCAGTA	RT-PCR for endogenous *RON* gene
RON A	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNNCTATAGGGAGACCCAAGCTT	Illumina fwd sequencing primer for DNA-seq
RON B	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNNGTTCCACTGAAGCCTGAG	Illumina fwd sequencing primer for DNA-seq and RNA-seq
RON C	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNNAGCTGCCAGCACGAGTTC	Illumina fwd sequencing primer for DNA-seq
RON D	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNNGAATCTGAGTGCCCGAGG	Illumina fwd sequencing primer for DNA-seq
RON E	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNNctactggctggtcctcatga	Illumina fwd sequencing primer for DNA-seq
P5 SOLEXA RON	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNATAGAATAGGGCCCTCTAGA	Illumina rev sequencing primer for DNA-seq and RNA-seq
RT1	NNAATANNNAGATCGGAAGAGCGTCGTGGATCCTGAACCGC	RT primer HNRNPH iCLIP for wt replicate 1
RT2	NNTTTCNNNAGATCGGAAGAGCGTCGTGGATCCTGAACCGC	RT primer HNRNPH iCLIP for wt replicate 2
RT3	NNCGATNNNAGATCGGAAGAGCGTCGTGGATCCTGAACCGC	RT primer HNRNPH iCLIP for wt replicate 3
RT4	NNTTCTNNNAGATCGGAAGAGCGTCGTGGATCCTGAACCGC	RT primer HNRNPH iCLIP for G305A replicate 1
RT5	NNCTCGNNNAGATCGGAAGAGCGTCGTGGATCCT	RT primer HNRNPH iCLIP

	GAACCGC	for G305A replicate 2
RT6	NNACGCNNNAGATCGGAAGAGCGTCGTGGATCCT GAACCGC	RT primer HNRNPH iCLIP for G305A replicate 3
RT7	NNTTCTNNNAGATCGGAAGAGCGTCGTGGATCCT GAACCGC	RT primer HNRNPH iCLIP for G331C replicate 1
RT8	NNGGCGNNNAGATCGGAAGAGCGTCGTGGATCCT GAACCGC	RT primer HNRNPH iCLIP for G331C replicate 2
RT9	NNTGTGNNNAGATCGGAAGAGCGTCGTGGATCCT GAACCGC	RT primer HNRNPH iCLIP for G348C replicate 1
RT10	NNGTATNNNAGATCGGAAGAGCGTCGTGGATCCT GAACCGC	RT primer HNRNPH iCLIP for G348C replicate 2

Oligonucleotides were purchased either from Sigma-Aldrich or Integrated DNA Technologies.

Supplementary Figures 1 - 16



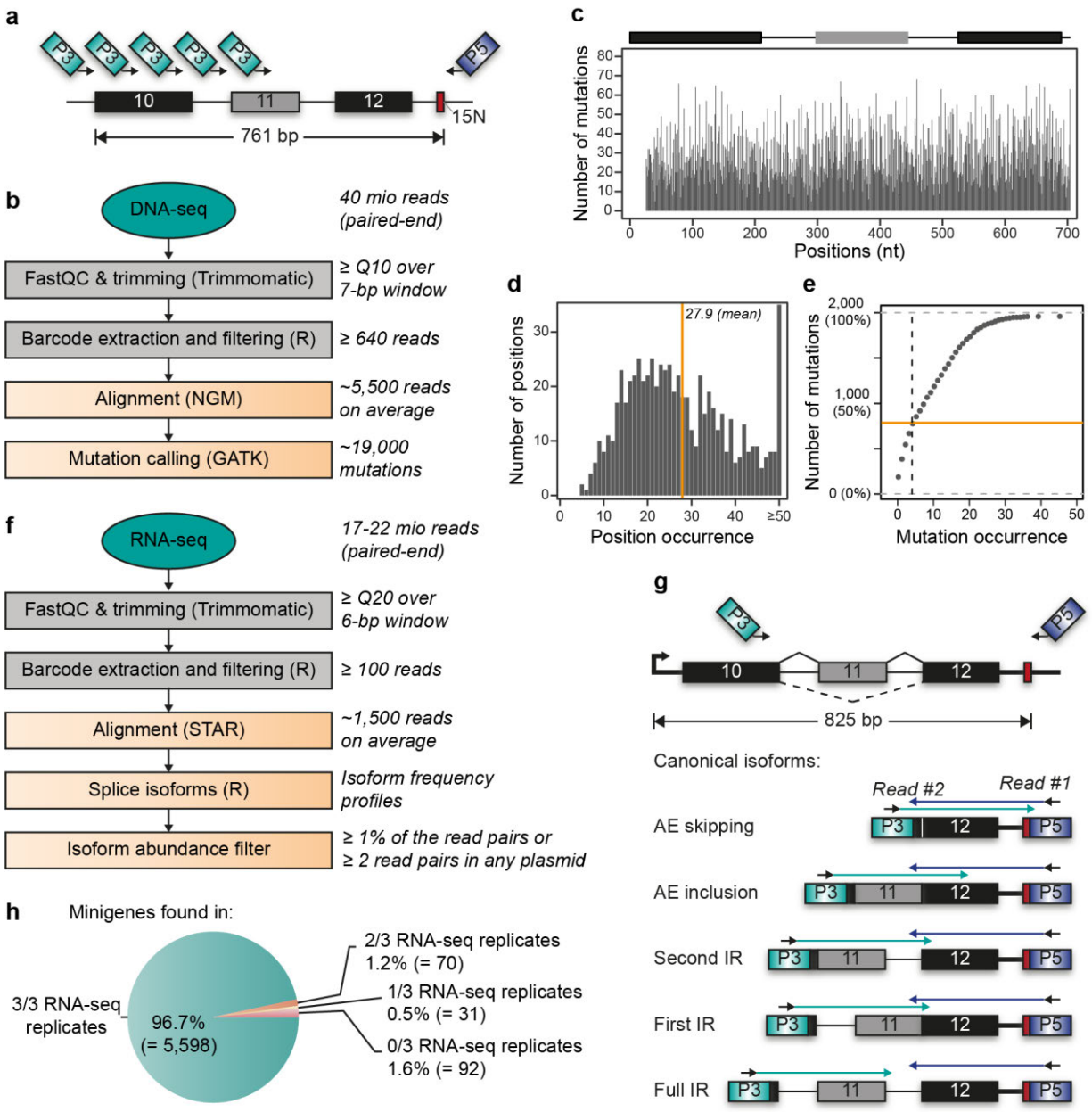
Supplementary Figure 1: Random mutagenesis generates a mutated *RON* minigene library. Related to Fig. 1a.

(a) The *RON* minigene harbours genomic sequence of the *RON* gene (*MST1R*, ENSG00000164078) including alternative exon 11 with the complete flanking introns and constitutive exons 10 and 12 (chr3: 49,933,098 - 49,933,837, GRCh37/hg19). Mutagenesis of a 679 bp region was performed using error-prone PCR and indicated forward (fwd) and reverse (rev) primers. TSS, transcriptional start site; pA, polyadenylation site; 15N, the 15-nt barcode as a unique identifier of each minigene variant.

(b) The wild type (wt) *RON* minigene gives rise to the same splicing isoforms as the endogenous *RON* gene in HEK293T and MCF7 cells. Gel-like representation of capillary electrophoresis of PCR products from semiquantitative RT-PCR monitoring *RON* exon 11 inclusion. Note that different primer combinations were used to differentiate between the endogenous *RON* gene and the *RON* wt minigene (**Supplementary Table 4**), resulting in a 52-bp difference in the RT-PCR products for the same isoforms.

(c) Introducing a previously published triple mutation[3] into the *RON* minigene (T565A, G566T, G569A; mut) triggers the expected splicing response. Gel-like representation of RT-PCR products from HEK293T cells as in (b). Bar diagram below shows quantification of isoform frequencies (in %) for alternative exon (AE) inclusion and skipping, as well as partial and full intron retention (IR). Individual data points from three independent biological replicates are displayed. Note that partial IR refers to the sum of first IR and second IR isoforms that cannot be discriminated in the RT-PCR analysis.

(d) Schematic overview of the experimental procedure to generate the mutated minigene library. Mutagenic PCR amplification of the wt *RON* minigene creates mutated amplicons that were ligated into the expression vector to obtain the mutated minigene library. The reverse primer used in the mutagenic PCR carries a 15-nt random sequence (15N) that is included as a unique identifier into each minigene variant. See Methods for details. Coloured vertical bars schematically indicate point mutations.



Supplementary Figure 2: Mutations and splicing products from the minigene library are characterised by high-throughput DNA and RNA sequencing. Related to Fig. 1.

(a) Schematic of amplicons for paired-end DNA sequencing. Reverse primer binds downstream of 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. Forward primers introduce P3 (Read #2).

(b) Bioinformatics workflow for DNA-seq analysis to characterise mutations. Quality control and trimming was performed with FastQC and Trimmomatic, respectively, followed by custom scripts (in R) to extract 15-nt barcode and filter for minigenes with ≥640 read pairs. Reads were aligned to wt *RON* minigene sequence using NextGenMap (NGM), and mutation calling was done using HaplotypeCaller tool from Genome Analysis Toolkit (GATK). See Methods for details.

(c) 18,948 point mutations evenly distribute across the *RON* minigene. Bar diagram showing the number of minigene variants (out of 5,791) harbouring a mutation in a given position.

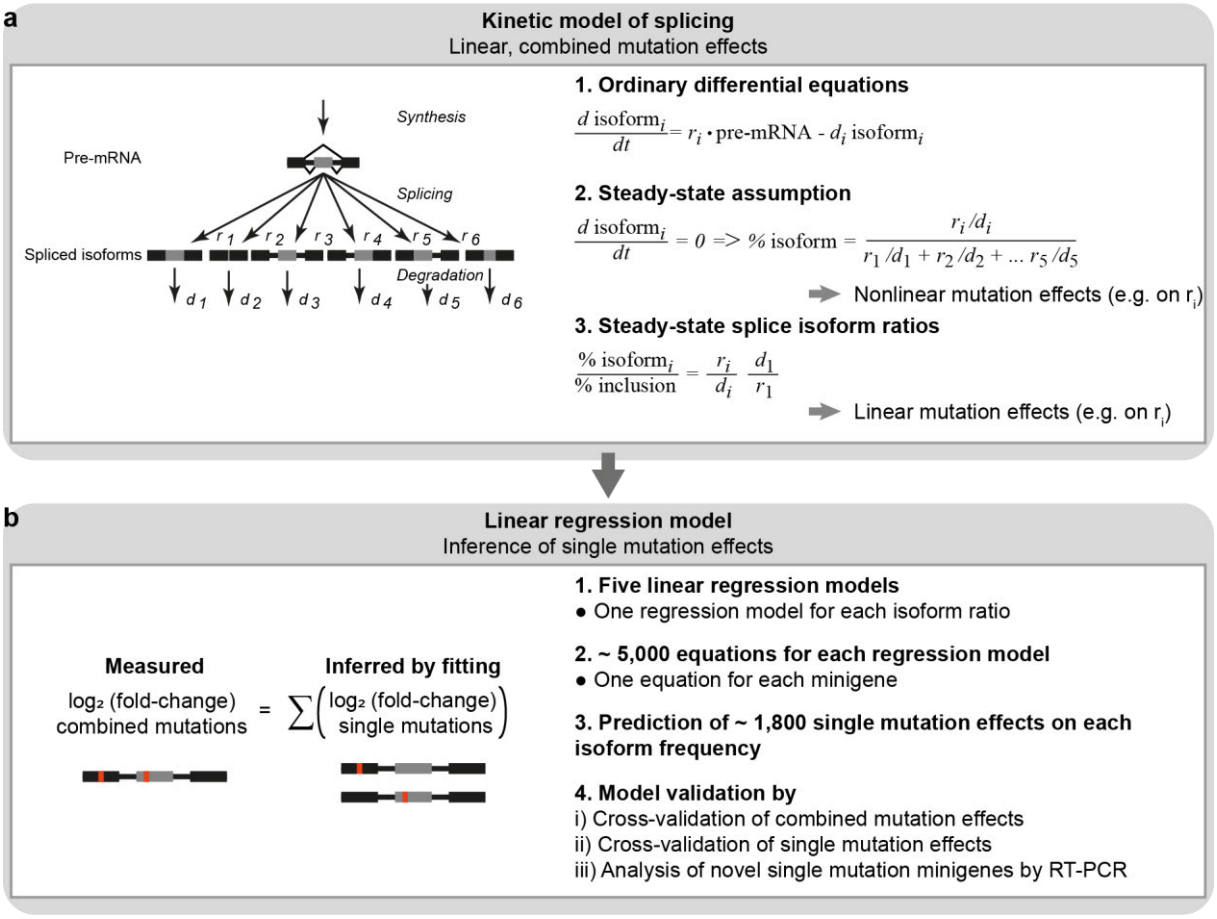
(d) Each position is on average mutated in 28 different minigene variants. Histogram summaries number of positions with a given mutation frequency. Orange line indicates mean mutation frequency across all positions.

(e) The majority of mutations occur in at least five different plasmid variants (labelled in orange). Cumulative distribution of mutations with a given mutation occurrence.

(f) Bioinformatics workflow for RNA-seq analysis to quantify splice isoforms. Upon quality control and filtering similar to (b), reads were aligned to wt *RON* minigene using splice-aware alignment software STAR. All isoforms present in RNA-seq library were reconstructed and filtered for minimum abundance using custom scripts (R). See Methods for details.

(g) Each canonical isoform is uniquely 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.

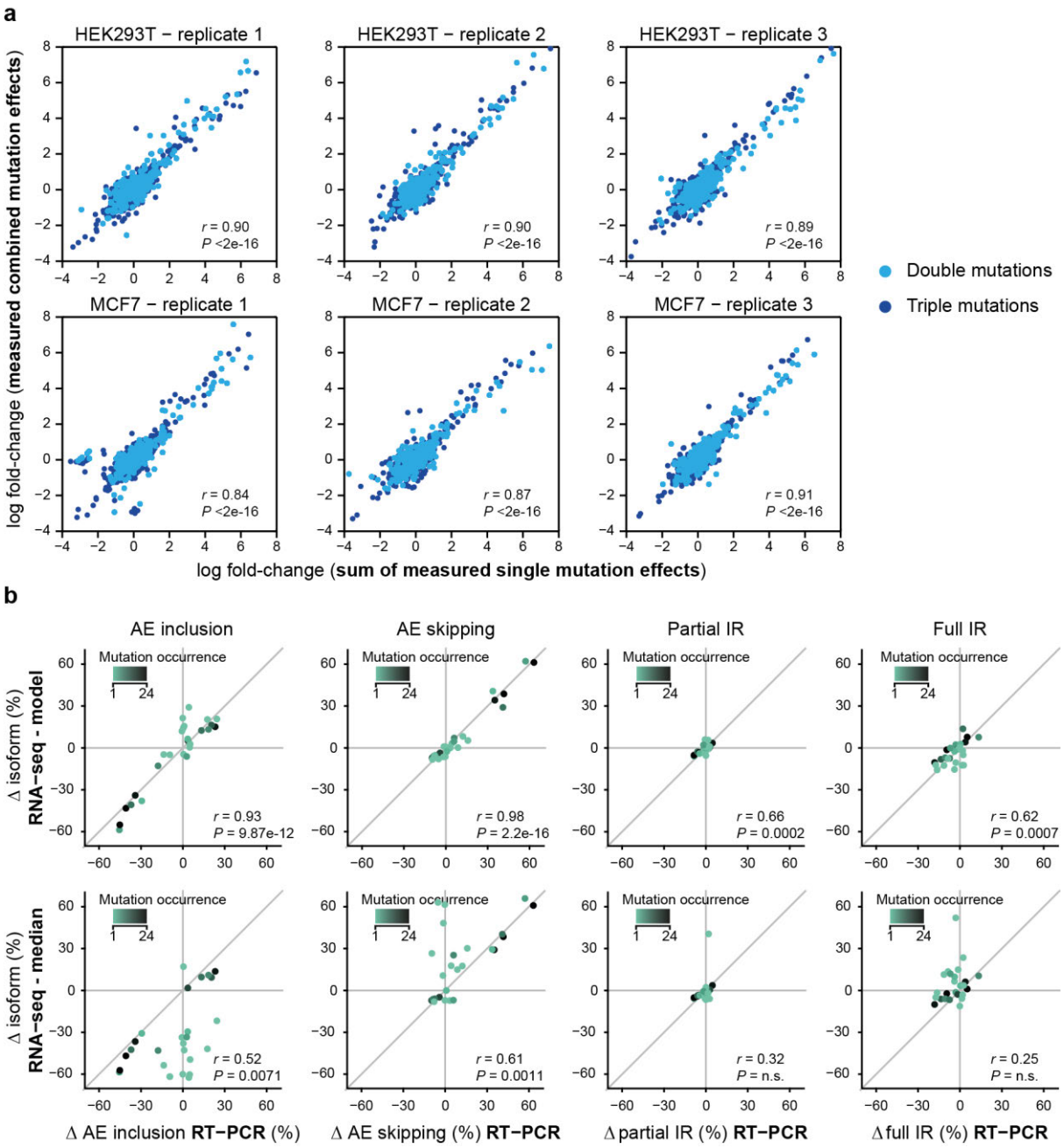
(h) The majority of minigene variants in the library is recovered in all three RNA-seq replicates from HEK293T cells. Pie chart displays the fraction of 5,791 minigene variants in the library that is recovered in 0-3 replicates.



Supplementary Figure 3: Modelling workflow for the inference of single mutation effects. Related to Fig. 2a.

(a) Kinetic model of splicing linearises splicing effects. Pre-mRNA synthesis, splicing reactions and mRNA degradation (scheme) are described by a set of ordinary differential equations (1). At steady state, each isoform frequency is described by a Michaelis-Menten-type equation (2), leading to non-linear mutation effects (e.g., effect of a mutation-induced change in r_1 depends on other parameters, i.e. other mutations). Mutation effects (e.g., on r_1) have linear effects when splice isoform ratios relative to a reference isoform are considered (3). See **Supplementary Note 1** for details.

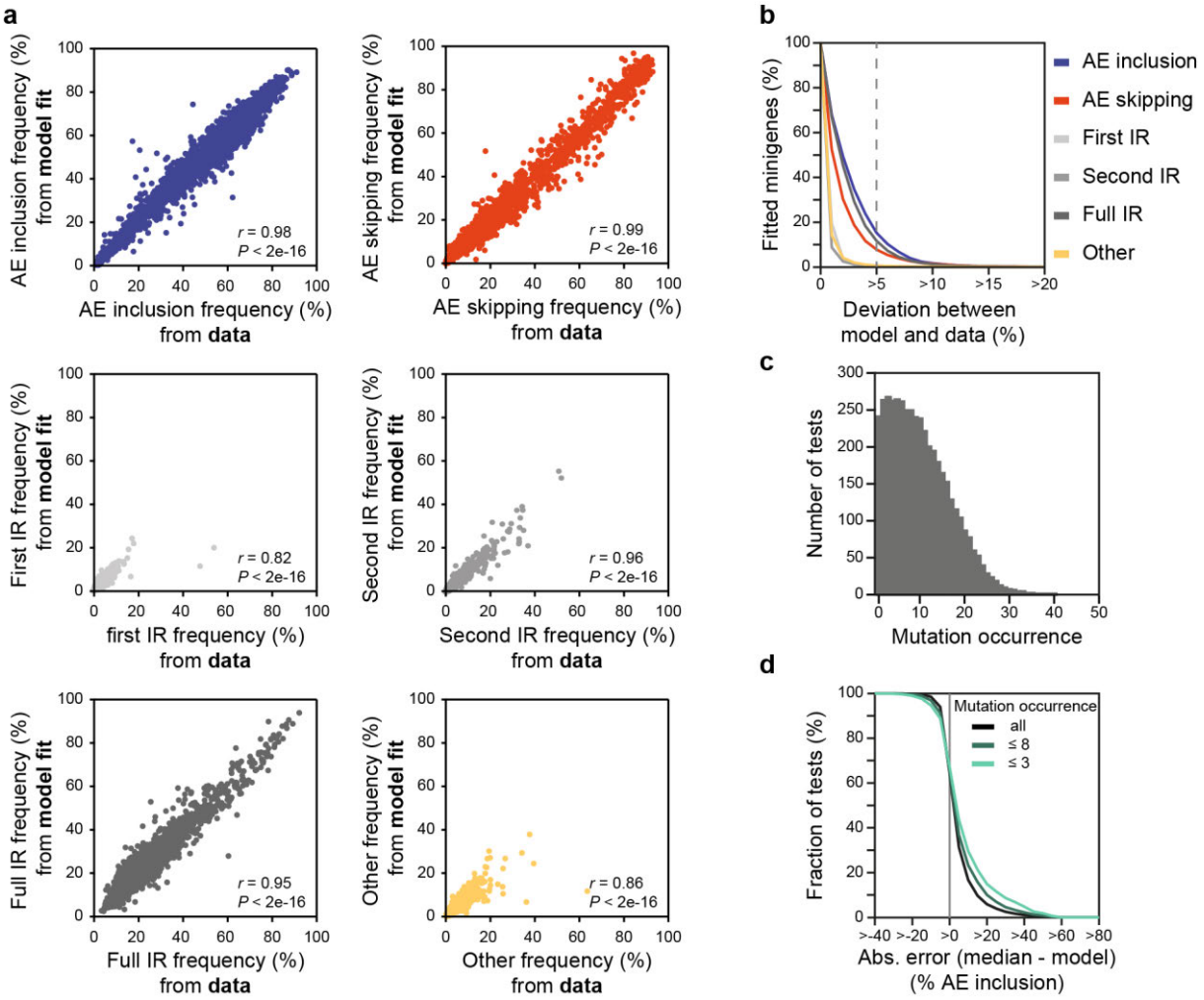
(b) Linear regression model infers single mutation effects. Effect of combined mutations (log fold-change) is formulated as sum of individual mutation effects. Five regression models (one per splice isoform), each containing ~5,000 equations (one per minigene), are formulated and fitted to the data. The models can be used to predict ~2,000 single mutation effects (700 nucleotides * 3 nucleotide exchanges) for each splice isoform. Model was subjected to cross-validation by leaving out 10% of the minigenes (i) or individual single mutation minigenes (ii) from the fit. Independent validation was performed by testing model predictions against RT-PCR for novel single mutation minigenes. See **Supplementary Note 2** for details.



Supplementary Figure 4: Single mutation effects are additive and confirmed by semiquantitative RT-PCR. Related to Fig. 2d.

(a) Single mutation effects are additive. Scatterplots show that sum of directly measured single mutation effects (from single-mutation minigenes; according to linear regression assumption, **Fig. 2a**; x-axes) agrees well with corresponding experimental measurements (y-axes) of minigenes containing two or three of these mutations (double and triple mutations, respectively). Analyses are shown for three replicates in HEK293T (top) and MCF7 cells (bottom).

(b) Regression model outperforms a median-based estimation of single mutation effects. Effects of mutations that rarely occur in the library (colour-coded) correlate better with model-inferred than median-based estimates. Scatterplots compare model-inferred (top row) and median-based (bottom row) estimations of single mutation effects relative to wt (y-axes) to semiquantitative RT-PCR measurements (x-axes) of targeted minigenes harbouring single point mutations, insertions and deletions (**Supplementary Data 8**). Separate plots are shown for different splice isoforms. First IR and second IR were summed up as 'partial IR', since these isoforms cannot be discriminated in RT-PCR. Pearson correlation coefficient and associated *P*-value are given in each panel. See Methods for description of median-based estimation.



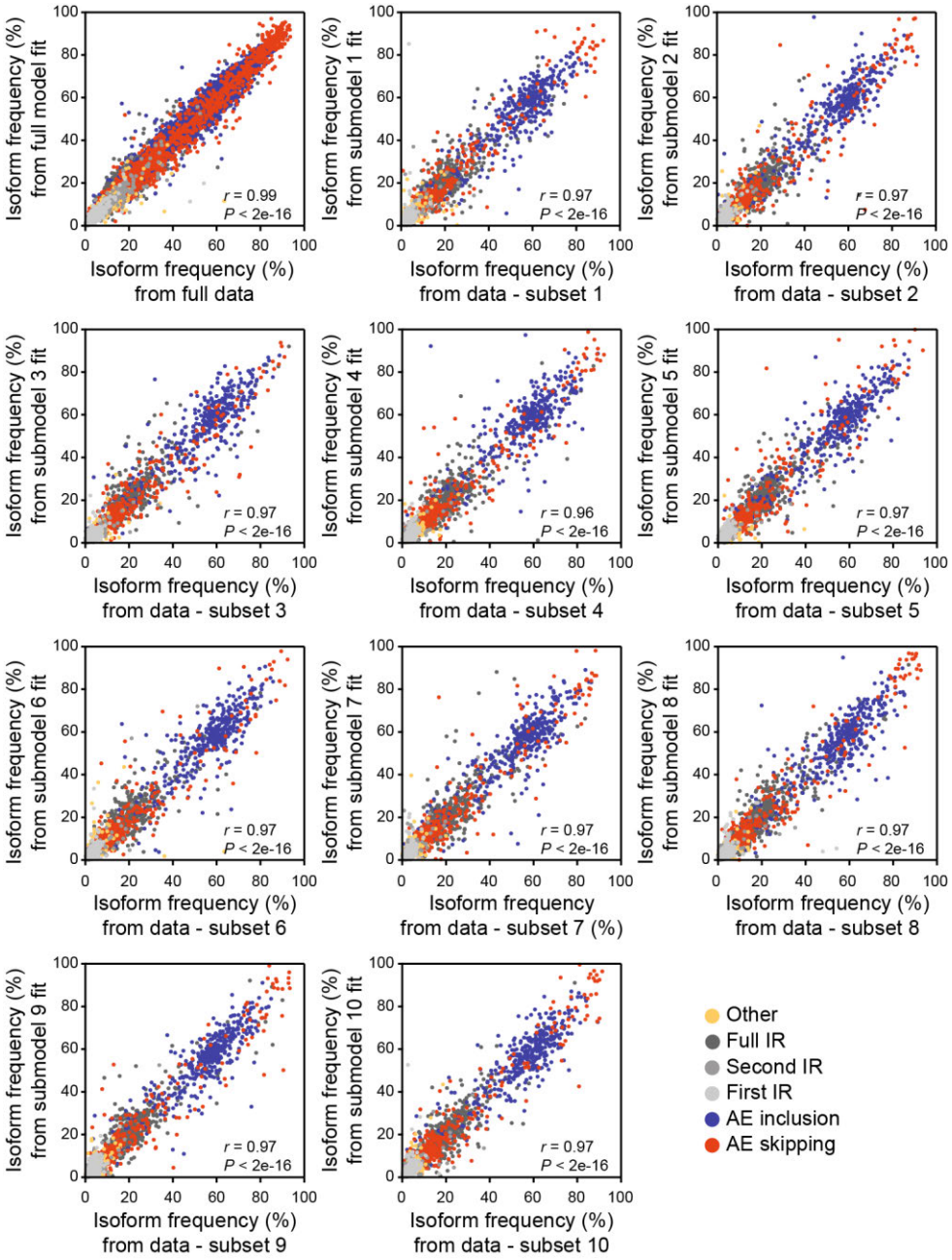
Supplementary Figure 5: The regression model increases the precision of isoform frequency estimations. Related to Fig. 2.

(a) Regression model describes experimentally measured isoform frequencies for each mutated minigene variant with high correlation (Pearson correlation coefficients r = 0.82-0.99, P-values < 2e-16). Scatterplots show frequencies of each of five canonical and non-canonical ('other') isoforms for combined mutations calculated from fitted model against measured data of one biological replicate (see **Supplementary Note 2**). Related to Fig. 2b.

(b) Majority of minigene variants are fitted within 5% deviation from measured value. For each isoform, fraction of fitted minigenes (y-axis) is shown for which model-derived isoform frequencies and measured data deviate more than a given %-value (x-axis). Related to Fig. 2b.

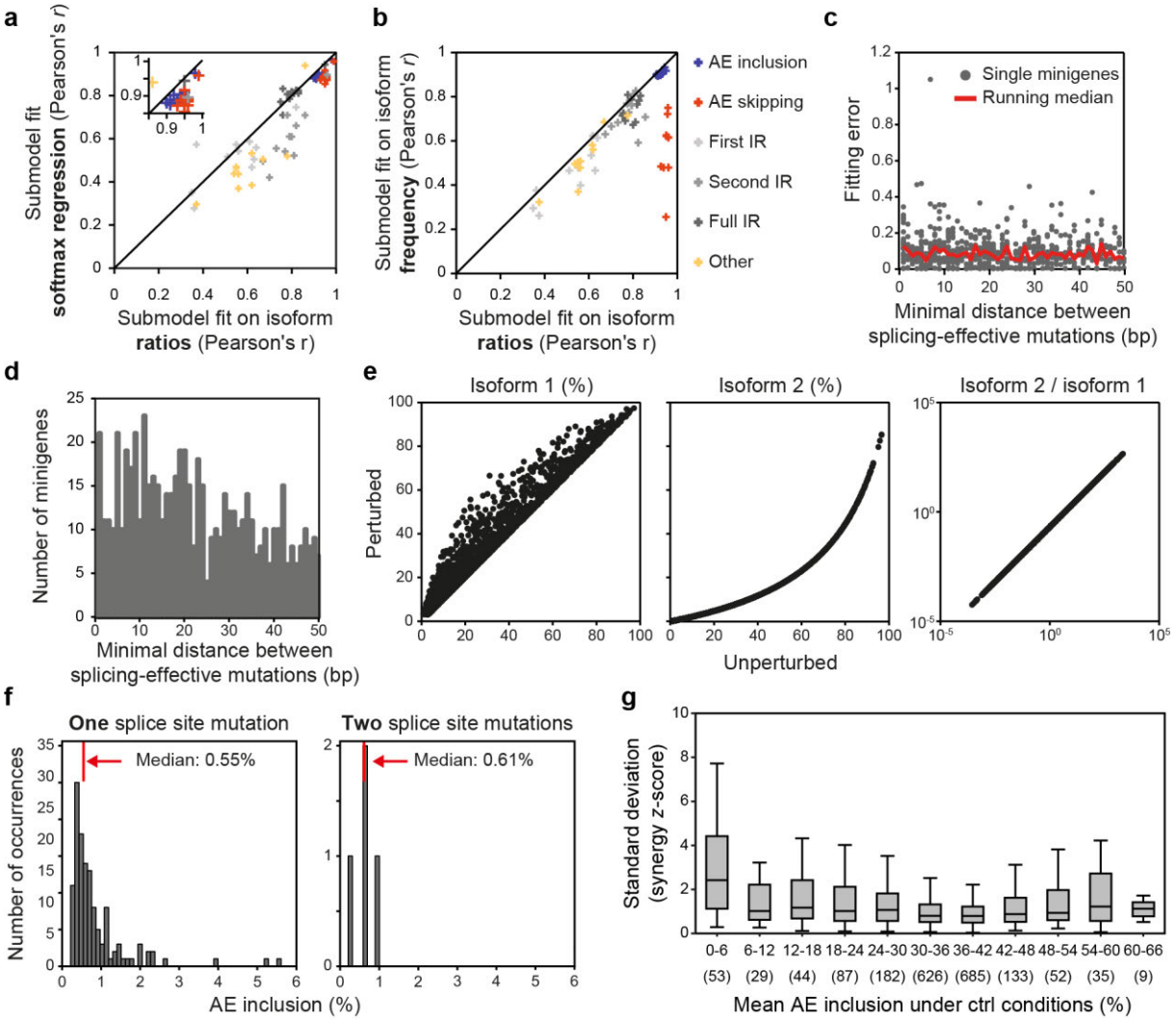
(c) Number of tests for different mutation occurrences that was used to calculate inference error of the model shown in Fig. 2c. Inference errors were estimated by separately benchmarking 561 mutation effects from single-mutation minigenes. To this end, minigenes containing the respective mutation were successively removed from the dataset, and subsequently model-inferred mutation effects were compared to isoform frequencies of single-mutation minigenes excluded from the analysis. Mutation occurrence shows number of different multi-mutation minigenes containing reference mutation used in one test. By successively reducing the dataset, we obtain the prediction accuracy for a particular mutation for different mutation occurrences. In some cases, estimation of mutational effects was not possible from a reduced dataset. These tests were left out, which explains the non-monotonical dependence of the number of tests on mutation occurrence. Related to Fig. 2c.

(d) Gain in accuracy for model-inferred isoform frequency estimations compared to median-based estimates. Difference of absolute errors in AE inclusion (%) between model and median-based calculation (x-axis) for a cumulative fraction of tests (y-axis) used in Fig. 2c. In 65% of tests, the model outperforms median-based estimation. Improvement of the model is more pronounced when considering only tests with low mutation occurrences (see legend). Related to Fig. 2c.



Supplementary Figure 6: Cross-validation underlines predictive power of the model for minigenes that were not used in training. Related to Fig. 2b.

The minigene library was randomly split into ten equal-sized subsets. During 10-fold cross-validation, regression models (one for each splice isoform) were fitted to all data excluding one subset. Scatterplots compare model-predicted splicing outcome for left-out subset to corresponding experimental data for all splice isoforms (see legend). In the first panel, full model fit is plotted against full dataset, followed by model prediction-data comparisons for ten different subsets. Pearson correlation coefficient and associated *P*-value are given in each panel.



Supplementary Figure 7: Linear regression modelling based on splice isoform ratios accurately infers single mutation effects in HEK293T cells. Related to Fig. 5b.

(a,b) Correlation between model-inferred isoform frequencies and experimental data improves when using linear regression on isoform ratios compared to softmax regression **(a)** or constrained linear regression on isoform frequencies **(b)**. Comparison of Pearson correlation coefficients (*r*) from 10-fold cross-validation (see **Supplementary Note 2**). Isoforms are colour-coded as indicated in (b).

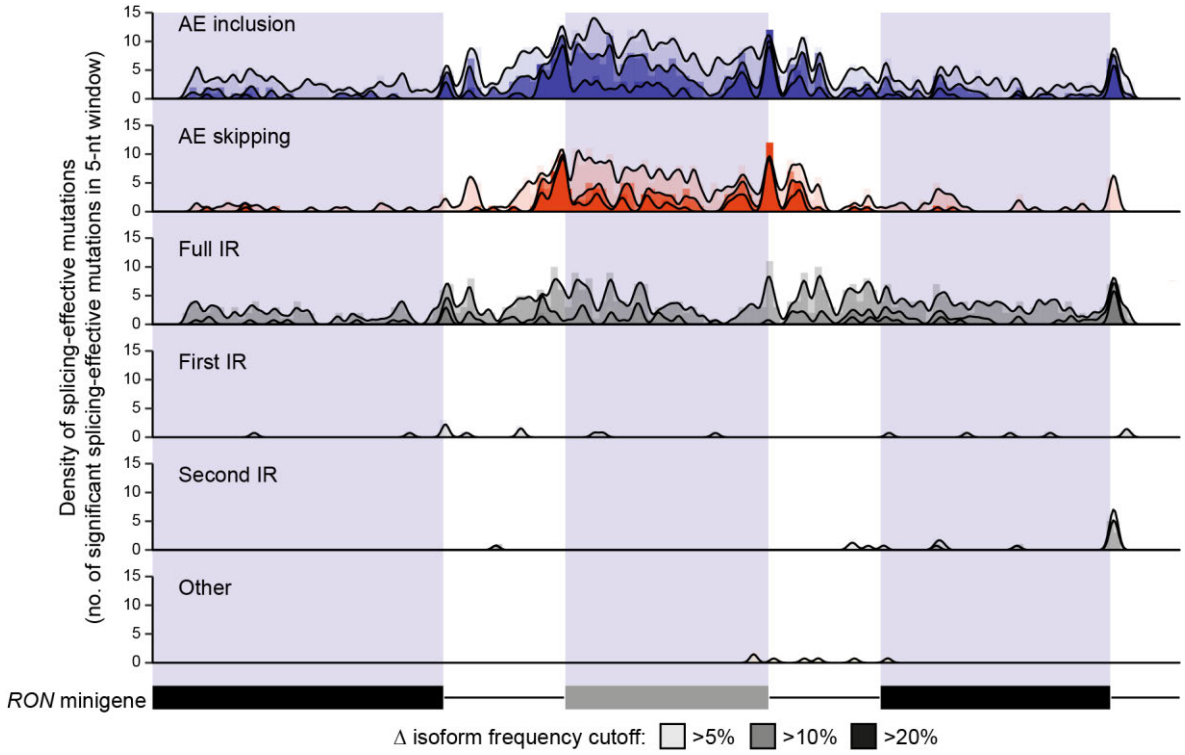
(c) Distance between mutations in the *RON* minigene does not influence the fitting error. Overall fitting error was computed by summing up the absolute deviation between fit and data for all isoforms. Only minigenes containing at least two mutations with significant effects on either isoform are plotted. The minimal distance between adjacent effective mutations contained in each minigene defines the x-axis.

(d) While 1,682 minigenes in our screen contained at least two splicing-effective mutations, only 84 of them occur within a distance of less than seven nucleotides. Histogram quantifies minigenes with a given minimal distance of splicing-effective mutations, corresponding to the number of data points for each value on the x-axis plotted in (c).

(e) Numeric simulation of competing splicing kinetic reactions reveals that perturbations of splicing rates have a linear effect on splice isoform ratios. Kinetic equations reflecting competing splicing reactions (Supplementary Equations 1 and 2 in **Supplementary Note 1**) were analysed *in silico*. The change of the steady-state after decreasing the production rate of one splicing isoform to 20% was simulated. The effect of this perturbation on all splicing isoforms is nonlinear and depends on the mutational context. In contrast, the perturbation has a linear effect on splice isoform ratios.

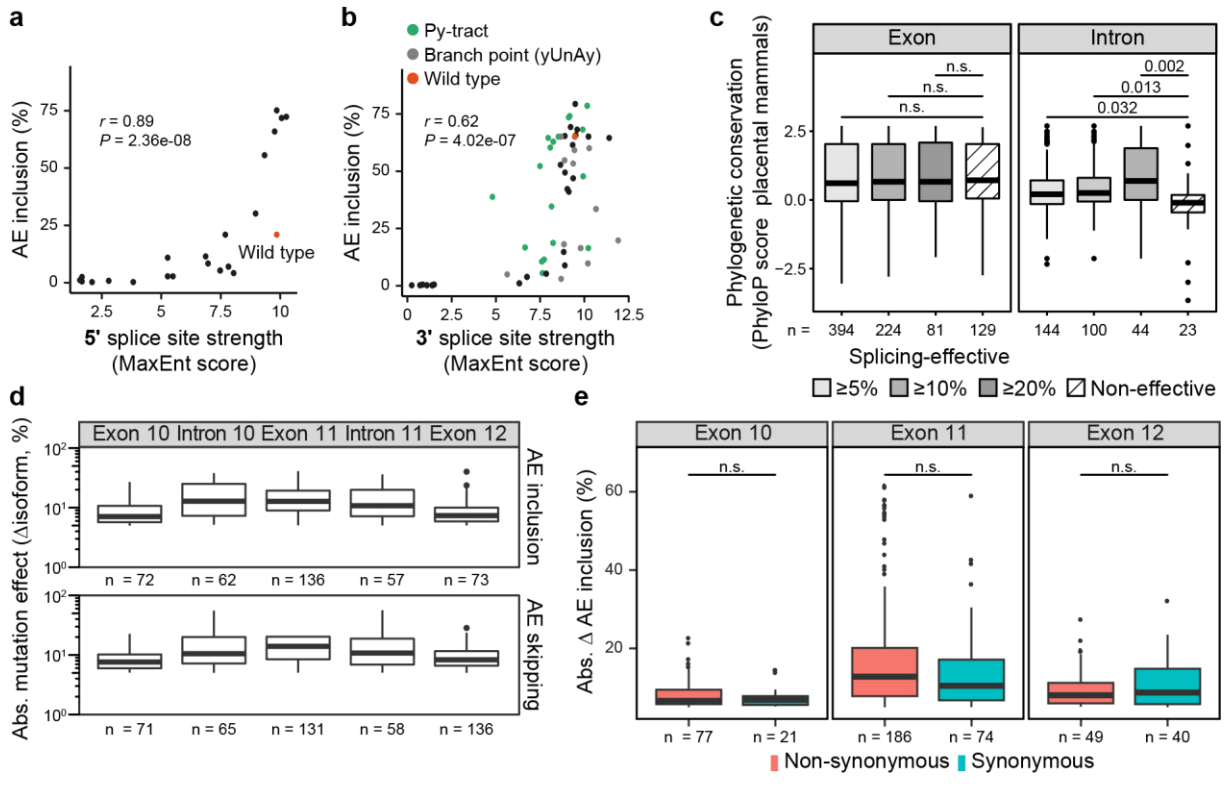
(f) The presence of two splice site mutations in a minigene does not further decrease AE inclusion compared to minigenes containing only one splice site mutation. Histograms of AE inclusion frequency in minigenes containing one or two splice site mutations.

(g) Computation of the synergy score is unstable for mutations abolishing AE inclusion. Boxplox shows the standard deviation of the synergy score for the AE skipping to AE inclusion ratio over the three replicates for mutations with mean control AE inclusion in different ranges. Bounds of each box represent quartiles, centre line denotes 50th percentile, and whiskers extend to most extreme data points. Mutations leading to control AE inclusion less than 6% show greatest uncertainty in the computation of the synergy *z*-score.



Supplementary Figure 8: Complete landscape of splicing-effective mutations in HEK293T cells. Related to Fig. 2e.

Bar diagrams for each isoform show the number of splicing-effective mutations in adjacent 5-nt windows across the *RON* minigene (FDR < 0.1%). Lines indicate the density of significant splicing-effective mutations in a 5-nt sliding window. Light to dark shading indicates cutoffs at >5%, >10%, and >20% change in isoform frequency, identifying a total of 778, 362 and 136 splicing-effective mutations, respectively. The alternative exon constitutes a regulatory hotspot for alternative exon (AE) inclusion and AE skipping. Mutations affecting full intron retention (IR) are dispersed across the alternative exon and the downstream constitutive exon.



f

	Effective mutations	Non-effective mutations	Percentage of effective mutations
Non-synonymous	312	568	35%
Synonymous	135	272	33%

Supplementary Figure 9: Splice site strength, evolutionary conservation and coding potential of splicing-effective positions. Related to Fig. 2.

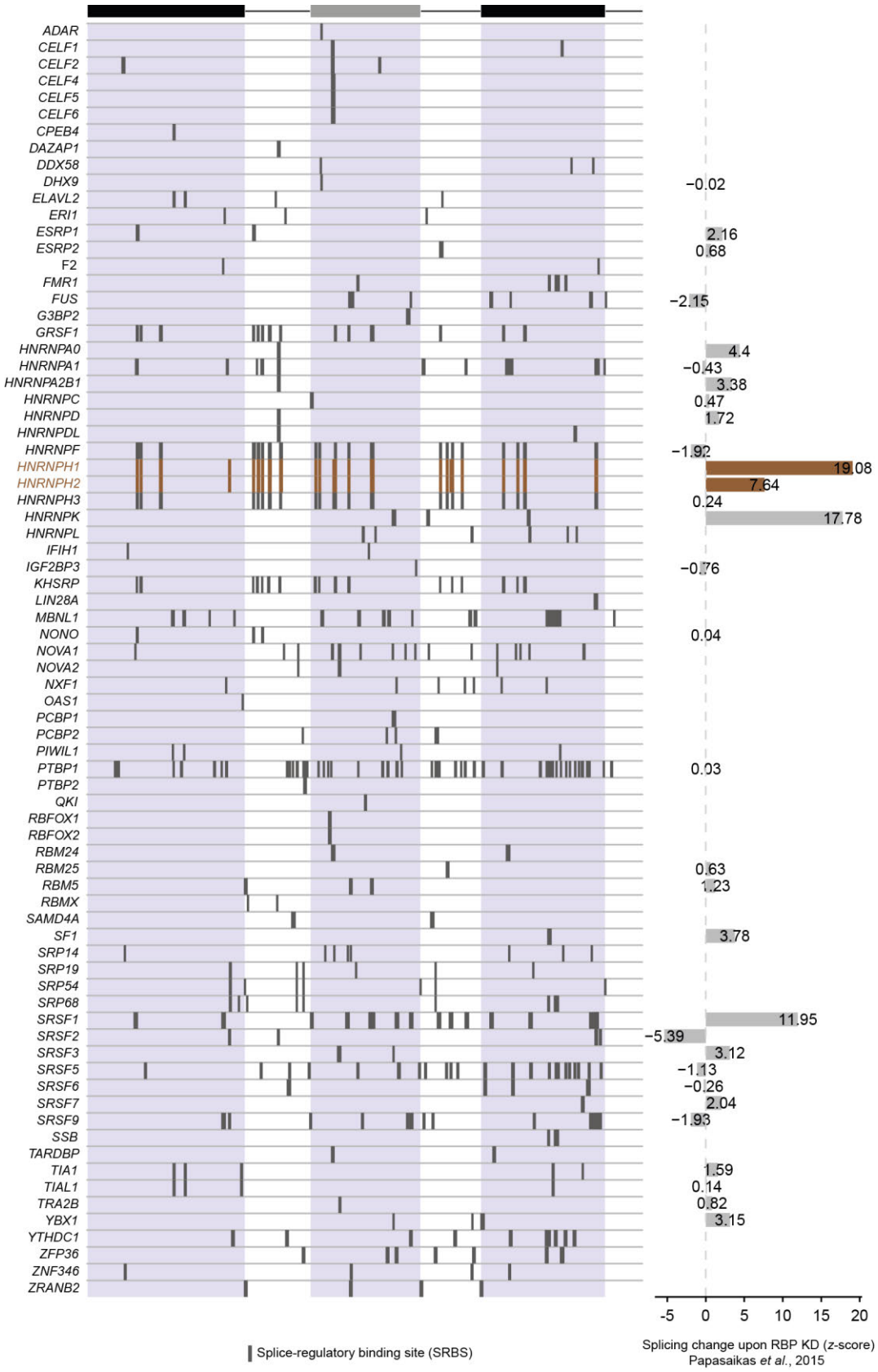
(a,b) Mutation effects at the 3' splice site **(a)** and 5' splice site **(b)** of *RON* exon 11 correlate with splice site strengths predicted by the MaxEntScan software. Scatterplots compare AE inclusion frequencies from HEK293T cells (y-axes) to the predicted splice site strength (MaxEnt score) for all mutations in positions considered by MaxEntScan (278-300 nt and 442-450 nt for 3' and 5' splice site, respectively). Red, green and grey dots indicate wt minigene and variants with mutations in polypyrimidine tract (Py-tract; 286-293 nt) and branch point motif (yUnAy, where y is pyrimidine and n is any base; 279-283 nt), respectively. *r*, Spearman correlation coefficient and corresponding *P*-value.

(c) Splicing-effective positions are significantly more conserved evolutionarily than permissive mutations within introns, but not exons. Boxplot shows distribution of conservation scores (PhyloP score across 46 placental mammals) for splicing-effective (light to dark shading indicating cutoffs at >5%, >10%, and >20% change in isoform frequency) and permissive positions in MCF7 cells in exons (left) and introns (right) of the *RON* minigene. Number of positions in each box indicated below. Centre line and bounds of each box denote 25th, 50th and 75th percentile, and whiskers extend to most extreme values within 1.5x interquartile range (IQR). *P*-values correspond to two-sided Mann-Whitney-Wilcoxon test. n.s., not significant.

(d) Splicing-effective mutations in *RON* exon 11 and the flanking introns are comparably strong. Boxplots summarise absolute changes in AE inclusion (top) or AE skipping (bottom) for significant splicing-effective mutations (>5%) in the different transcript regions (number given below). Box representation as in (c).

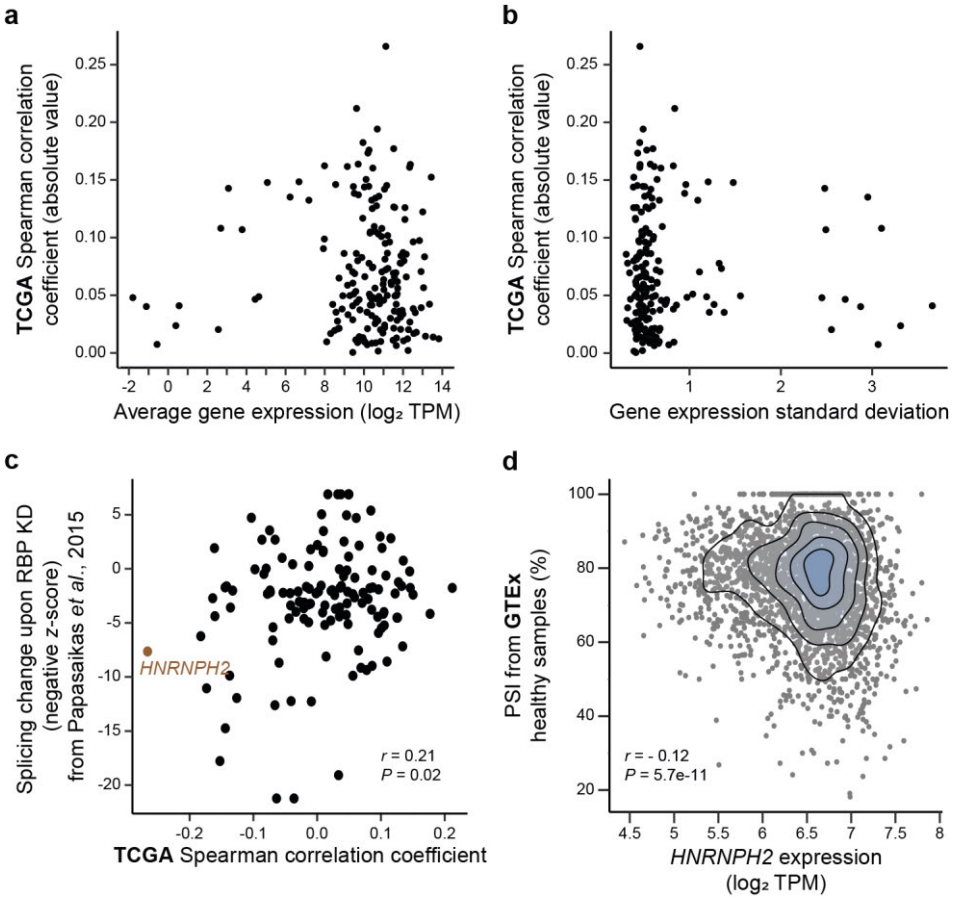
(e) Synonymous and non-synonymous mutations show similar effect sizes. Boxplots show absolute changes in AE inclusion in HEK293T cells for synonymous and non-synonymous mutations in exons 10-12. Number of positions in each box indicated below. Box representation as in (c). Significance was tested using two-sided Mann-Whitney-Wilcoxon test. n.s., not significant.

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



Supplementary Figure 10: Putative RBP regulators of *RON* exon 11 splicing and their predicted splice-regulatory binding sites. Related to Fig. 3e.

in silico binding site predictions for RNA-binding proteins (RBPs) identify splice-regulatory binding sites (SRBS; predicted binding sites that show substantial mutation effects, see Methods). Boxes indicate the location of SRBS for the 76 putative RBP regulators that were identified by ATtRACT. Predicted binding sites for HNRNPH1 and HNRNPH2 are highlighted in brown. Bar diagram (right) shows splicing effects (*z*-scores, values indicated at each bar) for 31 RBPs that are present in published data[2] on *RON* exon 11 splicing upon RBP knockdown (KD). Positive and negative *z*-scores correspond to increased and decreased *RON* exon 11 inclusion upon RBP KD, respectively.

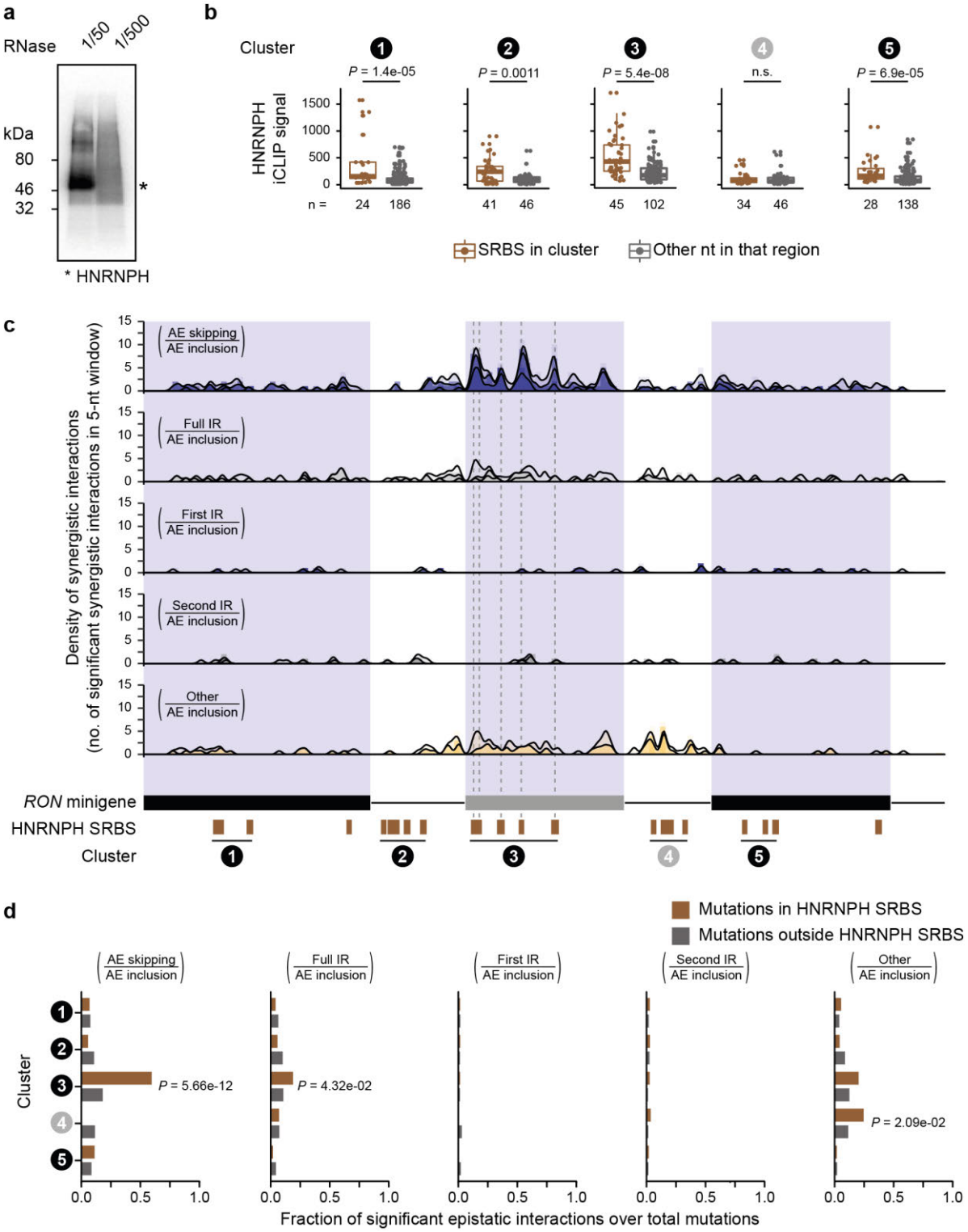


Supplementary Figure 11: Expression correlation of *HNRNPH2* and other RNA-binding proteins (RBPs) with *RON* exon 11 inclusion in TCGA and GTEx samples. Related to Fig. 3c,d,f.

(a,b) Absolute Spearman correlation coefficients of RBP expression (in transcripts per million, TPM) and *RON* exon 11 inclusion (in percent spliced-in, PSI) across TCGA samples do not depend on the average expression levels across samples **(a)** nor on the associated standard deviations **(b)**.

(c) Correlation between RBP expression and *RON* exon 11 inclusion in TCGA tumour samples partially recapitulates the observed effect of those RBPs in a previous knockdown (KD) screen[2]. Scatterplot compares Spearman correlation coefficients from TCGA samples with published *z*-scores (inverted sign) upon RBP KD. *HNRNPH2* is highlighted. *r*, Pearson correlation coefficient and corresponding *P*-value.

(d) *RON* exon 11 inclusion inversely correlates with *HNRNPH2* expression across 2,743 samples derived from 24 different healthy human tissues. Density scatterplot shows *HNRNPH2* expression (in TPM) and *RON* exon 11 inclusion (in PSI) across healthy samples from the Genotype-Tissue Expression (GTEx) project. *r*, Spearman correlation coefficient and corresponding *P*-value.



Supplementary Figure 12: HNRNPH iCLIP and synergistic interactions reveal functional HNRNPH binding sites. Related to Figs 4a and 5c.

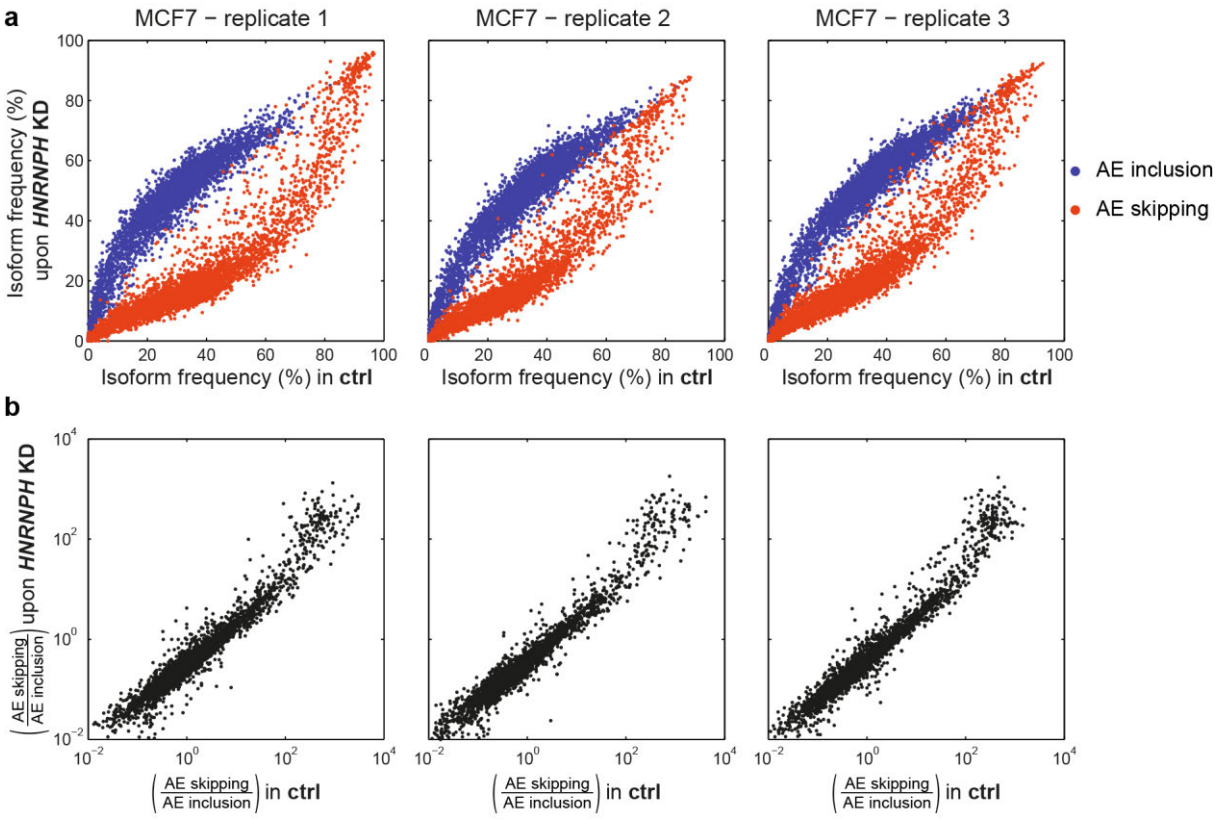
(a) Autoradiograph shows crosslinked HNRNPH/RNA complexes that were treated with increasing RNase I dilutions prior to immunoprecipitation for optimisation of partial RNase digestion. Protein-RNA complexes run above expected molecular weight of HNRNPH (53 kDa; labelled by asterisk).

(b) HNRNPH crosslink events are significantly enriched in four out of five clusters of HNRNPH splice-regulatory binding sites (SRBS). Boxplots summarise HNRNPH iCLIP crosslink events on all nucleotides (nt) within SRBS ± 2 nt (brown) of each cluster (labelled by numbered circles) compared to all other positions within same exon/intron (grey). Number of positions in each box indicated below. Centre line and bounds of each box denote 25th, 50th and 75th percentile, and whiskers extend to most extreme values within 1.5x interquartile range (IQR). *P*-values correspond to two-sided Wilcoxon Rank-Sum test.

(c) Synergistic interactions between point mutations and *HNRNPH* KD are predominantly observed for AE inclusion, AE skipping, and 'other' isoforms. Bar diagrams for each splice isoform ratio show number of significant synergistic interactions (FDR < 0.1%) in adjacent 5-nt windows. Lines indicate the density in a 5-nt sliding window. Each panel

shows an overlay of increasing *z*-score cutoffs (|*z*|>2 >3, and >5), identifying a total of 354, 222 and 66 significant synergistic interactions, respectively (**Supplementary Table 2**). Splice sites ± 2 nt were excluded from this analysis. *RON* minigene structure and predicted HNRNPH SRBS clusters are given below.

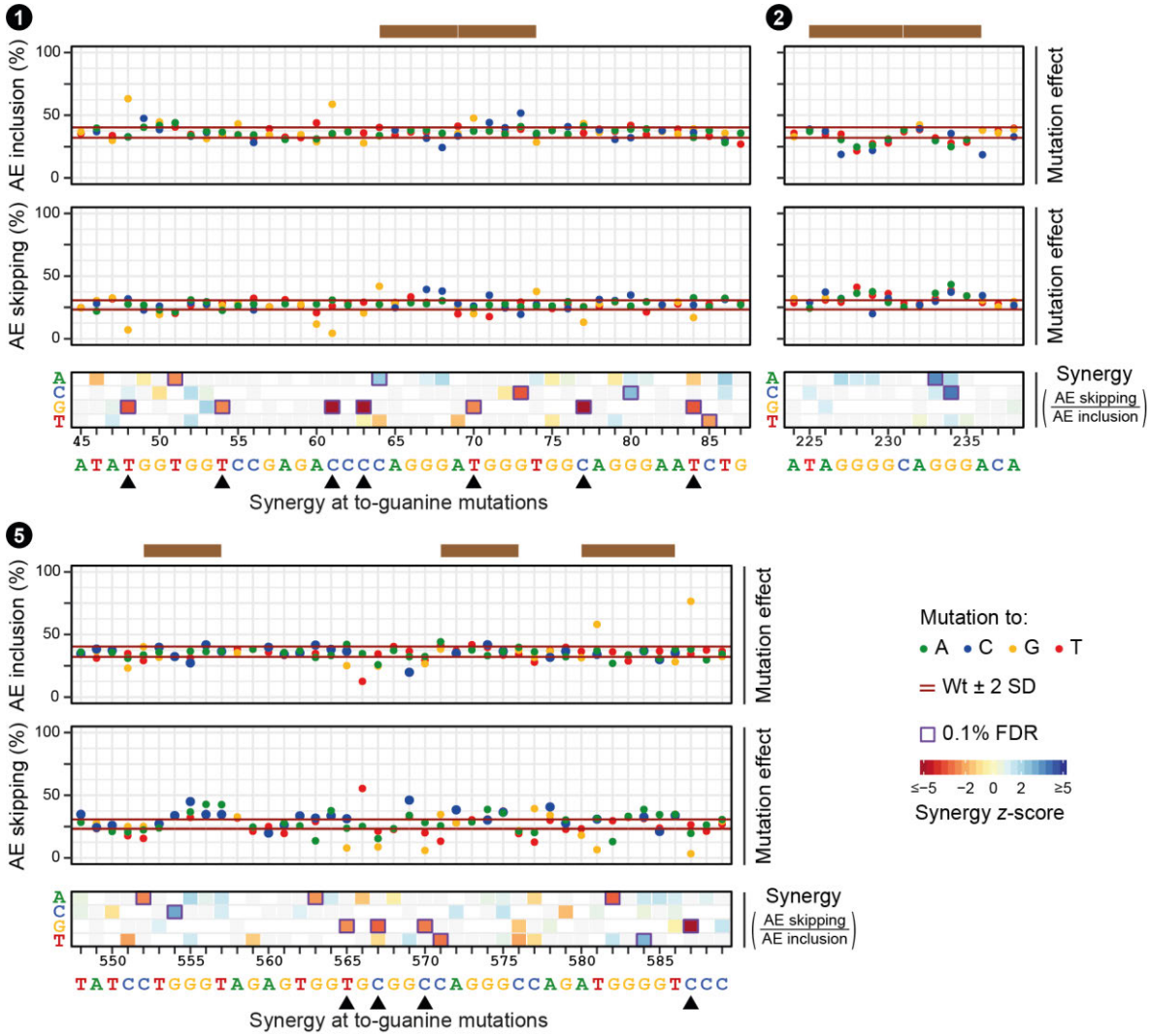
(d) Synergistic interactions are significantly enriched within HNRNPH SRBS in cluster 3. Bar diagrams for each splice isoform ratio display the fraction of significant synergistic interactions over all mutations for SRBS within the five clusters (brown) compared to all other positions within same exon/intron (grey). Significant differences are shown with *P*-values correspond to one-sided Wilcoxon Rank-Sum test.



Supplementary Figure 13: *HNRNPH* KD shows non-linear effects on splice isoforms, while splice isoform ratios respond linearly.

(a) AE inclusion (blue) and AE skipping (red) isoform frequencies in MCF7 cells under control (ctrl) and *HNRNPH* KD conditions are shown for all individual minigene variants in three biological replicates. Depending on baseline frequency under control conditions, strength of KD-induced effect varies (top).

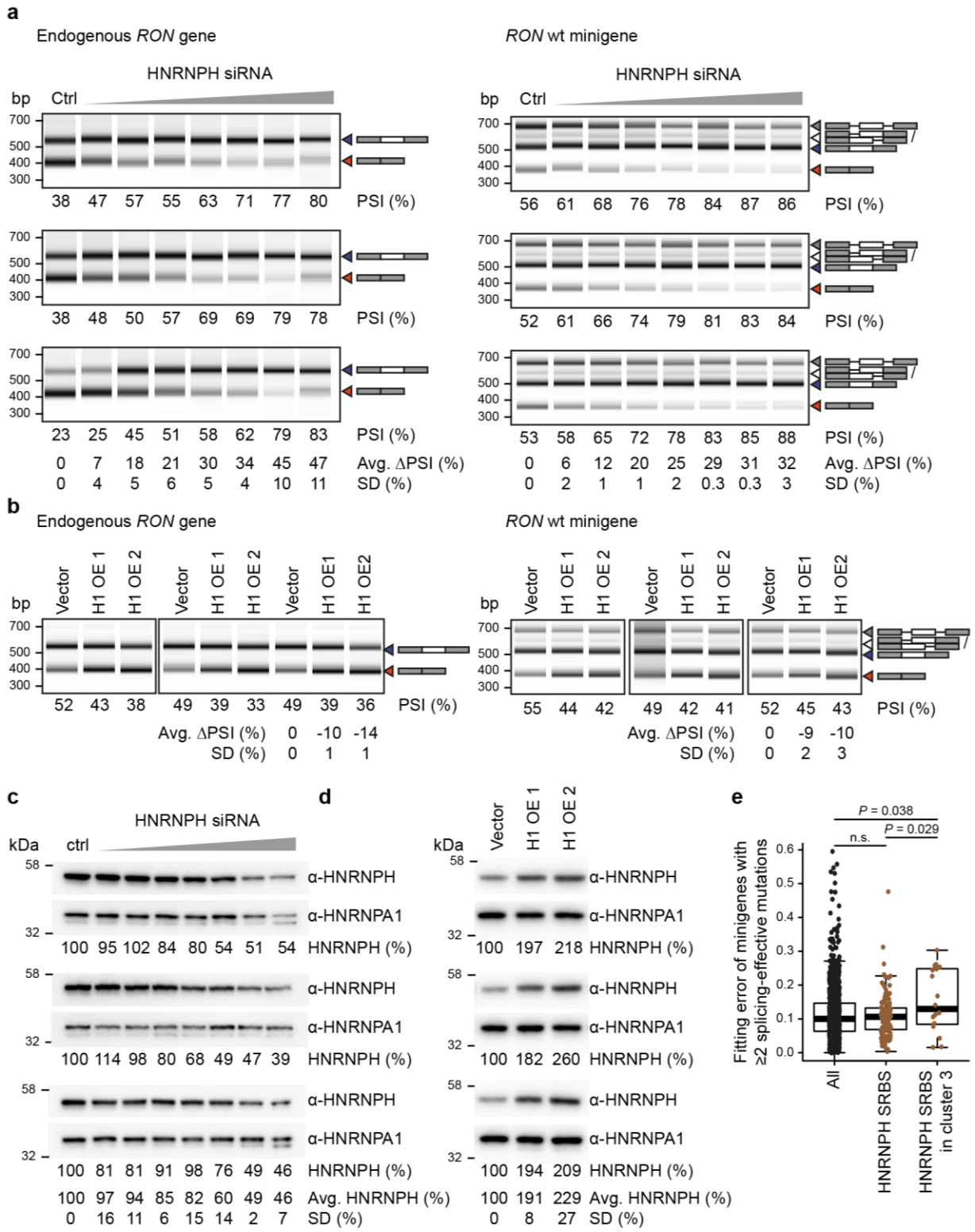
(b) Corresponding splice isoform ratios (AE skipping over AE inclusion) for individual minigene variants (black) are independent of baseline frequency and behave linearly.



Supplementary Figure 14: Mutation effects and synergistic interactions between *HNRNPH* KD and single point mutations highlight mutations that reinforce HNRNPH binding. Related to Fig. 5d.

Within HNRNPH splice-regulatory binding sites (SRBS) of clusters 1 and 5 (indicated by numbered circles) in constitutive exons 10 and 12, respectively, mutations to guanines generally lead to increased AE inclusion, while AE skipping levels are reduced. Strong synergistic interactions of these mutations (highlighted by arrowheads) suggest that strengthening HNRNPH binding at these sites enhances its splicing-regulatory function. HNRNPH SRBS cluster 2 in first intron regulates AE skipping and AE inclusion in opposite direction compared to HNRNPH SRBS cluster 3 (**Fig. 5d**).

For each SRBS cluster, three plots are shown summarising single mutations effects on AE inclusion (top) and AE skipping (middle) as well as synergistic interactions of mutations with *HNRNPH* KD (based on splice isoform ratio of AE skipping over AE inclusion; bottom). Single mutation effects are displayed as dot plot, with y-axis showing the isoform frequency (mean of three biological replicates) resulting from each individual mutation in a given position along the y-axis. Each dot represents one mutation, with colours indicating inserted nucleotide (green, mutation to A; blue, to C; yellow, to G; red, to T). Red lines indicate median isoform frequency of wt minigenes ± 2 standard deviations (SD). HNRNPH SRBS (brown) are given above. Synergistic interactions are displayed as a heatmap of *z*-scores (mean of three biological replicates) as a quantitative measure of synergy between indicated mutation and *HNRNPH* KD. Each row represents one type of inserted nucleotide (indicated on the left). White and grey fields indicated mutations that were either not present or filtered out due to inconsistent signs (see Methods). Purple boxes highlight significant synergistic interactions (0.1% FDR).



Supplementary Figure 15: *RON* exon 11 splicing is sensitive to reduced HNRNPH levels. Related to Fig. 6b.

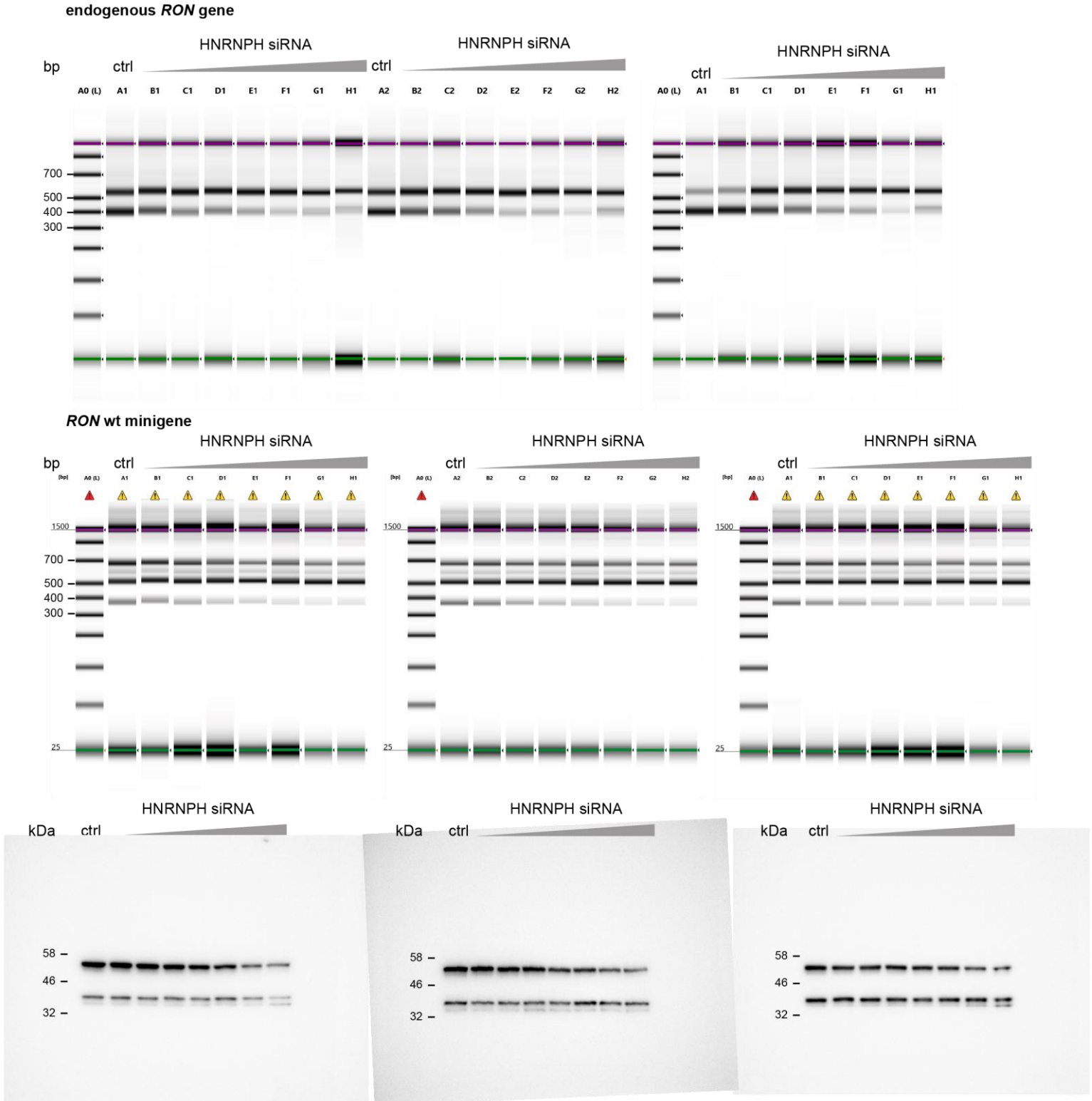
(a) *RON* exon 11 inclusion for endogenous *RON* gene and wt *RON* minigene upon gradual reduction of HNRNPH using increasing concentrations of HNRNPH-specific siRNA. Semiquantitative RT-PCR results in MCF7 cells are visualised in a gel-like representation of capillary electrophoresis. Splice products are indicated on the right. Percent spliced-in (PSI) for each condition is given below. Average (Avg.) and standard deviation (SD) of splicing change (ΔPSI against non-targeting control siRNA, Ctrl) across the three replicates are given below. Whole gel images for these experiments are shown in Supplementary Fig. 16.

(b) *RON* exon 11 inclusion for the endogenous *RON* gene and the wt *RON* minigene upon gradual overexpression of *HNRNPH1* (H1 OE1/OE2) compared to a transfection with an empty vector control (Vector). Semiquantitative RT-PCR results of three biological replicates in MCF7 cells. Visualisation as in (a). Whole gel images for these experiments are shown in Supplementary Fig. 16.

(c) Western Blot analysis to quantify amount of HNRNPH upon gradual *HNRNPH* knockdown using increasing concentrations of HNRNPH-specific siRNA in three biological replicates. HNRNPA1 served as loading control. Relative HNRNPH abundance normalised against HNRNPA1 (in %) is given below. Average (Avg.) and standard deviation (SD) of HNRNPH abundance relative to non-targeting control siRNA (Ctrl) across the three replicates are given below. Whole gel images for these experiments are shown in Supplementary Fig. 16.

(d) Western Blot analysis to quantify amount of HNRNPH upon gradual *HNRNPH1* overexpression (H1 OE1/OE2) compared to empty vector transfection (Vector) in three biological replicates. Loading control and visualisation as in (c). Whole gel images for these experiments are shown in Supplementary Fig. 16.

(e) Minigenes with combination of splicing-effective mutations in HNRNPH SRBS cluster 3 show increased fitting errors, evidencing cooperative HNRNPH binding. Fitting error of minigenes with multiple splicing-effective mutations in HNRNPH SRBS cluster 3 is larger than for other minigenes containing splicing-effective mutations within other HNRNPH SRBS or elsewhere in the *RON* minigene. *P*-values correspond to one-sided Student's t-test. Whole gel images for these experiments are shown in Supplementary Fig. 16.



Supplementary Figure 16: Whole gel images for experiments shown in Supplementary Figure 15.

Semiquantitative RT-PCR analyses of *RON* exon 11 splicing upon gradual reduction of HNRNPH visualised in a gel-like representation of capillary electrophoresis (top) and corresponding Western blot analyses (bottom).

Supplementary References

1. Raue, A. *et al*. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923-1929 (2009).
2. Papasaikas, P., Tejedor, J. R., Vigevani, L. & Valcárcel, J. Functional splicing network reveals extensive regulatory potential of the core spliceosomal machinery. *Mol. Cell* **57**, 7-22 (2015).
3. Bonomi, S. *et al*. HnRNP A1 controls a splicing regulatory circuit promoting mesenchymal-to-epithelial transition. *Nucleic Acids Res.* **41**, 8665-8679 (2013).

2.2 Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts

2.2.1 Abstract

Recently, aberrant *CD19* pre-mRNA splicing has been associated with relapse in leukemia patients following CAR-T cell therapy. One such splicing isoform comprises a cryptic intron within *CD19* exon 2 (*CD19*ex2part) that does not contain classical splice sites, but instead contains an eight nucleotide long repetitive sequence at both splice sites.

Using a splicing reporter assay and direct long-read RNA sequencing, we demonstrated that this putative cryptic intron is actually an artifact caused by reverse transcription. Analyzing other available data sets, we discovered dozens of other examples of such "exitrons" of suspect nature. They appear in long-read cDNA sequencing but not in direct RNA sequencing. These results stress the need for double validation of unexpected junctions in RNA isoforms by direct RNA sequencing.

2.2.2 Zusammenfassung

In letzter Zeit wurde fehlerhaftes Spleißen von *CD19*-pre-mRNA immer wieder mit dem Rückfall von Leukämiepatienten nach einer CAR-T-Zelltherapie in Verbindung gebracht. Eine solche Spleißisoform ist *CD19*ex2part. Sie beinhaltet ein kryptisches Intron innerhalb von Exon 2. Dieses kryptische Intron hat keine klassischen Spleißstellen; stattdessen liegt an beiden Spleißstellen eine acht Nukleotid lange repetetive Sequenz vor.

Mithilfe eines Spleiß-Reporter-Assays und direkter “Long-Read“-RNA-Sequenzierung konnten wir nachweisen, dass dieses vermeintliche kryptische Intron in Wirklichkeit ein Artefakt ist, das durch reverse Transkription verursacht wird. Bei der Analyse anderer verfügbarer Datensätze entdeckten wir außerdem Dutzende weiterer Beispiele für solch verdächtige "Exitrons". Sie tauchen bei der “Long-Read“-cDNA-Sequenzierung auf, nicht aber bei der direkten RNA-Sequenzierung. Diese Ergebnisse unterstreichen die Notwendigkeit einer doppelten Validierung von unerwarteten “Splice Junctions“ in RNA-Isoformen durch direkte RNA-Sequenzierung.

2.2.3 Statement of contribution

I performed the majority of the experimental work of this publication. To prove the non-existence of *CD19*ex2part as an RNA molecule, I generated a splicing reporter plasmid (as well as several control plasmids). The plasmid contained exon 2 of *CD19* and the coding sequence of the fluorescent proteins eGFP (upstream of the exon) and mCherry (downstream of the exon). I transfected this reporter into NALM-6 cells. I analyzed the fluorescence signal of the cells using flow cytometry. I also performed all RT-PCRs of the reporter RNA samples and additionally performed the thapsigargin assay to prove that unconventional splicing by IRE1 endoribonuclease also plays no role in the generation of *CD19*ex2part. In addition, I contributed in particular to the design of the study, the writing of the manuscript, and the preparation of the figures in this publication.

Schulz et al. Genome Biology (2021) 22:190
https://doi.org/10.1186/s13059 021 02411 1

Genome Biology

SHORT REPORT Open Access

Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts

Laura Schulz[1†], Manuel Torres Diz[2†], Mariela Cortés López[1†], Katharina E. Hayer[3†], Mukta Asnani[2], Sarah K. Tasian[4], Yoseph Barash[5], Elena Sotillo[2,6], Kathi Zarnack[7], Julian König[1*] and Andrei Thomas Tikhonenko[2,4,8*]

* Correspondence: j.koenig@imb mainz.de; andreit@pennmedicine. upenn.edu
†Laura Schulz, Manuel Torres Diz, Mariela Cortés López and Katharina E. Hayer contributed equally to this work.
[1]Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz, Germany
[2]Division of Cancer Pathobiology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
Full list of author information is available at the end of the article

Abstract

Resistance to CD19 directed immunotherapies in lymphoblastic leukemia has been attributed, among other factors, to several aberrant *CD19* pre mRNA splicing events, including recently reported excision of a cryptic intron embedded within *CD19* exon 2. While "exitrons" are known to exist in hundreds of human transcripts, we discovered, using reporter assays and direct long read RNA sequencing (dRNA seq), that the *CD19* exitron is an artifact of reverse transcription. Extending our analysis to publicly available datasets, we identified dozens of questionable exitrons, dubbed "falsitrons," that appear only in cDNA seq, but never in dRNA seq. Our results highlight the importance of dRNA seq for transcript isoform validation.

Keywords: Long read sequencing, Oxford Nanopore Technologies, Alternative splicing, mRNA isoforms, Exitrons, Reverse transcription, CD19, Immunotherapy, Blinatumomab

Background

Aberrant splicing plays an important role in therapeutic resistance either by generating protein isoforms resistant to treatment or by eliminating target proteins entirely. A prime example of this phenomenon is B cell acute lymphoblastic leukemia (B-ALL) acquiring resistance to chimeric antigen receptor-armed autologous T cells (CART-19), which are engineered to target the CD19 surface antigen of B cells [1]. We previously demonstrated that skipping of exon 2 of *CD19* pre-mRNA generates a protein variant inherently resistant to killing by CART-19 and mis-localized in the endoplasmic reticulum [2, 3]. Subsequently, we and others have shown that retention of the *CD19* intron 2 containing a premature termination codon contributes to CART-19 resistance as well [4, 5]. Of note, several publications reported that apparent removal of a cryptic intron fully embedded within *CD19* exon 2 generates a novel isoform in healthy individuals and B-ALL patients (termed Δex2part) [2, 6–8]. One study further suggested that this event could mediate resistance to blinatumomab, a CD19-CD3-bispecific T

cell engager ([6]; commentary by [9]). The same publication hypothesized that excision of the embedded intron might be catalyzed by the IRE1 (ERN1) endoribonuclease, which is responsible for unconventional splicing of the *XBP1* transcript during the unfolded protein response [10].

Such "exitrons" are known to exist in hundreds of human transcripts and are thought to evolve from ancestral coding exons, often preserving the open reading frames [11]. Given the potential significance of the reported *CD19* exitron, we began to investigate its nature using long-read Oxford Nanopore Technologies (ONT) sequencing. Long-read applications allow sequencing of complete transcript isoforms and have re-shaped our understanding of the complexities of human transcriptomes [12–14]. Different ONT protocols are currently available. In cDNA-seq, reverse transcribed (and often PCR-amplified) cDNA molecules are sequenced, while in dRNA-seq, polyadenylated mRNA molecules themselves are passed through the pores and read [15]. Both protocols can capture full transcripts, including alternatively spliced isoforms. However, dRNA-seq typically yields fewer reads and thus is most commonly used for detecting RNA modifications, such as adenine methylation [16]. Our data presented here indicate that the use of this method also avoids mis-identification of questionable exitrons (dubbed "falsitrons"), including but not limited to the one in *CD19* exon 2.

Results and discussion

To investigate the processing of *CD19* exon 2, we treated the NALM-6 B-ALL cell line with thapsigargin, which induces unfolded protein response and IRE1 activity [10], and profiled select transcripts by RT-PCR. As anticipated, the levels of the spliced *XBP1* isoform were increased, but we did not detect changes in the reported *CD19* Δex2part product (Additional File 1: Fig. S1a). This called into question the role of IRE1 in exon 2 processing. We therefore decided to investigate aberrant splicing of *CD19* mRNA in B-ALL in more detail. To this end, we performed dRNA-seq and cDNA-seq on the same RNA sample from a therapy-resistant patient-derived xenograft [17] using long-read ONT sequencing. Both datasets documented the occurrence of several previously reported pathological *CD19* isoforms, including exon 2 skipping [2] and intron 2 retention [4]. Surprisingly, we failed to detect the Δex2part product in dRNA-seq, even though it was clearly observed in cDNA-seq (Fig. 1a). This suggested that it may be an artifact of the reverse transcription (RT)/PCR amplification-based protocol. Close examination of the *CD19* exon 2 sequence revealed that the putative exitron could be folding into a stable hairpin flanked by two 8-nt direct repeats (Fig. 1b), hinting at possible RT or PCR slippage at the base of the hairpin and ensuing product truncation.

To test this hypothesis, we engineered a dual-fluorescence GFP/RFP reporter (Fig. 1c) that would allow detection of *CD19* exitron excision by standard RT-PCR, and the corresponding protein product - via restoring the RFP open reading frame detectable by flow cytometry. Consistent with the *CD19* exitron excision being an RT-PCR artifact, we readily observed the corresponding RT-PCR product, but no RFP/GFP double-positive cells upon transfection into HEK293T cells (Fig. 1d, e). In addition, we introduced point mutations that were predicted to either increase the stability of the secondary structure (mut+; $\Delta\Delta G = -5.1$ kcal/mol) or disrupt one of the direct repeats (mut–; Fig. 1b). Consistent with our hairpin hypothesis, these reporter variants altered the levels of the Δex2part product in the RT-PCR-based assay. Namely, they were 82% higher in the case of mut+ or

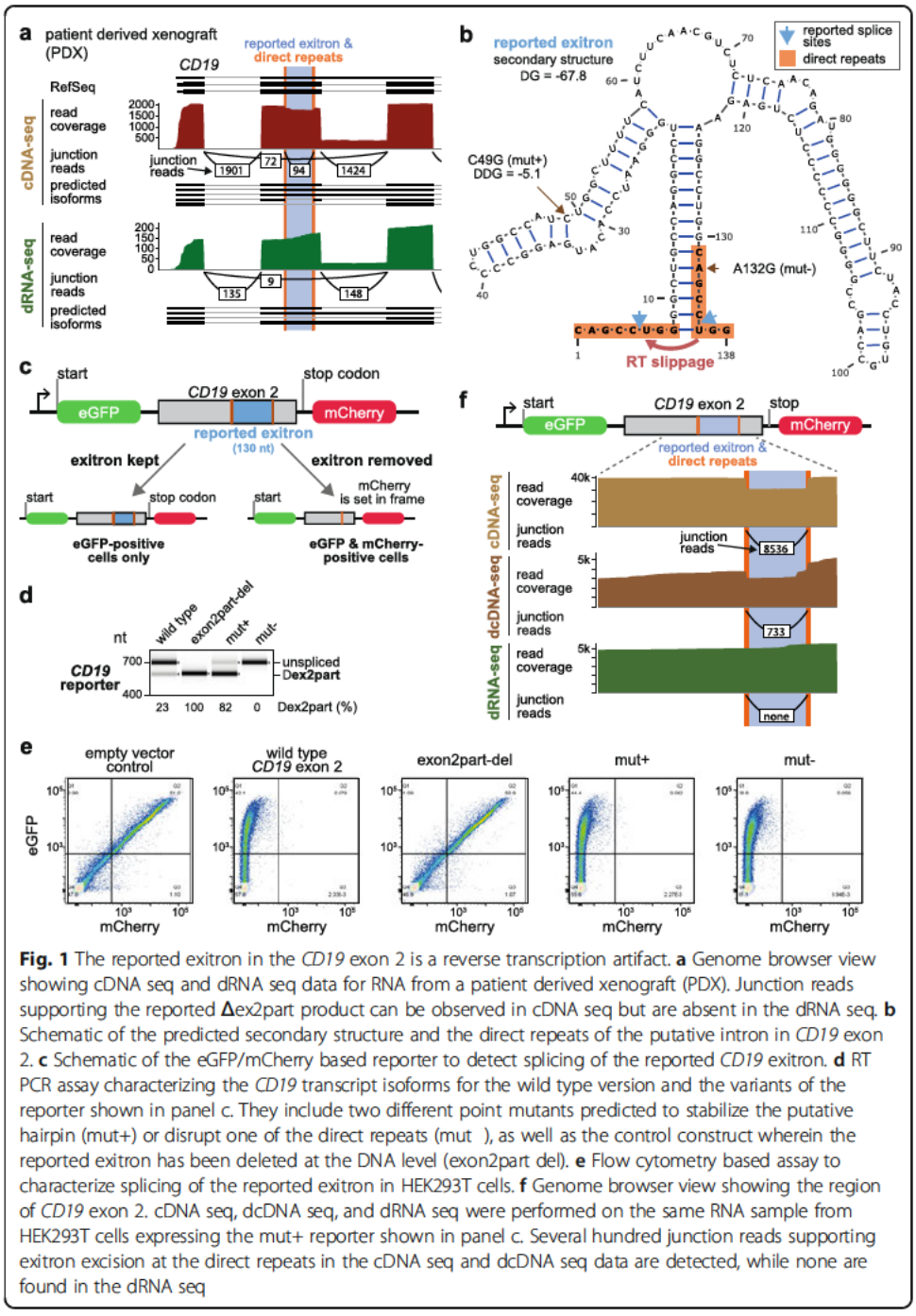


Fig. 1 The reported exitron in the *CD19* exon 2 is a reverse transcription artifact. **a** Genome browser view showing cDNA seq and dRNA seq data for RNA from a patient derived xenograft (PDX). Junction reads supporting the reported Δex2part product can be observed in cDNA seq but are absent in the dRNA seq. **b** Schematic of the predicted secondary structure and the direct repeats of the putative intron in *CD19* exon 2. **c** Schematic of the eGFP/mCherry based reporter to detect splicing of the reported *CD19* exitron. **d** RT PCR assay characterizing the *CD19* transcript isoforms for the wild type version and the variants of the reporter shown in panel c. They include two different point mutants predicted to stabilize the putative hairpin (mut+) or disrupt one of the direct repeats (mut), as well as the control construct wherein the reported exitron has been deleted at the DNA level (exon2part del). **e** Flow cytometry based assay to characterize splicing of the reported exitron in HEK293T cells. **f** Genome browser view showing the region of *CD19* exon 2. cDNA seq, dcDNA seq, and dRNA seq were performed on the same RNA sample from HEK293T cells expressing the mut+ reporter shown in panel c. Several hundred junction reads supporting exitron excision at the direct repeats in the cDNA seq and dcDNA seq data are detected, while none are found in the dRNA seq

completely abolished in the case of mut– (Fig. 1d). Again, neither of them, not even mut+, yielded GFP/RFP double-positive cells (Fig. 1e). As a positive control, we removed the reported exitron from the reporter at the DNA level (exon2part-del) and readily observed both truncated RT-PCR product (Fig. 1d, e; Additional File 1: Fig. S1b, c) and robust expression of RFP (Fig. 1e).

To differentiate between RT and PCR artifacts, we performed dRNA-seq, direct cDNA (dcDNA)-seq omitting PCR amplification, and regular PCR-aided cDNA-seq on the reporter-transfected cells. To rule out the sensitivity issue, we used the mut+

reporter variant, which yields the highest levels of the Δex2part product in RT-PCR (Fig. 1e). Strikingly, in the long-read ONT data, the Δex2part product accounted for > 25% of dcDNA-seq and almost 30% of cDNA-seq reads, but was undetectable using dRNA-seq (Fig. 1f). This direct comparison of sequencing protocols indicated that excision of the reported *CD19* exitron occurs not in live cells, but in the test tube during the RT step, possibly due to the two direct repeats brought together at the base of the predicted hairpin structure. A similar phenomenon has been previously observed in the human *LIP1* and *FOXL2* genes [18, 19].

Our results indicate that RT-based sequencing protocols can lead to the widespread mis-identification of exitrons. Indeed, the *CD19* exitron was recently reported to yield a new isoform in the long-read full-length cDNA-seq dataset obtained using the Rolling Circle Amplification to Concatemeric Consensus (R2C2) method serving to increase detection accuracy [7, 8]. To determine whether other transcripts are prone to such RT artifacts, we performed a targeted search in publicly available ONT sequencing datasets. Specifically, we screened for transcript isoforms that are present only in cDNA-seq but not in the matching dRNA-seq. This was achieved using several filtering steps, such as adjusting for read coverage and excluding the presence of canonical splice sites (Fig. 2a, Additional File 1: Fig. S2a, also see Methods). We first applied this comparison to cDNA-seq and dRNA-seq data for the B-lymphoblastoid cell line GM12878 from the Nanopore RNA Consortium [20]. We readily rediscovered the *CD19* exitron along with 19 other questionable exitrons, which we dubbed "falsitrons" (Fig. 2b, c, Additional File 1: Fig. S2b, Additional File 2: Data 1, Additional File 3: Table S1), supporting the common nature of such artifacts. We then extended our search to ONT sequencing data for five commonly used cell lines from the Singapore Nanopore Expression Project (SG-NEx) [21]: A549, HCT116, HepG2, K562, and MCF-7. In total, we discovered 100 candidate events corresponding to 57 unique falsitrons in 43 genes, for which "spliced" reads were present in the cDNA-seq (up to 70% of reads) but completely absent in the matched dRNA-seq (Fig. 2c, Additional File 2: Data 1, Additional File 3: Table S1). Many of these falsitrons were short (median length 353 nt; Fig. 2d), with the "spliced" regions flanked by direct repeats (35 out of 57; Fig. 2c, e). This discovery strengthens our hypothesis that falsitrons in many instances arise from RT slippage. These artifacts are not restricted to ONT data, but occur in other long-read sequencing protocols such as Iso-Seq (Isoform Sequencing, PacBio) as well [13]. We detected 33 out of 57 falsitrons in the reconstructed isoforms from publicly available Iso-Seq data for several human RNA samples (Alzheimer brain, lymphoblastoid cell line COLO829BL, melanoma cell line COLO829T and Human Universal Reference RNA—see the "Methods" section and Additional File 1: Fig. S2c).

Conceptually, such RT artifacts would not be restricted to long-read cDNA-seq data either and should also be found in conventional short-read RNA-seq protocols. To test this hypothesis, we screened the Cancer Genome Atlas (TCGA) database [22] and immediately found six of the falsitrons in several cancer types. Overall, the abundance of the corresponding isoforms was low (< 5%), but could rise up to > 90% for certain samples and tumor types (Fig. 2f). This is potentially important, because a recent paper reported more than 100,000 exitrons in the TCGA database and suggested that the corresponding isoforms are novel cancer drivers and neoepitopes [23]. To learn whether such analyses might be affected by RT artifacts, we overlaid the falsitrons from

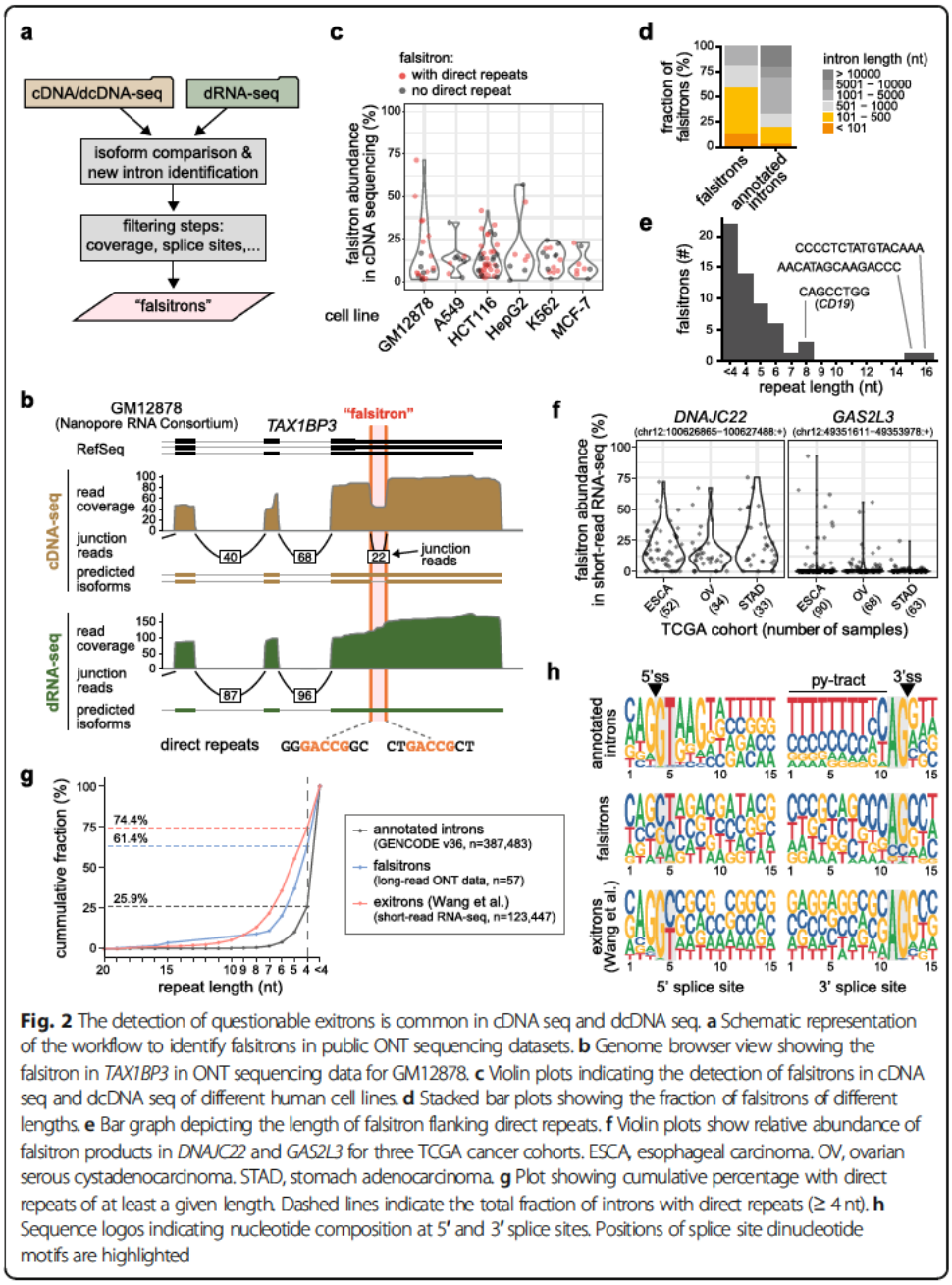


Fig. 2 The detection of questionable exitrons is common in cDNA seq and dcDNA seq. **a** Schematic representation of the workflow to identify falsitrons in public ONT sequencing datasets. **b** Genome browser view showing the falsitron in *TAX1BP3* in ONT sequencing data for GM12878. **c** Violin plots indicating the detection of falsitrons in cDNA seq and dcDNA seq of different human cell lines. **d** Stacked bar plots showing the fraction of falsitrons of different lengths. **e** Bar graph depicting the length of falsitron flanking direct repeats. **f** Violin plots show relative abundance of falsitron products in *DNAJC22* and *GAS2L3* for three TCGA cancer cohorts. ESCA, esophageal carcinoma. OV, ovarian serous cystadenocarcinoma. STAD, stomach adenocarcinoma. **g** Plot showing cumulative percentage with direct repeats of at least a given length. Dashed lines indicate the total fraction of introns with direct repeats (≥ 4 nt). **h** Sequence logos indicating nucleotide composition at 5′ and 3′ splice sites. Positions of splice site dinucleotide motifs are highlighted

our ONT data comparison onto these reported exitrons. We found that five falsitrons, including the *CD19* one, overlapped with reported exitrons. To our surprise, we further detected direct repeats (≥ 4 nt) overlapping the putative splice sites in almost 75% of the reported exitrons (91,852 out of 123,337; median length 5 nt), i.e. even more than in our falsitron list (with the shorter median length of 4 nt; Fig. 2g). In contrast, only ~ 25% of all annotated introns harbored such direct repeats at their splice sites (median length < 4 nt). Moreover, even though exitrons had been selected for canonical splice site dinucleotides (GU/GC-AG), they lacked other characteristics of 5′ and 3′ splice sites such as U1 complementarity and the polypyrimidine tract (Fig. 2h). This finding indicates that a significant fraction of the reported exitrons could also be RT artifacts.

Although this observation awaits experimental validation, it suggests that caution is required when interpreting RNA-seq mapping data. We envision that as more dRNA-seq data become available, the unequivocal classification of cryptic introns as exitrons or falsitrons will be possible.

Conclusions

Here, we show that RT artifacts can lead to the detection of questionable exitrons ("falsitrons") and non-existing transcript isoforms. Such artifacts are not limited to one study and occur reproducibly in all protocols which rely on RT, including standard RT-PCR and short-read RNA-seq, but also in ONT-based sequencing of cDNA (PCR-amplified or not). For laboratories looking to validate specific exitrons, utilization of thermo-stable reverse transcriptases (as in TGIRT-Seq [24]) and Northern blotting can be used to avoid artifacts, especially when exitrons in question are reasonably long. Moreover, at least one computational tool (SQANTI) has been developed to flag suspicious introns by implementing a machine learning classifier based on a variety of transcript descriptors [25]. For example, in the publicly available Iso-Seq dataset (PacBio) from the lymphoblastoid cell line COLO829BL derived from a melanoma patient [26], SQANTI2 correctly filters out the *CD19* falsitron (Additional File 1: Fig. S2c). However, such flagging could come at the expense of filtering out real exitrons. Thus, in our opinion, dRNA-seq should be utilized beyond RNA modification detection as a reliable validation tool for high-throughput transcriptome analysis. While it requires significant amount of input RNA and typically yield fewer reads, it does not pick up falsitrons and allows for a more accurate cataloging of bona fide transcript isoforms. As our work illustrates, the accuracy is particularly important when putative isoforms have clinical correlates, such as resistance to life-saving immunotherapies.

Methods

Cell lines and patient-derived xenografts

HEK293T cells were obtained from DSMZ. They were cultured in DMEM (Life Technologies) with 10% fetal bovine serum (Life Technologies) and 1% L-glutamine (Life Technologies). NALM-6 cells were obtained from ATCC and cultured in RPMI medium with the same additives as for HEK293T cells. All cells were kept at 37 °C in a humidified incubator containing 5% CO_2. They were routinely tested for mycoplasma infection. Viably-cryopreserved cells from a patient-derived xenograft model of human B-ALL harboring a TCF3-HLF fusion (ALL1807) were established as previously described [17] and used for downstream sequencing studies.

Cloning

The backbone of the splicing reporter (including both fluorophores) was generously provided by Ramanujan S. Hegde (MRC Laboratory of Molecular Biology, Cambridge, UK) [27]. We introduced exon 2 and part of exon 3 of the human *CD19* gene between GFP and mCherry. To this end, we amplified the *CD19* exon 2 insert sequence from human genomic DNA (Promega) with the following primers:

5′-GATGACGATGACAAGGCCGGATCTGGAGATAACGCTGTGCTGCA-3′ and 5′-GCCAACTTTGAGCCCAGGTGAATCGGTCCGAAACATTCCACCGGAACAGC

TCCCCGCTGCCCTCCACATTGACT-3′. The backbone was amplified with the following primers 5′-GATTCACCTGGGCTCAAAGT-3′ and 5′-AGATCCGGCCTTGT CATCGT-3′. The amplification products were combined using Gibson assembly ready-made master mix from IMB Protein Production Core Facility. The generation of point mutations in the splicing reporter was achieved with the Q5 Site-Directed Mutagenesis Kit (New England Biolabs) according to the manufacturer's recommendations.

Dual-fluorescence splicing reporter assay via flow cytometry

Overexpression of the reporter plasmid was performed using Lipofectamine 2000 (Life Technologies) according to the manufacturer's recommendation. Samples were transfected with reporter plasmids 48 h prior to flow cytometric analysis. Cells were washed in DPBS and trypsinized. After centrifugation, cells were washed twice with Dulbecco's phosphate-buffered saline (DPBS) and resuspended in FACS buffer (DPBS, 1% BSA and 2 mM EDTA). Experiments were performed on the LSRFortessa SORP (BD Biosciences) and analyzed via the FlowJo (v10) software (FlowJo, LLC).

Thapsigargin assay

Thapsigargin (Biomol GmbH) was used after 24 h post-transfection at a concentration of 250 nM for 2, 6, and 24 h on NALM-6 cells. Afterwards, cells were harvested and washed twice in PBS. RNA was isolated with the RNeasy Plus Mini Kit (Qiagen).

Quantification of splicing isoforms with RT-PCR

Semiquantitative RT-PCR was used to quantify ratios of *CD19* and *XBP1* mRNA isoforms. To this end, reverse transcription was performed on 500 ng RNA with RevertAid Reverse Transcriptase (Thermo Fisher Scientific) according to the manufacturer's recommendations. Subsequently, 1 μl of the cDNA was used as template for the RT-PCR reaction with the OneTaq DNA Polymerase (New England Biolabs) (Cycler conditions: 94 °C for 30 s, 28 cycles [reporter PCR] or 34 cycles [endogenous *CD19*, *XBP1*] of [94 °C for 20 s, 53 °C [reporter assay] or 55 °C [*CD19* endogenous] or 54 °C [*XBP1*] for 30 s, 68 °C for 30 s] and final extension at 68 °C for 5 min). The primers 5′-CGCGATCACA TGGTCCTTAA-3′ and 5′-CATGTTATCCTCCTCGCCCT-3′ were used for the reporter assay, 5′-ACCTCCTCGCCTCCTCTTCTTC-3′ and 5′-CCGAAACATTCCAC CGGAACAGC-3′ for the endogenous PCR on *CD19* and 5′-CCTGGTTGCTGAA-GAGGAGG-3′ and 5′-CCATGGGGAGATGTTCTGGAG-3′ for *XBP1*. The TapeStation 2200 capillary gel electrophoresis instrument (Agilent) was used for quantification of the PCR products on D1000 tapes.

Nanopore sequencing

For the ONT sequencing of the PDX sample ALL1807 or HEK293T cells transfected with the mut+ reporter construct, total RNA was extracted using Trizol reagent following manufacturer's recommendation. The mRNA was isolated from 100 μg of total RNA using Dynabeads mRNA DIRECT Kit (Invitrogen). The mRNA samples were subjected to PCR-cDNA (SQK-PCS109, ONT), direct-cDNA (SQK-DCS109, ONT) and direct-RNA (SQK-RNA002, ONT) library preparation in parallel using the equipment and consumables according to each library protocol. Subsequently, each library was

loaded into a Spot-ON flow cell R9 Version (FLO-MIN106D, ONT) and sequenced on a MinION Mk1B device (ONT) for 48 h. The RNA from the sample ALL1807 was submitted to the Sequencing Technologies and Analysis Core at Cold Spring Harbor Laboratory for PCR-cDNA library preparation and sequencing on a PromethION device (ONT).

Nanopore sequence analysis

Base calling was performed using the ONT data processing toolkit guppy (version 3.4.5). guppy basecaller was run with default settings providing the specific flow cell and library preparation pairs. The resulting reads were aligned to either the human reference genome (version hg38) or our custom *CD19* reporter (mut+) sequence using minimap2 (version 2.17-r941) [28], using the following flags "-k 12 -u b -x splice --secondary=no". For downstream transcriptome analysis, we used the ONT pipeline [github.com/nanoporetech/pipeline-nanopore-ref-isoforms], which implements preprocessing with pychopper (DNA only), mapping with minimap2 and transcriptome reconstruction with StringTie [29] in long-read mode. Finally, the annotation obtained from StringTie was compared back to the existing annotation using gffcompare [30]. This pipeline was modified to run StringTie without annotation to guide the reconstruction and we omitted the "--conservative" flag.

ONT data comparison to identify falsitrons

In order to identify additional falsitrons, we compared cDNA-seq and dRNA-seq data produced by the Nanopore RNA Consortium [20] and the Singapore Nanopore Expression Project (SG-NEx) [21]. The first dataset from the Nanopore RNA Consortium contains dRNA-seq and cDNA-seq data for the cell line GM12878. SG-NEx offers cDNA-seq, dcDNA-seq, and dRNA-seq for the five commonly used cell lines A549, HCT116, HepG2, K562 and MCF-7. For each dataset, we used StringTie for isoform reconstruction as described above. For read filtering, we used the default parameters specified in the pipeline: --minimum mapping quality 40, --poly context 24, and --max poly run 8. We then contrasted the GFF transcript output files from StringTie using gffcompare which provides a summary of all the distinct isoforms between two GFF files. We searched for falsitrons that are supported by "spliced" reads only in cDNA-seq but not in dRNA-seq. To do this, we inspected the pairs of "non-equal" isoforms for junction-spanning reads that were present only in cDNA-seq and were fully contained within an exon (filter 1a, Additional File 1: Fig. S2a) or had start and end coordinates that were resided in two adjacent exons detected in the dRNA-seq (filter 1b, Additional File 1: Fig. S2a). Based on the characteristics of *CD19* Δexon2part, we applied additional filters, i.e. a minimum coverage of five reads of both cDNA-seq and dRNA-seq (as reported by StringTie), and the lack of canonical GU-AG splice sites. Using these search criteria, we identified 100 candidate events arising from 57 unique putative falsitrons. Of those, 35 contained direct repeats in the splice sites ranging from 3 to 16 nt, similar to the 8-nt repeats in *CD19* Δex2part. Read numbers, mapping statistics, and gffcompare results for the samples are reported in Additional File 4: Table S2. Genome browser views showing ONT cDNA-seq and dRNA-seq data from all putative falsitron

events are shown in Additional file 2: Data 1. The code for the falsitron search is available in Zenodo/Github under an open source MIT license [31, 32].

Direct repeat search

For each candidate event, we searched for the presence of the same *k*-mers with length from 4 to 20 nt in a 40-nt window around each splice site. The *k*-mers were required to overlap at least 1 nt of the 5′ and 3′ dinucleotide motifs. The same analysis was applied to all the exitrons detected in Wang et al. [23] as well as for all unique annotated introns in GENCODE gene annotation (v36, genome version hg38) [33].

Junction search in TCGA

We use the R/Bioconductor package snapcount [https://github.com/langmead-lab/snapcount] to query the 57 putative falsitrons from our ONT data comparison in short-read RNA-seq data from the Cancer Genome Atlas (TCGA) database. As most of the putative falsitrons end in repetitive regions, like in the case of *CD19* Δex2part, we allowed the splice sites to be shifted outwards by an offset of up to 1 repeat length of that given intron, as long as the resulting junction did not differ by more than ± 1 repeat length from the original junction length. Following these filters, we detected six of our putative falsitrons in TCGA. These reside in the following genes (genomic coordinates of falsitron in brackets): *PHAX* (chr5:126625543-126625746:+), *CCDC86* (chr11:60842626-60842700:+), *DNAJC22* (chr12:49351611-49353978:+), *GAS2L3* (chr12:100626865-100627488:+), *CDC27* (chr17:47118517-47118594:−), and *H1F0* (chr22:37807089-37807354:+).

Relative isoform abundance estimates

For the long-read ONT data, relative isoform abundance was calculated by dividing the number of split reads supporting the falsitron junction over the total number of reads overlapping the junction coordinates. Operations were performed using the R/Bioconductor package GenomicAlignments [34]. For the TCGA data, we calculated relative isoform abundances by dividing the spliced reads (quantified using snapcount) over the mean of reads overlapping the junction region. The latter were quantified with data from the ReCount database [35] via the R/Bioconductor packages megadepth and recount3 [36].

Nucleotide composition at splice sites

For the sequence logos at splice sites, we retrieved the sequence in a 15-nt window (3 nt in the exon + 12 nt in the intron) of the 3′ and 5′ splice sites of the different sets of introns: our putative falsitrons from the ONT comparison (n = 57), all unique exitrons reported by Wang et al. [23] (n = 123,337) and all unique introns in GENCODE gene annotation (v36, genome version hg38) (n = 387,483). We used the R package ggseqlogo [37] to plot the frequency of nucleotides in each set.

Analysis of Iso-Seq data

Isoform predictions for Iso-Seq data (PacBio Sequel) before and after SQANTI2 filtering (v2.7) were taken from https://github.com/PacificBiosciences/DevNet/wiki/Melanoma%2D%2DCancer-Cell-Line-Iso-Seq-Data (for the lymphoblastoid cell line

COLO829BL and melanoma COLO829T; PacBio Sequel), and https://downloads.pacbcloud.com/public/dataset/Alzheimer2019 IsoSeq/ (for total RNA from an Alzheimer's Disease brain sample; PacBio Sequel II). The Universal Human Reference (Agilent; PacBio Sequel II) did not contain the SQANTI2 correction in the initial 2019 release (https://downloads-ap.pacbcloud.com/public/dataset/UHR IsoSeq/). Upon request, we obtained a 2021 version of the annotation, filtered with SQANTI3 (https://downloads.pacbcloud.com/public/dataset/UHRRisoseq2021/). In the filtered files only 4 falsitrons were detected, located in the following genes: *DNAJC22* (chr5:126625543-126625746:+), *GAS2L3* (chr12:49351611-49353978:+), *CDC27* (chr12:100626865-100627488:+), *PHAX* (chr17:47118517-47118594:−).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059 021 02411 1.

Additional file 1: Figure S1. Levels of the Δex2part product are not affected by thapsigargin treatment. a) RT PCR experiments followed by capillary electrophoresis to quantify different *CD19* and *XBP1* isoforms. NALM 6 cells were treated with thapsigargin for indicated time intervals. b) RT PCR experiments followed by capillary electro phoresis to quantify different *CD19* isoforms in HEK293T cells transfected with a mixture of mut (A; does not pro duce Δex2partband) and exon2part del (B; the reported intron is removed at the DNA level) reporter constructs. c) Flow cytometry based assay performed on the same cells. **Figure S2.** The workflow to detect falsitrons captures the truncated *CD19* Δex2part product. a) Extended schematic representation of the workflow to identify question able exitrons (dubbed "falsitrons"). b) Genome browser view depicting detection of the *CD19* falsitron (Δex2part) in ONT cDNA seq, but not dRNA seq data from the Nanopore RNA Consortium. c) Genome browser view shows that the *CD19* falsitron (Δex2part) is detected in PacBio Iso Seq experiments but is filtered out when applying SQANTI2.

Additional file 2: Data 1. Putative falsitrons in the genomic context.

Additional file 3: Table S1. Putative falsitrons detected from Oxford Nanopore Technologies (ONT) sequencing data for the five commonly used cell lines A549, HCT116, HepG2, K562 and MCF 7, as well as the B lymphoblastoid cell line GM12878.

Additional file 4: Table S2. Mapping and gffcompare statistics for Oxford Nanopore Technologies (ONT) sequencing datasets used in this study.

Additional file 5. Review history.

Acknowledgements

We thank all members of the Thomas Tikhonenko, König, and Zarnack groups for many helpful discussions. We also gratefully acknowledge support of the IMB Bioinformatics and Flow Cytometry Core Facilities. The results published here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Review history

The review history is available as additional file 5.

Peer review information

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

LS generated the *CD19* reporter assay and performed RT PCR and flow cytometry measurements. MTD and MA per formed ONT sequencing on HEK293T cells. MCL conceived and implemented computational analysis strategies for fal sitron identification in ONT sequencing data, TCGA search, Iso Seq analysis, and exitron characterization. KEH performed ONT data analysis for PDX and HEK293T data. SKT has developed the ALL1807 PDX model. YB contributed to the analysis of short read RNA seq data. All authors contributed to the design of the study. KZ, JK, ES, and ATT wrote the manuscript with input from all coauthors. All authors read and approved the final manuscript.

Authors' information

Twitter handles: @koenig lab (Julian König); @andrei thomas t (Andrei Thomas Tikhonenko).

Funding

Research in the ATT laboratory was supported by grants from the NIH (U01 CA232563), St. Baldrick's Stand Up to Cancer Pediatric Dream Team (SU2C AACR DT 27 17), the V Foundation for Cancer Research (T2018 014), and the Cookies for Kids' Cancer (CFKC) Foundation. ATT is Richard "Buz" Cooper Scholar of the Breakthrough Bike Challenge. SKT was supported by grants from NIH (U01 CA232486 and U01 CA243072), Department of Defense (CA180683P1), Philip A Sharp Award for Innovation in Collaboration, and Simutis family fund for childhood leukemia research. YB's work was supported by grants from the NIH (U01 CA232563, R01 LM013437, R01 GM128096). Research in the KZ

group and the JK laboratory was supported by grants from the German Research Foundation/Deutsche Forschungsgemeinschaft (ZA 881/2 1 and KO 4566/4 1, respectively).

Availability of data and materials

The long read ONT sequencing data for the PDX sample ALL1807 (cDNA seq and dRNA seq) and the HEK293T cells transfected with the mut+ reporter construct (cDNA seq, dcDNA seq and RNA seq) are avaible in NCBI Short Read Archive under accession numbers SRR14326969 14326973 [38]:
https://www.ncbi.nlm.nih.gov/sra/?term SRR14326969
https://www.ncbi.nlm.nih.gov/sra/?term SRR14326970
https://www.ncbi.nlm.nih.gov/sra/?term SRR14326971
https://www.ncbi.nlm.nih.gov/sra/?term SRR14326972
https://www.ncbi.nlm.nih.gov/sra/?term SRR14326973
The computational code for the detection of falsitrons in ONT Seq data is available in Zenodo/Github under an open source MIT license (https://doi.org/10.5281/zenodo.4906610) [31, 32].

Declarations

Ethics approval and consent to participate

Primary leukemia cells from the patient have been previously banked at the Children's Hospital of Philadelphia Center for Childhood Cancer biorepository with informed consent in accordance with the Declaration of Helsinki via IRB approved research protocols.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz, Germany. [2]Division of Cancer Pathobiology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [3]The Bioinformatics Group, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [4]Division of Oncology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [5]Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA. [6]Present address: Stanford Cancer Institute, 265 Campus Dr., Stanford, CA 94305, USA. [7]Buchmann Institute for Molecular Life Sciences (BMLS) and Faculty of Biological Sciences, Goethe University Frankfurt, Max von Laue Str. 15, 60438 Frankfurt, Germany. [8]Department of Pathology & Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA.

Received: 27 April 2021 Accepted: 16 June 2021
Published online: 28 June 2021

References

1. Maude SL, Laetsch TW, Buechner J, Rives S, Boyer M, Bittencourt H, et al. Tisagenlecleucel in children and young adults with B cell lymphoblastic leukemia. N Engl J Med. 2018;378(5):439 48. https://doi.org/10.1056/NEJMoa1709866.
2. Sotillo E, Barrett DM, Black KL, Bagashev A, Oldridge D, Wu G, et al. Convergence of acquired mutations and alternative splicing of CD19 enables resistance to CART 19 immunotherapy. Cancer Discov. 2015;5(12):1282 95. https://doi.org/10.1158/2159 8290.CD 15 1020.
3. Bagashev A, Sotillo E, Tang C HA, Black KL, Perazzelli J, Seeholzer SH, et al. CD19 alterations emerging after CD19 directed immunotherapy cause retention of the misfolded protein in the endoplasmic reticulum. Mol Cell Biol. 2018;38: e00383 18.
4. Asnani M, Hayer KE, Naqvi AS, Zheng S, Yang SY, Oldridge D, et al. Retention of CD19 intron 2 contributes to CART 19 resistance in leukemias with subclonal frameshift mutations in CD19. Leukemia. 2020;34(4):1202 7. https://doi.org/10.1038/s41375 019 0580 z.
5. Rabilloud T, Potier D, Pankaew S, Nozais M, Loosveld M, Payet Bornet D. Single cell profiling identifies pre existing CD19 negative subclones in a B ALL patient with CD19 negative relapse after CAR T therapy. Nat Commun. 2021;12(1):865. https://doi.org/10.1038/s41467 021 21168 6.
6. Zhao Y, Aldoss I, Qu C, Crawford JC, Gu Z, Allen EK, et al. Tumor intrinsic and extrinsic determinants of response to blinatumomab in adults with B ALL. Blood. 2021;137(4):471 84. https://doi.org/10.1182/blood.2020006287.
7. Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, et al. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full length single cell cDNA. Proc Natl Acad Sci USA. 2018; 115(39):9726 31. https://doi.org/10.1073/pnas.1806447115.
8. Cole C, Byrne A, Adams M, Volden R, Vollmers C. Complete characterization of the human immune cell transcriptome using accurate full length cDNA sequencing. Genome Res. 2020;30(4):589 601. https://doi.org/10.1101/gr.257188.119.
9. Boissel N. ALL in escape room. Blood. 2021;137(4):432 4. https://doi.org/10.1182/blood.2020008850.
10. Maurel M, Chevet E, Tavernier J, Gerlo S. Getting RIDD of RNA: IRE1 in cell fate regulation. Trends Biochemical Sci. 2014; 39(5):245 54. https://doi.org/10.1016/j.tibs.2014.02.008.
11. Marquez Y, Höpfler M, Ayatollahi Z, Barta A, Kalyna M. Unmasking alternative splicing inside protein coding exons defines exitrons and their role in proteome plasticity. Genome Res. 2015;25(7):995 1007. https://doi.org/10.1101/gr.186585.114.

12. Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, et al. Comprehensive transcriptome analysis using synthetic long read sequencing reveals molecular co association of distant splicing events. Nat Biotechnol. 2015;33(7): 736 42. https://doi.org/10.1038/nbt.3242.
13. Sharon D, Tilgner H, Grubert F, Snyder M. A single molecule long read survey of the human transcriptome. Nat Biotechnol. 2013;31(11):1009 14. https://doi.org/10.1038/nbt.2705.
14. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, et al. Nanopore long read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat Commun. 2017;8(1):16027. https://doi.org/10.1038/ncomms16027.
15. Hu T, Chitnis N, Monos D, Dinh A. Next generation sequencing technologies: An overview. Human Immunol. 2021. https://doi.org/10.1016/j.humimm.2021.02.012.
16. Liu H, Begik O, Lucas MC, Ramirez JM, Mason CE, Wiener D, et al. Accurate detection of m6A RNA modifications in native RNA sequences. Nat Commun. 2019;10(1):4079. https://doi.org/10.1038/s41467 019 11713 9.
17. Hurtz C, Wertheim GB, Loftus JP, Blumenthal D, Lehman A, Li Y, et al. Oncogene independent BCR like signaling adaptation confers drug resistance in Ph like ALL. J Clin Invest. 2020;130(7):3637 53. https://doi.org/10.1172/JCI134424.
18. Cocquet J, Chong A, Zhang G, Veitia RA. Reverse transcriptase template switching and false alternative transcripts. Genomics. 2006;88(1):127 31. https://doi.org/10.1016/j.ygeno.2005.12.013.
19. Zhang YJ, Pan HY, Gao SJ: Reverse transcription slippage over the mRNA secondary structure of the LIP1 gene. Biotechniques. 2001;31:1286.
20. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods. 2019;16(12):1297 305. https://doi.org/10.1038/s41592 019 0617 2.
21. Chen Y, Davidson NM, Wan YK, Patel H, Yao F, Low HM, Hendra C, Watten L, Sim A, Sawyer C, et al: A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. bioRxiv. 2021:2021. 2004.2021.440736.
22. Sanchez Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic signaling pathways in The Cancer Genome Atlas. Cell. 2018;173(2):321 37 e310. https://doi.org/10.1016/j.cell.2018.03.035.
23. Wang TY, Liu Q, Ren Y, Alam SK, Wang L, Zhu Z, et al. A pan cancer transcriptome analysis of exitron splicing identifies novel cancer driver genes and neoepitopes. Mol Cell. 2021;81(10):2246 2260.e12. https://doi.org/10.1016/j.molcel.2021.03.028.
24. Qin Y, Yao J, Wu DC, Nottingham RM, Mohr S, Hunicke Smith S, Lambowitz AM: High throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. RNA. 2015;22(1):111 28. https://doi.org/10.1261/rna.054809.115.
25. Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo Palacios FJ, Del Risco H, et al. SQANTI: extensive characterization of long read transcript sequences for quality control in full length transcriptome identification and quantification. Genome Res. 2018;28(3):396 411. https://doi.org/10.1101/gr.222976.117.
26. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010;463(7278):191 6. https://doi.org/10.1038/nature08658.
27. Juszkiewicz S, Hegde RS: Initiation of quality control during poly(A) translation requires site specific ribosome ubiquitination. Mol Cell. 2017; 65:743 750.e744.
28. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094 100. https://doi.org/10.1093/bioinformatics/bty191.
29. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA seq reads. Nat Biotechnol. 2015;33(3):290 5. https://doi.org/10.1038/nbt.3122.
30. Pertea G, Pertea M: GFF Utilities: GffRead and GffCompare. F1000Res 2020; 9.
31. Cortés López M:IntronArtifacts. Github. 2021. https://github.com/mcortes lopez/IntronArtifacts/tree/v1.0.1.
32. Cortés López M:IntronArtifacts: exon2part. Zenodo. 2021. https://zenodo.org/record/4906611.
33. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):D766 d773. https://doi.org/10.1093/nar/gky955.
34. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118. https://doi.org/10.1371/journal.pcbi.1003118.
35. Frazee AC, Langmead B, Leek JT. ReCount: A multi experiment resource of analysis ready RNA seq gene count datasets. BMC Bioinformatics. 2011;12(1):449. https://doi.org/10.1186/1471 2105 12 449.
36. Wilks C, Ahmed O, Baker DN, Zhang D, Collado Torres L, Langmead B. Megadepth: efficient coverage quantification for BigWigs and BAMs. Bioinformatics. 2021. https://doi.org/10.1093/bioinformatics/btab152.
37. Wagih O. ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics. 2017;33(22):3645 7. https://doi.org/10.1093/bioinformatics/btx469.
38. Schulz L; Torres Diz, M; Cortés López, M; Hayer, KE; Asnani, M; Tasian, SK; Barash, Y; Sotillo, E; Zarnack, K; König, J; Thomas Tikhonenko, A:Direct long read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts. Datasets. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/sra/?term SRR14326969; https://www.ncbi.nlm.nih.gov/sra/?term SRR14326970; https://www.ncbi.nlm.nih.gov/sra/?term SRR14326971; https://www.ncbi.nlm.nih.gov/sra/?term SRR14326972; https://www.ncbi.nlm.nih.gov/sra/?term SRR14326973. (2021)

Publisher's Note

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com

Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts

Laura Schulz[1*], Manuel Torres-Diz[2*], Mariela Cortés-López[1*], Katharina E. Hayer[3*], Mukta Asnani[2], Sarah K. Tasian[4], Yoseph Barash[5], Elena Sotillo[2&], Kathi Zarnack[6], Julian König[1#], and Andrei Thomas-Tikhonenko[2 7 #]

[1] Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz, Germany. [2] Division of Cancer Pathobiology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, US. [3] The Bioinformatics Group, Children's Hospital of Philadelphia, Philadelphia, PA 19104, US. [4] Division of Oncology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, US. [5] Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, US. [6] Buchmann Institute for Molecular Life Sciences (BMLS), Goethe University Frankfurt, Max-von-Laue-Str. 15, 60438 Frankfurt, Germany. [7] Department of Pathology & Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, US.

* These authors contributed equally.

[&] Present address: Stanford Cancer Institute, 265 Campus Dr., Stanford, CA 94305

[#] Corresponding authors: Julian König (j.koenig@imb-mainz.de) and Andrei Thomas-Tikhonenko (andreit@pennmedicine.upenn.edu)

SUPPLEMENTARY FIGURES

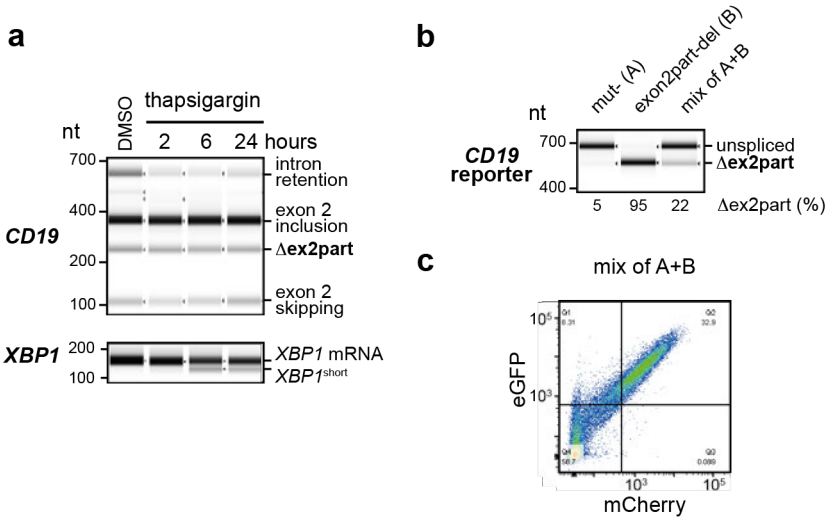


Figure S1. Levels of the Δex2part product are not affected by thapsigargin treatment. a) RT-PCR experiments followed by capillary electrophoresis to quantify different *CD19* and *XBP1* isoforms. NALM-6 cells were treated with thapsigargin for indicated time intervals. **b)** RT-PCR experiments followed by capillary electrophoresis to quantify different *CD19* isoforms in HEK293T cells transfected with a mixture of mut- (A; does not produce Δex2part band) and exon2part-del (B; the reported intron is removed at the DNA level) reporter constructs. **c)** Flow cytometry-based assay performed on the same cells.

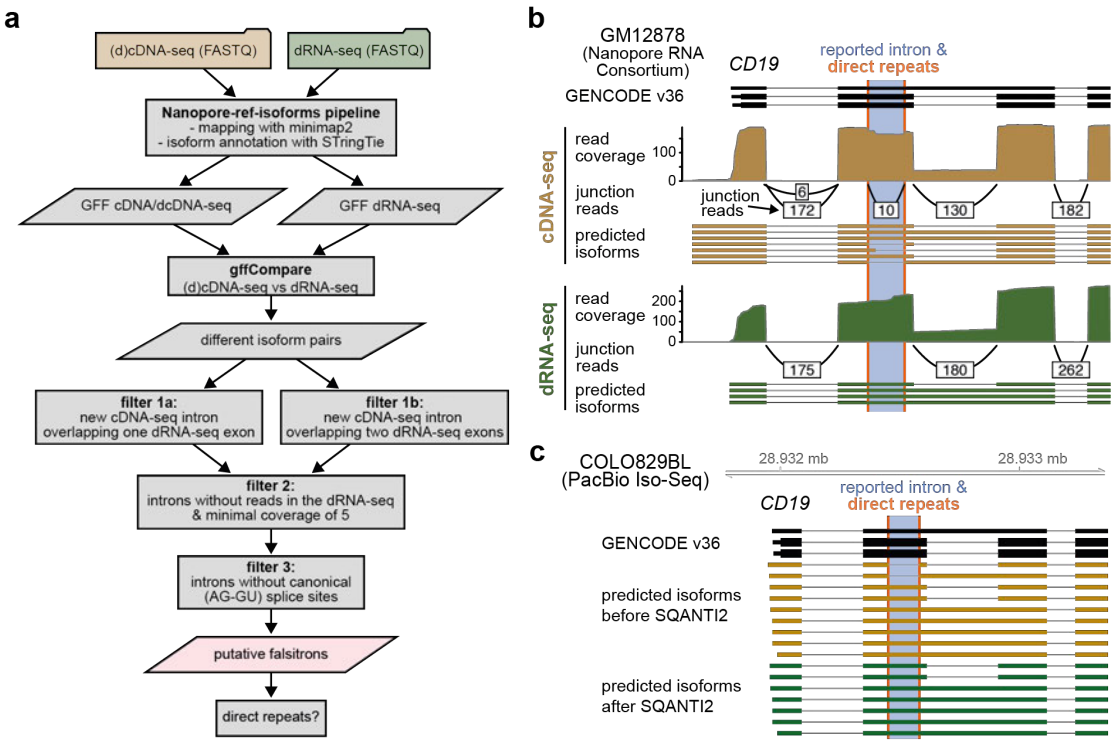


Figure S2. The workflow to detect falsitrons captures the truncated *CD19* Δex2part product. a) Extended schematic representation of the workflow to identify questionable exitrons (dubbed "falsitrons"). **b)** Genome browser view depicting detection of the *CD19* falsitron (Δex2part) in ONT cDNA-seq, but not dRNA-seq data from the Nanopore RNA Consortium. **c)** Genome browser view shows that the *CD19* falsitron (Δex2part) is detected in PacBio Iso-Seq experiments but is filtered out when applying SQANTI2.

2.3 Mutations and RNA-binding proteins controlling *CD19* splicing and CART-19 therapy resistance

2.3.1 Abstract

CD19-targeted CAR-T cell therapy is the first cell-based immunotherapy and represents a breakthrough in the treatment of patients with B cell acute lymphoblastic leukemia. Unfortunately, many patients still relapse after an initial response because the CD19 epitope on the surface of the cancer cells has been lost. *CD19* exon 2 is part of the coding sequence for the epitope. Therefore, incorrect splicing of exon 2 may be the reason for epitope loss and thus tumor escape.

To decipher the regulatory code of *CD19* splicing, we applied a massively parallel reporter assay to a *CD19* minigene comprising exons 1 to 3 including both interspersed introns. Computational modeling was then used to examine the splicing effects of all single-nucleotide mutations in the minigene. As a result, we found ~200 mutations that lead to a strong increase in therapy-relevant splice isoforms such as exon 2 skipping and second intron retention. These mutations could predispose patients undergoing therapy to relapse. In addition, *CD19* exons 1-3 give also rise to ~100 previously unknown cryptic splicing isoforms. These new cryptic isoforms most likely encode non-functional CD19 proteins. This in turn could decrease CD19 epitope presentation and affect the long-term success of CART-19.

Furthermore, we used bioinformatic binding motif analyses to search for possible *trans*-regulatory elements that might play a role in *CD19* splicing. Knockdown experiments revealed the importance of SRSF3, PTBP1, and other RBPs in affecting therapy-related splicing isoforms. This study presents potential prognostic markers important for assessing the risk of CART-19 resistance.

2.3.2 Zusammenfassung

Die CD19-gerichtete CAR-T-Zelltherapie ist die erste zellbasierte Immuntherapie und stellt einen Durchbruch in der Behandlung von Patienten mit akuter lymphatischer B-Zell-Leukämie dar. Trotz eines anfänglichen Ansprechens auf die Therapie erleiden viele Patienten immer noch einen Rückfall. Grund dafür ist unter anderem der Verlust des CD19-Epitops auf der Oberfläche der Tumorzellen. *CD19* Exon 2 ist Teil der kodierenden

Sequenz für das Epitop, das die CAR-T-Zelltherapie benutzt. Somit könnte ein fehlerhaftes Spleißen von Exon 2 der Grund für den Epitopverlust und damit für einen Rückfall sein.

Um das *CD19*-Spleißen zu entschlüsseln, haben wir einen Hochdurchsatz-Reporter-Assay auf ein *CD19*-Minigen angewandt, das die Exons 1 bis 3 einschließlich der beiden dazwischen gelegenen Introns umfasst. Mit Hilfe von mathematischer Modellierung wurden die Spleißeffekte aller Einzelnukleotid-Mutationen in dem Minigen entschlüsselt. Als Ergebnis fanden wir 193 Mutationen, die zu einer starken Zunahme von therapie-relevanten Spleißisoformen wie "Exon 2-Skipping" und "Intron 2-Retention" führen. Diese Mutationen könnten Patienten, die sich einer Therapie unterziehen, für einen Rückfall prädisponieren. Darüber hinaus fanden wir ~100 bisher unbekannte kryptische CD19 Spleißisoformen in den Daten. Diese neuen kryptischen Isoformen kodieren höchstwahrscheinlich nicht-funktionale CD19-Proteine. Dies wiederum könnte die Anzahl der CD19-Epitope auf den Zellen verringern und damit den langfristigen Erfolg von CART-19 beeinträchtigen.

Außerdem haben wir mit Hilfe von bioinformatischen Bindungsmotiv-Analysen nach möglichen *trans*-regulatorischen Elementen gesucht, die eine Rolle beim *CD19*-Spleißen spielen könnten. Knockdown-Experimente zeigten den Effekt von SRSF3, PTBP1 und anderen RNA-bindenden Proteinen auf therapie-relevante Spleißisoformen. Diese Studie stellt potenzielle prognostische Marker vor, die für die Bewertung des Risikos einer CART-19-Resistenz wichtig sind.

2.3.3 Statement of contribution

This was the main project shared between [redacted] and me. I performed most of the experimental work on this publication; [redacted] did the majority of the bioinformatics analyses. I generated the original *CD19* minigene, performed the mutagenesis, prepared the DNA-Seqs, the transfections, RNA extractions, emulsion PCRs, and Amplicon-RNA-Seqs. I created the minigenes for validation of the cryptic isoforms as well as the minigenes representing clinical patient mutations. I then performed all downstream experiments as well as their analysis.

In addition, I supervised [redacted], the master student in our lab at that time, and introduced her to the *CD19* experiments. Together with her, we created all shRNA-containing plasmids and performed all viral transduction-related experiments in the S2-lab leading to eleven inducible RBP knockdown cell lines of NALM-6. We verified their

functionality by qPCR. We examined these cell lines for their *CD19* splicing pattern by RT-PCR and analyzed the samples via capillary gel electrophoresis.

Overall, I contributed to the design of the study and interpretation of results during our biweekly project meetings, which [redacted] and I organized and prepared together. In addition, I wrote the main text and methods of the manuscript belonging to my experiments and contributed to and reviewed the other parts. I also created the corresponding figures and reviewed the other figures.

Mutations and RNA-binding proteins controlling *CD19* splicing and CART-19 therapy resistance

Mariela Cortés-López[1#], Laura Schulz[1#], Mihaela Enculescu[1#], Claudia Paret[2], Bea Spiekermann[1], Anke Busch[1], Anna Orekhova[1], Fridolin Kielisch[1], Mathieu Quesnel-Vallières[3], Manuel Torres Diz[4], Jörg Faber[2], Yoseph Barash[3], Andrei Thomas-Tikhonenko[4,5], Kathi Zarnack[6*], Stefan Legewie[1,7*], and Julian König[1*]

Affiliations:
[1] Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz, Germany
[2] Department of Pediatric Hematology/Oncology, Center for Pediatric and Adolescent Medicine, University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz, Germany & University Cancer Center (UCT), University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz & German Cancer Consortium (DKTK), site Frankfurt/Mainz, Germany, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[3] Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, US and Department of Biochemistry and Biophysics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, US
[4] Division of Cancer Pathobiology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, US
[5] Department of Pathology & Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, US
[6] Buchmann Institute for Molecular Life Sciences (BMLS) and Faculty Biological Sciences, Goethe University Frankfurt, Max-von-Laue-Str. 15, 60438 Frankfurt, Germany
[7] Department of Systems Biology and Stuttgart Research Center for Systems Biology (SRCSB), University of Stuttgart, Stuttgart, Germany

# These authors contributed equally.
* Corresponding authors: Kathi Zarnack (kathi.zarnack@bmls.de), Stefan Legewie (legewie@iig.uni-stuttgart.de), Julian König (j.koenig@imb-mainz.de)

Keywords: *CD19*, B-ALL, CART-19 therapy, relapse, massively parallel reporter assay, *cis*-regulatory mutations, *trans*-acting regulators, cryptic splice sites, modelling, PTBP1

Abstract

Despite the great success of CART-19 (CD19-directed chimeric antigen receptor-armed autologous T-cells) immunotherapy to treat B-cell acute lymphoblastic leukaemia (B-ALL), many patients relapse due to loss of the cognate CD19 epitope. Since epitope loss can be caused by *CD19* exon 2 mis-splicing, we set out to learn the regulatory code that controls *CD19* alternative splicing. To this end, we combine a massively parallel reporter assay with mathematical modelling and determine the splicing effects of all mutations in the region comprising *CD19* exons 1-3. Thereupon, we quantitatively disentangle the effects of complex mutation combinations and identify 83 single point mutations that lead to the accumulation of known therapy-relevant isoforms (exon 2 skipping or intron 2 retention) and thus could predispose B-ALL patients to CART-19 therapy resistance. In addition, we report 31 single point mutations that give rise to previously unknown splice isoforms. These isoforms emerge from cryptic splice site activation and likely encode for non-functional CD19 proteins, thereby potentially playing a role in CART-19 therapy resistance as well. We further use our mutagenesis data to learn about *cis*-regulatory elements and *trans*-acting RNA-binding proteins (RBPs) that control *CD19* splicing. By knockdown experiments, we confirm splicing-regulatory roles for several RPBs (e.g., SRSF3, PTBP1) and show that loss of these factors leads to enhanced exon 2 skipping and intron 2 retention. Taken together, our dataset represents a comprehensive resource for potential prognostic factors predicting CART-19 therapy success.

Highlights

- Mutations in relapsed CART-19 patients lead to *CD19* mis-splicing
- High-throughput mutagenesis uncovers ~200 mutations with a potential role in CART-19 therapy resistance
- Many mutations generate non-functional CD19 proteins by activating cryptic splice sites
- RNA-binding proteins such as PTBP1 promote the therapy-relevant isoforms

Introduction

B-cell acute lymphoblastic leukaemia (B-ALL) is a hematologic malignancy which causes a significant number of childhood and adult cancer deaths. In CART-19 immunotherapy, chimeric antigen receptor-armed autologous T-cells (CARTs) are engineered to target the surface antigen CD19 on B-cells by linking the single-chain variable fragment (scFv) of an anti-CD19 antibody to the intracellular signalling domain of the T-cell receptor {Maude, 2014 #42}. Upon CD19 recognition, the chimeric antigen receptors activate the cytotoxic T-cells to attack the tumour cells. CART-19 therapy was recently approved for the treatment of paediatric B-ALL in the US and Europe. Unfortunately, up to 50% of children relapse under CART-19 therapy, and response rates are even worse in adults {Wudhikarn, 2021 #58}{Roberts, 2018 #59}. Several studies reported that in 40-60% of cases the cancerous B-cells get invisible to the CARTs due to loss of detectable CD19 epitope (CD19-negative) {Orlando, 2018 #10;Park, 2018 #36;Gardner, 2017 #37;Maude, 2018 #38}. This recurrently involves alternative splicing of the *CD19* pre-mRNA {Sotillo, 2015 #11;Shah, 2019 #35}{Asnani, 2020 #20}.

Splicing comprises the excision of introns and the joining of exons by the spliceosome to generate mature mRNAs. During alternative splicing, certain exons can be either included or excluded ('skipped'), thus leading to different transcript isoforms. The splicing outcome at each exon is controlled by a large set of *cis*-regulatory elements in the RNA sequence which are recognised by *trans*-acting RNA-binding proteins (RBPs) that guide the spliceosome activity. It is increasingly recognised that widespread alterations in splicing are a molecular hallmark of cancer and often contribute to therapeutic resistance (reviewed in {Bonnal, 2020 #24}). For instance, intron retention, i.e., the failure to remove certain introns, often disrupts the open reading frame with premature termination codons (PTCs) and thereby compromises the expression of the encoded proteins. Consistent with the widespread splicing changes, cancer-causing 'driver' mutations frequently occur in splice-regulatory *cis*-elements, and many splicing factors have oncogenic properties, being commonly mutated or dysregulated in cancer {Bonnal, 2020 #24;Dvinge, 2016 #44;El Marabti, 2021 #45}.

Multiple alternative splicing events in *CD19* mRNA have been described to interfere with CART-19 therapy {Asnani, 2020 #20;Bagashev, 2018 #12;Fischer, 2017 #13;Rabilloud, 2021 #25;Sotillo, 2015 #11;Zhao, 2021 #39}. Most prominently, skipping of exon 2 results in a truncated CD19 protein which is no longer presented on the cell surface and hence fails to trigger CART-19-mediated killing {Sotillo, 2015 #11;Bagashev, 2018 #12}. In addition, it was reported that relapsed patients showed retention of intron 2 which introduces a PTC, thereby disrupting CD19 expression {Asnani, 2020 #20}. Similarly, simultaneous skipping of exons 5 and 6 introduces a PTC {Sotillo, 2015 #11}. The splicing alterations can be caused by mutations within the *CD19* gene or by changes in the expression of *trans*-acting RBPs. For instance, it has been suggested that the known splicing regulator SRSF3 binds to *cis*-regulatory elements within *CD19* exon 2 to promote its inclusion {Sotillo, 2015 #11}. Of note, alternative *CD19* isoforms showing exon 2 skipping were observed to pre-exist in patients prior to CART-19 therapy {Fischer, 2017 #13;Rabilloud, 2021 #25}, suggesting that *CD19* splicing patterns may harbour prognostic information and could be modulated to re-establish sensitivity to CART-19 mediated killing. However, Orlando and co-workers suggested that alternative splicing changes in B-ALL patients are present in diagnostic samples and may not contribute meaningfully to CD19 epitope loss {Orlando, 2018 #10}. We therefore set out to investigate *CD19* alternative splicing and its molecular determinants in B-ALL in more detail.

High-throughput mutagenesis screens combined with next-generation sequencing provide comprehensive insights into the regulatory code of splicing {Braun, 2018 #14;Baeza-Centurion, 2019 #15;Baeza-Centurion, 2020 #16;Ke, 2018 #17}. The interpretation of such data is challenging, as the mutation effects often depend on other mutations and are typically most pronounced at intermediate exon inclusion levels {Braun, 2018 #14;Baeza-Centurion, 2019 #15;Glidden, 2021 #18}. We and others have shown by mathematical modelling that kinetic models account for the context-dependence of mutation effects on splice isoforms {Braun, 2018 #14;Baeza-Centurion, 2019 #15}. By these models, systems-level insights can be gained into complex *cis*-regulatory landscapes, effects of *trans*-acting RBPs and principles of splicing regulation {Braun, 2018 #14;Baeza-Centurion, 2019 #15;Enculescu, 2020 #19}.

In this manuscript, we combine B-ALL patient data with high-throughput mutagenesis, mathematical modelling and RBP knockdowns to comprehensively characterise *cis*-regulatory mutations and *trans*-acting RBPs controlling *CD19* exon 2 splicing. Unlike previous mutagenesis screens, we determine all intronic and exonic mutation effects in a 1.2 kb region and quantify the abundance of 100 alternative isoforms, including intron 2 retention and alternative 3'/5' splice site usage. Many of these isoforms encode for a non-functional CD19 protein and are therefore likely to impair CART-19 therapy. By *in silico* analyses and RBP knockdowns, we identify *trans*-regulators of *CD19* splicing that promote the production of the therapy-relevant isoforms. Taken together, our dataset is a comprehensive resource for prognostic markers of CART-19 therapy resistance and for a systems-level understanding of the splicing code.

Results

CART-19 patients show increased *CD19* intron 2 retention after relapse

To resolve the contribution of *CD19* splicing in CART-19 therapy, we re-analysed RNA-seq data from Orlando and co-workers {Orlando, 2018 #10}, in which B-ALL cells of 17 patients were sequenced at initial screening and after relapse. In contrast to the original study, we expanded the analyses to intron retention surrounding *CD19* exon 2. We found that the average frequency of retention of intron 2 across patients significantly increases from 63% before therapy to 82% after relapse (*P* value = 0.022, Wilcoxon signed-rank test; **Figure 1A, B**). The trend towards higher intron 2 retention is preserved in 7 out of 10 individual patients that were sequenced both before therapy and after relapse (**Figure 1B**). Since the resulting isoform does not encode the CD19 epitope, this suggests that increased intron 2 retention contributes to CART-19 therapy relapse as reported in a recent study {Asnani, 2020 #20}.

Somatic mutations in relapsed patients cause splicing alterations

The majority of relapsed patients in the Orlando study (12 out of 17) {Orlando, 2018 #10} harbour somatic mutations within the *CD19* gene, including frameshift insertions, deletions and single nucleotide missense variants. We selected nine mutations in exons 2 or 3 from eight patients for further analysis (**Table S1**). To test for effects on splicing, we constructed a minigene reporter that harbours *CD19* exon 1-3 including the two intervening introns 1 and 2 (**Figure 1C**). We confirmed that the minigene gives rise to the same transcript isoforms as the endogenous gene in the human B-ALL cell line NALM-6 (**Figure 1D, E**). When introducing the patient mutations into our minigene reporter, we found that six out of nine tested mutations lead to the production of alternative *CD19* isoforms linked to CART-19 therapy resistance (**Figure 1F, G**): The mutation from patient #2 induces exon 2 skipping, while mutations from patients #4 and #14.2 cause intron 2 retention. In addition, three mutations enhance the production of an additional isoform that uses an alternative 3' splice site in exon 2 (termed alt-exon2; mutations from patients #5, #14.1 and #15). The alternative splice junction in alt-exon2 introduces a frameshift causing a PTC and will hence abolish the production of a targetable CD19 epitope. We note that as reported by Orlando and co-workers {Orlando, 2018 #10}, most of the tested mutations also introduce frameshifts, making it difficult to discriminate between PTC-induced and splicing-mediated defects. For instance, the alternative 3' splice site of alt-exon2 that is used in response to the deletion in patient #5 in fact compensates the frameshift that is introduced by the deletion, i.e., it restores the frame (**Figure Sxx**). Thus, taking the splicing information into account changes the interpretation of what CD19 protein variants are likely present in this patient. Overall, these results suggest that *CD19* mutations in CART-19 relapse patients frequently trigger splicing changes that potentially interfere with the therapy success.

High-throughput screening of *CD19* exons 1-3 alternative splicing

To systematically study the effects of point mutations on *CD19* exons 1-3 splicing, we adopted our previously developed massively parallel splicing reporter assay {Braun, 2018 #14} (**Figure 2A**). To this end, we randomly introduced point mutations as well as short insertions and deletions into the *CD19* minigene reporter by error-prone PCR. This yielded a pool of 10,295 minigene variants, each with a different set of mutations and tagged with a unique 15-nt barcode sequence. As an internal control, 194 wild type (WT) minigenes with distinct barcodes were added. Mutations in all minigene variants were mapped using targeted long-read DNA sequencing (DNA-seq, PacBio SMRT-seq, **Figure S1A, B**) and validated for 30 minigene

clones via Sanger sequencing. The DNA-seq data shows that the minigene variants contain on average 9.7 mutations (**Figure S1C**). This allows for a comprehensive characterisation of the mutation landscape, as each position is on average mutated in 80 different minigene variants and 90% of the mutations are present in at least four distinct minigene variants (**Figure S1D, E**). To measure splicing outcomes, the minigene pool was transfected into NALM-6 cells and the resulting transcripts were quantified by targeted RNA sequencing (RNA-seq) using 350 nt + 250 nt paired-end reads (Illumina MiSeq, **Figure S1B, S2A**). We detected around 100 different splice isoforms (see below) which were unambiguously identified by paired-end sequencing. Two replicate experiments showed high correlation in the measured isoform frequencies (R between 0.91 and 0.98 for the different isoforms, **Figure S2B**). Based on the common barcode sequence, information from DNA and RNA sequencing could be combined, linking mutations at the DNA level to frequencies of RNA splice isoforms for a total of 10,295 minigenes in two replicate experiments (**Table S2**).

Therapy-relevant isoforms accumulate in response to numerous point mutations

To our surprise, the screen revealed a high complexity of *CD19* exon 1-3 splicing, with a total of 101 alternative isoforms occurring with a frequency of ≥5% of all transcripts in at least two minigene variants (**Table S3**). Out of these, the five major isoforms exceed 1% in WT minigenes, whereas the others, termed cryptic isoforms, only accumulate in mutated minigene variants (**Figure 2B**). In WT, the by far most abundant major isoform is exon 2 inclusion (termed ‚inclusion'), followed by intron 2 retention (termed ‚intron2-retention') and exon 2 skipping (termed ‚skipping'). Two additional major isoforms in WT originate from alternative 3' splice site usage within exon 2 (alt-exon2) and 3 (alt-exon3) (**Figure 2B, C**). Notably, alt-exon2 is the same isoform as observed upon patient mutations above. As expected, the measured frequencies for the major isoforms show little variance for the 194 unmutated WT minigenes (standard deviation < 6%, **Figure 2C**). In contrast, many mutated minigene variants show strong changes relative to WT, suggesting a large impact of specific mutations on splicing outcomes (**Figure 2C**). For instance, all minigenes with a mutation in the 3' splice site of exon 2 lose the inclusion isoform, accompanied by strong alterations in the remaining major isoforms (**Figure 2C**). Taken together, these observations support the accuracy of our screening results.

All major isoforms, except exon 2 inclusion, encode for a truncated CD19 receptor lacking a functional CART-19 epitope and could thus contribute to therapy resistance. Our unbiased screening approach extends the list of potentially therapy-relevant *CD19* mutations, since 1,721 out of 9,127 mutated minigenes show exon 2 skipping, intron 2 retention and/or alt-exon2 isoform frequencies of >25% (**Figure 2C**). However, since the minigene variants carry on average 9.7 point mutations, the observed splicing changes represent the combined effects of several mutations. To extract the impact of individual mutations, we adapted our previous mathematical modelling framework {Braun, 2018 #14} and implemented a multinomial logistic regression approach. Here, the splicing change in each minigene variant is described as the sum of the underlying point mutation effects (**Figure 3A**, see Methods). These single mutation effects are unknown and are determined by simultaneously fitting the model to all minigene measurements. Thereby, we were able to infer the individual effects of 4,255 point mutations on the five major isoforms (**Figure 3B, S3A**). We validated the reliability of this model in describing combined mutations using a 10-fold cross-validation approach, in which we left out 10% of all minigene variants from fitting and were able to accurately predict them after model fitting (Pearson correlation coefficients 0.65-0.95; **Figure 3C, S3B**). Furthermore, the model

performed well in predicting single mutation effects, as soon as a mutation occurred in three or more minigenes in the dataset (**Figure S4C**), which applied to 90% of all mutations (**Figure S1E**).

Out of 4,255 quantified single mutation effects, we find 193 splicing-effective mutations that significantly alter the frequency of at least one isoform in the two replicates beyond the 2.5 and 97.5% quantiles of the WT minigene distribution (**Figure 3A, Table S4, Data S1**). The strongest effects accumulate around the four main splice sites and throughout exon 2 and correspond to the core *cis*-regulatory elements, such as splice-site dinucleotides, branchpoint and polypyrimidine tract, as well as auxiliary elements (**Figure 3B**). Inspecting in more detail the 83 mutations that specifically impact on *CD19* exon 2 skipping, we find them to cluster within and around exon 2. In particular, 21% of all positions within exon 2 (55 out of 267 nt) harbour at least one splicing-effective mutation, suggesting that *CD19* exon 2 is densely packed with *cis*-regulatory elements controlling its inclusion. In addition, we observe smaller clusters of mutations within the introns and flanking constitutive exons which likely represent more distal *cis*-regulatory elements (**Figure 3B**). Similarly, we explored the 54 splicing-effective mutations that impact on intron 2 retention. As expected, strongest effects are observed at the splice sites of intron 2. In addition, we find clusters of mutations in intron 2 and exon 3 that might reflect important *cis*-regulatory elements. The effect of all mutations on the five major isoforms can be explored in **Data S1**.

In conclusion, our combined screening and modelling approach quantitatively describes alternative splicing of *CD19* exons 1-3 by predicting the effects of all individual point mutations and combinations thereof. 34 of the splicing-effective mutations overlap with single nucleotide variants (SNVs) that were previously reported from whole-genome or exome sequencing data (**Table S5**). These variants are generally rare in the population, but could become relevant under CART-19 therapy. Our screen thereby represents a comprehensive resource for the identification of mutations with clinical relevance in CART-19 therapy resistance.

Cryptic isoforms destroy the *CD19* ORF and are associated with recurrent mutations

Besides the five major isoforms, the *CD19* exons 1-3 can give rise to 96 cryptic isoforms which are rare (<1%) in WT, but accumulate upon certain mutations (**Figure 2B, Table S3**). The cryptic isoforms involve a total of 71 cryptic splice sites (**Figure 4A**). Of note, 33 of these cryptic isoforms make up more than 50% of total transcripts and are therefore dominant in certain minigene variants (**Figure 2C**). To assess whether these cryptic isoforms impact on CD19 epitope presentation, we analysed their coding potential and found that the vast majority of cryptic *CD19* isoforms (78 out of 96) show a frameshift and/or carry a PTC (**Figure 4B**). This will either lead to the production of truncated CD19 peptides that likely do not allow for presentation on the cell surface {Bagashev, 2018 #12} or will induce nonsense-mediated mRNA decay of the cryptic isoforms and will hence reduce *CD19* transcript and protein levels.

To derive a mechanistic understanding of cryptic isoform biogenesis, we analysed the underlying point mutations. To this end, we calculated a prevalence score which quantifies the degree of association between an isoform and a point mutation. This was done based on the measured isoform frequencies in the minigene library by multiplying: (i) the frequency of a mutation being present if the isoform level is high (>5%), and (ii) the frequency of the isoform level being high given that the mutation is present. A prevalence score of 1 indicates perfect correspondence between mutation and isoform, whereas a prevalence score of 0 is observed if they are unrelated. This score-based analysis showed that 36 cryptic isoforms are specifically associated with 31 specific point mutations (38 mutation-isoform pairs with

prevalence score > 0.25, **Figure S4A, Table S3**). The remaining 60 cryptic isoforms do not show a specific association, implying that they can either be generated by multiple redundant mutations, or that our screen lacks sufficient coverage to support a reliable association. To directly test the predicted associations, we introduced five mutations with a specific association to a cryptic isoform in our minigene reporter (C535G, chr16:28932405, prevalence score = 0.18; C806A, chr16:28932676, 0.68; A827T, chr16:28932697, 0.93; C864G, chr16:28932875, 1; G1005A, chr16:28932734, 0.89). Semi-quantitative RT-PCR confirmed that all five tested mutations lead to the appearance of the associated cryptic isoform (**Figure 4C, D**).

Altogether, our analysis provides a list of 31 mutations that are likely to trigger cryptic isoform formation. Importantly, the resulting cryptic isoforms show a maximum usage of up to 91% (**Table S3**), which is likely to drastically interfere with normal *CD19* splicing, protein production and subsequent epitope presentation. The associated mutations may thus provide predictive biomarkers for CART-19 therapy response in the future.

The cryptic isoforms are caused by mutations that disrupt or create splice sites

Due to their potential clinical relevance, we wanted to learn more about how the mutations activate the cryptic isoforms. We found that the majority of mutations with a prevalence score > 0.25 are either in close proximity or directly overlap with the associated cryptic splice site (78.9% with distance < 5 nt; **Figure 4E**). Further inspection showed that the underlying mutations either destroy the original splice site (7.9%) or generate a new cryptic splice site (57.9%). Hence, the cryptic isoforms do originate from the generation or destruction of core *cis*-regulatory elements rather than affecting auxiliary elements.

Currently, major efforts are ongoing to implement artificial intelligence (AI) tools to predict the effect of clinical variants on the splicing outcome. We therefore tested whether the state-of-the-art neural network SpliceAI {Jaganathan, 2019 #6}, which predicts changes in the splicing patterns induced by single point mutations, captures the gain and loss of splice sites in *CD19*. To this end, we applied SpliceAI using all possible single point mutations in the *CD19* minigene as an input. Similar to the results from our mutagenesis screen (**Figure 4A**), SpliceAI predicts cryptic splice site activation by mutations throughout the minigene, with an increased density around the 3' splice site of exon 3 (**Figure S4B**). All SpliceAI-predicted mutations are close to the affected cryptic splice sites (**Figure S4C**). With respect to the accuracy of the predictions, we found that almost half of the mutations with strong SpliceAI predictions indeed lead to the activation of cryptic splice sites (23 out of 48 with SpliceAI score > 0.5, **Figure 4F**). Generally, it appears that SpliceAI does not miss any major effects, but possibly slightly over-predicts, since 52.1% of mutations with a SpliceAI score > 0.5 do not match the cryptic sites induced by mutations detected in the experimental data (either showing no effect at all or activating a different cryptic splice site than predicted). Overall, the comparison supports that SpliceAI can guide the interpretation of mutation effects in clinical samples. In addition, our data can be used to benchmark new tools for splicing prediction.

The cryptic isoforms arise from numerous 3' and 5' cryptic splice sites that distribute over the entire minigene and accumulate at exon 3 (**Figure 4A**). In line with a high penetrance, 26 cryptic splice sites reach more than 50% usage upon certain mutations, particularly around the start of exon 3. We hypothesised that cryptic splice site activation occurs in exon 3 because its canonical splice site can be outcompeted by neighbouring cryptic sites. To test this, we scored the strength of local consensus sequences using MaxEntScan {Yeo, 2004 #5}, and indeed found that the 3' splice site of exon 3 is weak compared to all other canonical splice

sites of *CD19* exons 1-3 (**Figure 4G, S4D, E**). In line with our hypothesis, mutations around the 3' splice site of exon 3 frequently create stronger splice sites than elsewhere in the minigene that exceed the strength of the canonical 3' splice site of exon 3 (**Figure 4G**). This suggests that weak splice sites are particularly vulnerable for the activation of competing cryptic splice sites and should be of particular interest when assessing the impact of clinical variants on splicing outcomes.

An extensive network of RBP regulators may drive *CD19* mis-splicing

Besides *CD19* mutations, CART-19 therapy resistance may also originate from an altered expression of *trans*-acting RBPs which bind to the *CD19* pre-mRNA to control alternative splicing. To identify putative RBP regulators, we explored publicly available databases containing experimentally determined RBP binding motifs (ATtRACT {Giudice, 2016 #7}, oRNAment {Benoit Bouvrette, 2020 #8}). Furthermore, we employed DeepRiPe {Ghanbari, 2020 #9}, a neural network-based algorithm trained on PAR-CLIP and ENCODE eCLIP datasets that predicts changes of RBP binding upon mutation. In combination, these tools predict a total of 198 RBPs to bind within *CD19* exons 1-3 (ATtRACT: 62 RBPs; oRNAment: 70 RBPs) or to change binding upon mutation (DeepRiPe: 128 RBPs) (**Figure 5A, S5**).

To link the putative RBP regulators to the observed splicing changes, we overlaid the predicted binding sites (or predicted mutations for DeepRiPe) with splicing-effective mutations from our screen. Overall, we find that 79% and 60% of ATtRACT and oRNAment binding sites, respectively, overlap with a splicing-effective mutation (affecting any of the five major isoforms). Furthermore, 48% of mutations predicted to change RBP binding by DeepRiPe overlap with splicing-effective mutations, suggesting that modulating RBP binding at these sites may have a functional impact on *CD19* splicing (**Figure 5A, S5A**). By merging these sets, we obtained a list of 119 RBPs that may regulate splicing by binding to *CD19* exons 1-3 (**Table S5**). Most of these are stably expressed in cancerous B-cells from B-ALL patients from {Gu, 2019 #46} (80 with mean FPKM [fragments per kilobase of transcript per million mapped reads] > 10; **Figure S5B**) and could thus interfere with CART-19 therapy. Among these RBPs are SRSF3, a previously reported regulator of *CD19* splicing {Sotillo, 2015 #11}, but also new candidates such as PTBP1. Altogether, the *in silico* predictions suggest the presence of an extensive RBP network controlling *CD19* splicing that may impact on the CART-19 therapy success.

Depletion of PTBP1 and several other RBPs results in non-functional *CD19* isoforms

Based on our experimental data, *in silico* predictions, expression and literature information, we shortlisted 11 RBP candidates for further analysis, including *SRSF3* as a positive control. To test their impact on endogenous *CD19* splicing, we generated NALM-6 cell lines stably expressing shRNAs against the shortlisted RBPs (depletion to <40% transcripts; **Figure S6A**). As previously described {Sotillo, 2015 #11}, knockdown of *SRSF3* leads to increased exon 2 skipping in the endogenous *CD19* transcripts, confirming that this SR protein is required for exon 2 inclusion (**Figure 5E, F**). Importantly, we find that knockdown of six additional RBPs (PTBP1, PCBP2, SF3B4, HNRNPK, MBNL1 and HNRNPM*)* has significant effects on *CD19* alternative splicing (**Figure 5E, F, S6B, C**). The knockdown of these factors reduces *CD19* exon 2 inclusion, while promoting intron 2 retention and/or exon 2 skipping, thus shifting the cells towards expression of relapse-associated *CD19* isoforms. This implies that reduced levels of these factors can impair targetable CD19 epitope expression.

PTBP1 stands out among the putative regulators as it shows the strongest effects on intron 2 retention, which emerged as the most prominent *CD19* mis-splicing isoform in our reanalysis of B-ALL patient data (**Figure 1A, B**). PTBP1 recognises clusters of UC-rich motifs {Spellman, 2006 #66}{Haberman, 2017 #67}. Remarkably, ATtRACT predicts almost 100 such PTBP1 binding motifs across the studied *CD19* region, including 25 that overlap with splicing-effective mutations (**Figure 5D, Table S5**). Moreover, DeepRiPe predicts 78 mutations in 63 positions that change PTBP1 binding, out of which 10 are splicing-effective in our screen. The high number of predicted binding sites suggests a partial redundancy, indicating that PTBP1 regulation might be difficult to disrupt with individual point mutations as introduced in our screen. To experimentally test if PTBP1 binds to the predicted sites, we performed PTBP1 iCLIP2 experiments in NALM-6 cells. In line with a role in intron 2 retention, we find extensive PTBP1 binding particularly in intron 2, where it spreads over an extended cluster of predicted binding sites (**Figure 5G**). The broad binding at splicing-effective positions and beyond supports that PTBP1 is a direct and central regulator of *CD19* alternative splicing.

Discussion

Massively parallel reporter assays such as our high-throughput mutagenesis screen provide comprehensive insights into the regulatory code of splicing, as they characterise the complete set of *cis*-acting sequence mutations and reveal the binding sites of *trans*-acting RNA-binding proteins (e.g., {Baeza-Centurion, 2019 #15;Baeza-Centurion, 2020 #16;Braun, 2018 #14;Mikl, 2019 #47}{Cheng, 2019 #48}). The interpretation of these datasets is challenging due to nonlinear interactions of individual mutation effects. For instance, competition effects in splicing reduce the impact of individual mutations at low and high isoform frequencies, i.e., depending on the mutational background {Braun, 2018 #14;Baeza-Centurion, 2019 #15}. Using kinetic modelling, we and others derived regression models taking this competition into account, thereby showing that splicing effects of complex mutation combinations can be quantitatively described as the sum of individual mutation effects {Braun, 2018 #14;Baeza-Centurion, 2019 #15}. Thus, mutations seem to control splicing additively rather than synergistically, and this principle also holds for *CD19* splicing.

In our *CD19* mutagenesis dataset, we comprehensively characterised the full set of splice isoforms generated in response to thousands of sequence mutations. In particular, we find that cryptic splice site activation and thus alternative 3' and 5' splice site usage are common modes of alternative splicing. Intriguingly, such events do not require extensive sequence remodelling, but can often be triggered by single point mutations, as indicated by strong associations between cryptic isoforms and certain nucleotide substitutions. This suggests, in accordance with previous reports {Yu, 2008 #60}, that neighbouring splice sites frequently compete for spliceosome assembly, especially if the canonical splice site is comparably weak (**Figure 4G**). While this finding shows the enormous isoform complexity that can arise already from such a simple exon configuration, it raises the question of how protein function can be robustly maintained, since most cryptic *CD19* splicing isoforms likely encode non-functional proteins.

Unlike previous mutagenesis screens, which mainly focused on exonic sequence mutations, the present *CD19* dataset characterises the complete set of intronic and exonic mutations in a 1,200 nt sequence stretch. The complete characterisation of *CD19* exons 1-3 required the use of long-read sequencing technology. Given that introns in the human genome on average span 1,700 nt (refxx), the long-read sequencing methodology described in this work opens the approach for broad applications. For *CD19*, we find that strong mutation effects are mainly centred around canonical and cryptic splice sites, whereas mutation effects seem to be dispersed for highly regulated exons such as *MSTR1* exon 11 {Braun, 2018 #14}. This suggests that (near-)constitutive exons like *CD19* exon 2 may exhibit stronger and redundant splicing enhancers and that their inclusion is therefore less sensitive to individual point mutations {Baeza-Centurion, 2019 #15}. More generally, constitutive exons may require more specific perturbations and as we show here, do not respond with only exon skipping, but tend to employ alternative splice site usage and intron retention, both of which are clinically relevant in the case of *CD19* splicing and CART-19 therapy resistance.

Our retrospective analyses of clinical B-ALL samples implicate both *CD19* mutations and *CD19* splice isoforms in the development of CART-19 therapy resistance. Using minigene assays, we directly show that *CD19* mutations that are observed in relapsed patients lead to exon 2 skipping, intron 2 retention or an additional isoform that uses an alternative 3' splice site in exon 2. Furthermore, based on our mutational scan, we report 83 additional point mutations significantly affecting these therapy-relevant isoforms. Taken together, our results

strongly suggest that *CD19* mutations contribute to CART-19 therapy resistance by inducing splicing changes and likely do so by changing RBP binding sites in the *CD19* pre-mRNA. The detection of such mutations in longitudinal samples may provide predictive biomarkers for therapy response in the future.

At the same time, alterations in the expression of *trans*-acting RBPs can induce aberrant *CD19* splicing, explaining the presence of CD19-negative relapses in samples with a low allelic frequency of mutations or without mutations in the *CD19* locus. Mutations in splicing factors such as SRSF2, SF3B1 and U2AF1 are common in myelodysplastic syndrome/acute myelogenous leukaemia {Yoshida, 2011 #21} and chronic lymphocytic leukaemia {Quesada, 2011 #22}, and are associated with aberrant splicing. In B-ALL, mutations in splicing factors are not common, but previous work suggests that several splicing factors are deregulated {Black, 2018 #23}. In the context of *CD19*, we confirm that SRSF3 deregulation induces exon 2 skipping {Sotillo, 2015 #11} and identify several other RBPs that promote CD19 protein isoforms invisible to the immunotherapeutic agent, including PTBP1, PCBP2, SF3B4, HNRNPK, MBNL1 and HNRNPM. Several of the newly identified regulators have been found as deregulated in other cancer types and discussed as potential targets for anti-cancer therapy {Desterro, 2020 #65} (xxref). Moreover, an upregulation of PTBP1 has been implicated in the acquired resistance of pancreatic ductal carcinoma cells to the chemotherapeutic drug gemcitabine {Calabretta, 2016 #61}. In the context of lymphocytes, PTBP1 is upregulated in B cells and required for early B cell selection {Monzón-Casanova, 2020 #62}. It was reported, however, that treatment of leukemic cells with the chemotherapeutic drug imatinib lowers PTBP1 levels {Shinohara, 2016 #63}. In the light of our finding that *PTBP1* knockdown increases *CD19* intron 2 retention and thereby most likely reduces CD19 epitope exposure, a previous treatment with imatinib may have a negative impact on the subsequent CART-19 therapy response. In addition, a recent study showed that the repeat RNA *PNCTR* sequesters substantial amounts of nuclear PTBP1 in various cancers {Yap, 2018 #64}. Thus, other factors like cellular availability may further impact on PTBP1 function in B-ALL cells under CART-19 therapy.

Currently, we cannot predict which patients with a CD19-positive B-ALL have a high risk of developing a CD19-negative relapsed disease. The pre-existence of isoforms skipping exon 2 or exons 5-6 has been previously discussed as a possible biomarker {Fischer, 2017 #13;Rabilloud, 2021 #25}. Our results indicate the necessity to extend the analysis to more isoforms and possibly to include the expression of splicing factors in screening approaches to identify patients at risk to relapse under CART-19 therapy. Notably, the same biomarkers might also be relevant for other malignancies arising from B-cell lineage, such as large B-cell lymphoma. Here, sequential loss of CD19 following CART-19 therapy has been described as a mechanism for relapse following immunotherapy {Shalabi, 2018 #27}, accounting for 29% of relapses in recent clinical studies {Spiegel, 2021 #28}. Our data show that *CD19* splicing is highly complex, with already ~100 alternative isoforms concerning just exons 1-3. Of them, about 80% encode for a CD19 receptor lacking a functional CART-19 epitope and are thus expected to contribute to therapy resistance. The specific detection of alternative splicing might serve as a reliable biomarker and may provide a novel approach to monitor disease progression as already suggested in other tumour entities {Venables, 2008 #26}.

The contribution of aberrant splicing to CART-19 resistance may further be relevant for future combination therapies. Small-molecule splicing modulators are currently in clinical trials for myeloid neoplasms and splice site-switching antisense oligonucleotides are in development

for different targets (reviewed in {Bonnal, 2020 #24}). Our mutagenesis dataset provides a strong basis for designing and systematically evaluating splice-switching oligonucleotides for the modulation of *CD19* splicing. The combined application of these splicing modulators with immunotherapy may represent a way to limit the generation of resistance to CART therapies.

Methods

Cell lines

NALM-6 cells were obtained from ATCC and cultured in RPMI medium (Life Technologies) with 10% foetal bovine serum (Life Technologies) and 1% l-glutamine (Life Technologies). HEK293T cells were obtained from DSMZ and grown with the same additives as for NALM-6. All cells were kept at 37 °C in a humidified incubator containing 5% CO_2. They were routinely tested for mycoplasma infection.

Cloning

The *CD19* minigene was amplified from human genomic DNA (Promega) with the primers 5'-catAAGCTTgaccaccgccttcctctctg-3' and 5'-catGAATTCNNNNNNNNNNNNNNNGGATCCttcccggcatctccccagtc-3'. pcDNA3.1 was used as the vector backbone for the *CD19* minigene plasmid. Both the backbone as well as the minigene amplicons were digested with the restriction enzymes *EcoR*I and *Hind*III (New England Biolabs). The backbone was extracted from a 1% agarose gel using QIAquick Gel Extraction Kit (Qiagen) and the minigene insert was cleaned up using QIAquick PCR Purification Kit (Qiagen). Ligation was conducted overnight at 16 °C with T4 DNA Ligase (New England Biolabs). All minigene mutations were introduced via Q5 Site-Directed Mutagenesis Kit (New England Biolabs). The nine mutations from eight patients in Orlando et al. {Orlando, 2018 #10} are listed in **Table S1**. All kits were used according to the manufacturers' recommendations.

Mutagenesis of minigene and library construction

For the random mutagenesis of the *CD19* minigene, GeneMorph II Random Mutagenesis Kit (Agilent) was used according to manufacturer's recommendations using 500 ng *CD19* minigene for 30 cycles at 56 °C with the amplification primers 5'-catAAGCTTgaccaccgccttcctctctg-3' and 5'-catGAATTCNNNNNNNNNNNNNNNGGATCCttcccggcatctccccagtc-3'. PCR products were purified using QIAquick Gel Extraction Kit (Qiagen), digested with *EcoR*I and *Hind*III (New England Biolabs) and then ligated into the backbone. To raise the baseline level of exon 2 inclusion in the *CD19* minigene to a similar level as in the endogenous *CD19* gene, position 748 (nucleotide 6 of intron 2) was exchanged from G to T.

Transfection of minigene

Cells were twice washed in Dulbecco's phosphate buffered saline (DPBS, Gibco Thermo Fisher Scientific) and then collected in R buffer with a density of 2×10^7 cells/ml. For electroporation, we used 5 µg plasmid DNA (with a concentration of at least 1 µg/µl) to 2×10^6 cells in R buffer for a 100 µl NEON electroporation pipette tip (Thermo Fisher Scientific) at 1600 V for 30 ms and 1 pulse. Cells were harvested 24 h later.

Quantification of splicing isoforms using semi-quantitative RT-PCR

Semi-quantitative RT-PCR was used to quantify ratios of *CD19* mRNA isoform variants. To this end, reverse transcription was performed on 500 ng RNA with RevertAid Reverse Transcriptase (Thermo Fisher Scientific) according to the manufacturer's recommendations. Subsequently, 1 µl of the cDNA was used as template for the RT-PCR reaction with OneTaq DNA Polymerase (New England Biolabs). PCRs were run at the following conditions: 94 °C for 30 s, 28 cycles (minigene) or 34 cycles (endogenous *CD19*)

of [94 °C for 20 s, 55 °C for 30 s, 68 °C for 30 s] and final extension at 68 °C for 5 min. The primers 5'-ACCTCCTCGCCTCCTCTTCTTC-3' and 5'-GCAACTAGAAGGCACAGTCG-3' were used for the *CD19* minigene, and 5'-ACCTCCTCGCCTCCTCTTCTTC-3' and 5'-CCGAAACATTCCACCGGAACAGC-3' for the endogenous *CD19* gene. The TapeStation 2200 capillary gel electrophoresis instrument (Agilent) was used for quantification of the PCR products on D1000 tapes.

Generation of stable and inducible shRNA knockdown cell lines

Production and preparation of lentivirus

Oligonucleotides with shRNA inserts against eleven RBPs (**Table S7**) were ordered as Ultramer DNA Oligos from Integrated DNA Technologies (Leuven, Belgium). All sequences were based on {Fellmann, 2011 #29}. Oligonucleotides containing shRNA inserts were PCR-amplified with primers 5'-TCTCGAATTCTAGCCCCTTGAAGTCCGAGGCAGTAGGC-3' and 5'-TGAACTCGAGAAGGTATATTGCTGTTGACAGTGAGCG-3' and purified with QIAquick PCR Purification Kit (Qiagen). shRNA inserts and miRE18_LT3GEPIR_Ren714 backbone (inducible via Tet-On system) were cut with *EcoR*I and *Xho*I (New England Biolabs). Backbone was purified from agarose gel with QIAquick Gel Extraction Kit (Qiagen). The fragments were then ligated with T4 DNA Ligase (New England Biolabs) at 16 °C overnight.

Constructs were transduced into NALM-6 via HEK293T-produced lentiviruses. To this end, 10 cm dishes of HEK293T were transfected using 30 µl Lipofectamine 2000 (Thermo Fisher Scientific) with three plasmids: 4 µg shRNA-producing constructs + 2 µg psPAX2 (lentiviral packaging) + 1 µg pMD2.G (lentiviral envelope) at 72 h prior to transduction. On the first day after transfection, the medium was changed. Work with cells used for lentiviral production was conducted in the S2 laboratory.

Transduction of NALM-6 cells

Lentiviral production was confirmed with Lenti-X GoStix (Takara) and lentiviruses were concentrated with Lenti-X Concentrator (Takara) according to the manufacturer's recommendations. For transduction, 1 x 10^6 NALM-6 cells in 500 µl of medium were added to the concentrated virus. 5 µg/ml polybrene (Sigma-Aldrich) was added. The cells were centrifuged at 800 g and 32 °C for 30 min. Cells were then transferred into 6-well plates and cultivated in normal growth medium without antibiotics. Selection was started after 48 h with 0.5 µg/ml puromycin (Thermo Fisher Scientific). Antibiotic medium was exchanged every 2 to 3 days. As soon as cells were not dying under selection anymore and the population was stable, induction experiments were started. After transduction, cells remained in the S2 laboratory for at least 6 weeks. Then, Lenti-X GoStix was used to check for any remaining lentivirus.

Induction of stable shRNA-expressing NALM-6 cells

Controlled by the Tet-responsive *TRE3G* promoter, the expression of shRNA was induced by addition of doxycycline (Thermo Fisher Scientific). To this end, 2 x 10^6 NALM-6 cells were seeded into a 6-well plate in 2 ml medium containing 0.5 µg/ml puromycin and induced with 0.5 µg/ml doxycycline, diluted in RPMI 1640 medium (Thermo Fisher Scientific). Induction was conducted at 37 °C and 5% CO_2 and cells were harvested after 48 h. During induction, the shRNA expression system is coupled to the production of eGFP, which was examined by fluorescence microscopy before harvesting.

Quantitative real-time PCR (qPCR)

RNA was extracted from the induced harvested cells using the RNeasy Plus Mini Kit (Qiagen). This RNA was used for qPCR to validate the RBP knockdown as well as for semi-quantitative RT-PCR experiments to check the splicing pattern of endogenous *CD19*. The qPCR was conducted using the Luminaris HiGreen qPCR Master Mix, low ROX (Thermo Fisher Scientific) according to the manufacturer's recommendations. Oligonucleotide sequences of all qPCR primers are given in **Table S6**.

Targeted DNA sequencing

DNA-seq of the minigene library was performed on the PacBio SMRT sequencing platform at MPI-CBG Dresden. For this purpose, the minigene plasmid library was digested with *EcoR*I and *Hind*III (New England Biolabs) and run on an agarose gel. The desired band at the size of 1,301 nt was cut out and purified using QIAquick Gel Extraction Kit (Qiagen). For the run on the PacBio SMRT cell, a standard library preparation was performed.

Targeted RNA sequencing

NALM-6 cells were electroporated with the mutated minigene library (see above). 24 h later cells were harvested and RNA was isolated via the RNeasy Mini Kit (Qiagen). 20 µg isolated RNA was poly-A-selected using Dynabeads Oligo $(dT)_{25}$ beads (Invitrogen) according to the manufacturer's recommendations. Reverse transcription was performed on 500 ng poly-A-selected RNA with RevertAid Reverse Transcriptase (Thermo Fisher Scientific) according to the manufacturer's recommendations. To prevent chimeric amplicons, the RNA-seq libraries were amplified via emulsion PCR {Williams, 2006 #30} using the Phusion DNA Polymerase (New England Biolabs). The following primers containing Illumina adapters were used in the PCR: 5'-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGAACCTCTAGTGGTGAAGG-3' (fwd) 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNCCGCCAGTGTGATGGATATC-3' (rev) under following conditions: 98 °C for 30 s, 25 cycles of [98 °C for 10 s, 63 °C for 20 s, 72 °C for 1 min] and final extension at 72 °C for 5 min. Amplicons were purified using Agencourt AMPure XP beads (Backman Coulter). Purified products were analysed on the TapeStation 2200 capillary gel electrophoresis instrument (Agilent) and quantified using the Qubit assay (Thermo Fisher Scientific). RNA-seq was carried out on the Illumina MiSeq platform using paired-end reads of 350 nt + 250 nt length and a 10% PhiX spike-in to increase sequence complexity.

Re-analysis of RNA-seq data from Orlando et al.

We re-analysed RNA-seq data of B-ALL patients at screening and after CART-19 therapy relapse from Orlando et al. {Orlando, 2018 #10} to quantify intron 2 retention in *CD19*. Since raw data were not available, we obtained BAM files for the different patients deposited in the Short Read Archive (SRA) under the accession SRP141691. For 10 patients, matched data were available at screening and relapse. The data contained the aligned reads mapped to several genes from the immune system including *CD19*. Using custom scripts, we extracted the sequence of the reads, reformatted them and generated fastq files. We then mapped the fastq files to our minigene sequence using STAR (v2.6.1) {Dobin, 2013 #41}. We used the re-mapped reads to quantify the levels of intron 2 retention in the different samples using the R/Bioconductor package ASpli {Estefania, 2021 #31}.

DNA-seq barcode demultiplexing

We obtained the circular consensus sequences (CCS), stored as fastq files. Two rounds of sequencing yielded a total of 337,215 CCS. We kept only reads with a length of 150-1,150 nt. We adapted the demultiplexing procedure described in {Braun, 2018 #14}. In this case, we searched for the 15-nt barcode in the last 50 nt of the read. If the barcode was not found, we searched in the last 50 nt of the reverse complementary strand. We only allowed the recovery of barcodes ranging from 14 to 16 nt, which would account for barcodes containing one nucleotide inserted or deleted. Before proceeding with the variant calling, we determined a cutoff to decide the minimal number of CCS to call variants on. Here, we kept only barcodes supported by at least 4 CCS. In total, we recovered 68.5% of all the demultiplexed barcodes which corresponded to 10,558 different minigenes, closely resembling the ~10,000 minigene clones that were used to generate the library.

DNA-seq mapping and variant calling

We use BLASR {Chaisson, 2012 #1} with the standard parameters to map the de-multiplexed minigene sequences to the minigene reference. We performed variant calling in the aligned BAM files using the GATK {McKenna, 2010 #2} HaplotypeCaller (v4.0.10) with the parameters *--kmer-size 10 --kmer-size 15 --kmer-size 25 --allow-non-unique-kmers-in-ref*. We used different k-mer sizes to improve the detection of problematic regions. Mixed barcodes, i.e., barcodes containing two classes of mutations, were removed based on the "penetrance score", reported as allele frequency (AF) in the GATK vcf output files, such that barcodes with more than 25% variants of low penetrance (AF < 0.8) were discarded. Using this strategy, we were able to recover 100,135 mutations of high quality coming from 10,295 distinct minigenes plus an additional 194 unmutated WT minigenes with distinct barcodes. 57.4% of the mutations appeared in at least ten different minigenes.

RNA-seq barcode demultiplexing

RNA-seq libraries were sequenced on Illumina MiSeq as 350 nt + 250 nt paired-end reads, yielding approximately 23 million reads. We controlled their quality using FastQC (v0.11.5, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and removed bad quality ends of reads using Trimmomatic {Bolger, 2014 #40} (v0.36, parameters: SLIDINGWINDOW:6:10 MINLEN:0). After trimming, we filtered for read pairs with a minimal length of 305 nt (read1) and 157 nt (read2) and, as done in Braun et al. {Braun, 2018 #14}, we used matchLRPatterns() and trimLRPatterns() from the R/Bioconductor package 'Biostrings' to extract the 15 nt barcode in read1 between the two flanking restriction sites (Lpattern = TGCAGAATTC, Rpattern = GGATCC) allowing one mismatch. All read pairs with barcode length between 14 and 16 nt were kept for further processing. Barcode sequences were added to the read names in the fastq file and 5' ends of reads were trimming (read1: everything until the second anchor sequence GGATCC, read2: the first 12 nt). After identifying and trimming the barcode and other regions, we used Cutadapt {Martin, 2011 #3} (v1.6, parameters: --adapter=TAGAGGTTCC --overlap=3 --error-rate=0.1 --no-indels --minimum-length=244 --pair-filter=both) to remove remaining primer sequences from read1. Lastly, the barcode information attached to the read names was used to demultiplex all read pairs into individual fastq files for each minigene.

Isoform quantification from RNA-seq data

Only barcodes/minigenes also detected in the DNA-seq library were kept for further analysis. All minigenes with insertions or deletions of 10 or more base pairs were removed from further analysis. For better mapping results, we shortened read1 to at most 260 nt. Read pairs of

each minigene were mapped to the respective minigene (including all mutations, but excluding insertions and deletions) using STAR {Dobin, 2013 #41} (v2.6.1b). An annotation of three isoforms (exon 2 inclusion and skipping, as well as the artefact PCR product Δex2part which lacks an internal fragment of exon 2 due to a reverse transcription artefact {Schulz, 2021 #49}) was provided to STAR during mapping and an --sjdbOverhang of 259 was set. When running STAR, all SAM attributes were written, up to ten mismatches were allowed, soft-clipping was prohibited on both ends of the reads and only uniquely mapping reads were kept for further analysis. BAM files were sorted and indexed using SAMtools {Li, 2009 #50} (v1.5).

Properly and consistently mapped pairs were used for isoform reconstruction using a custom Perl script. Read pairs were considered properly mapped if they mapped with the right orientation on opposite strands. Read pairs mapped consistently if they either did not overlap or in case of an overlap, agreed in their predicted splice junctions. Besides, only read pairs for which both mates exceeded the constitutive exon boundaries by at least 10 nt were used for isoform reconstruction. All other pairs were removed since they did not provide any isoform information. Only minigenes covered by at least 100 read pairs usable for isoform reconstruction were kept for further analysis. For each read pair, the CIGAR strings of the two mates were used to reconstruct their splicing isoform. Regarding the artefact product Δex2part, we combined the eight possible mappings of the missing internal fragment of exon 2 which are possible due the associated 8-nt repeat sequence {Schulz, 2021 #49}. Only isoforms, which were supported by >=1% of the read pairs and at least two read pairs in at least one minigene, were kept for further analysis.

The analysis described above was done separately for two replicates. All isoforms occurring with a frequency of at least 5% in two or more minigene variants in either of the two replicates were kept as individual isoforms. All other detected isoforms were summarised into a category ‘discarded’. Isoforms with Δex2part, i.e., excluding the internal intron in exon 2, were combined with their ‘real’ counterparts without Δex2part by merging isoforms that only differed in the exclusion of the internal fragment of exon 2. In total, this leads to a set of 101 individual isoforms.

Estimation of single mutation effects and splicing-effective mutations

Since the majority of the minigenes in the dataset exhibit more than one mutation, with a mean of 9.6 mutations per minigene, the splicing-effective mutations cannot be read out directly from the data. We used multinomial logistic regression to infer the effects of single mutations from combined measurements. The regression is based on hypothetical minigenes containing only one mutation, and on the assumption that mutation effects (log fold-changes compared to WT) add up into combined ones at the levels splice isoform ratios {Braun, 2018 #14}.

For regression, we focused on the five major isoforms that are already present in the WT minigene (see main text). Therefore, minigenes exhibiting more than 5% cryptic isoforms were removed from the dataset, and for the remaining minigenes the cryptic isoforms were merged into a lumped splicing category which we termed ‘other’. Thus, six categorical splicing outputs (inclusion, skipping, intron2-retention, alt-exon2, alt-exon3, other) were considered in the regression model, and the probability of each these outputs to be observed was assumed to equal the measured isoform frequencies. The regression was formulated as a softmax regression problem using the LogisticRegression command from the Python package scikit-learn {Pedregosa, 2011 #51}.

Given the large number of mutations per minigene in the dataset, the regression was prone to overfitting (i.e., mutations with weak effects on splicing were assigned non-zero coefficients to fit random fluctuations in the data; not shown). To avoid this problem, we employed L1 penalisation. The strength of the penalty was optimised by tenfold cross-validation, and the resulting inverse regularisation strength was C=10 for both replicates.

The goodness of the model in describing the measured combined mutation effects (minigenes) was tested by assessing the correlation between model and data in training and test datasets (**Figure S3A**). Tenfold cross-validation at the final penalisation strength showed that the method performs very well in estimating the minigene isoform frequencies of the test dataset (**Figure S3B**). In the cross-validation, the Pearson correlation coefficients between softmax predictions of combined mutation effects and measurements lie for the single isoforms between 0.68-0.95 for the first replicate and between 0.71-0.93 for the second replicate (**Figure 3C**).

The accuracy of the model-predicted single mutation effects in the softmax regression was assessed by leaving out 56 directly measured single mutation minigenes (i.e., minigenes bearing only one mutation) from the training data. Since most of these 56 mutations are not splicing-effective, we focused our analysis on the seven mutations that change the inclusion isoform level beyond two standard deviations of the WT minigene distribution: For each of the seven mutations, we performed multiple softmax fits in which the training data: (i) contained all minigenes not harbouring the mutation of interest, (ii) excluded its single mutation minigenes, and (iii) comprised varying numbers of combined mutation minigenes containing the mutation. For each mutation occurrence between 1 and 10, we used up to 7 different, randomly chosen combinations of multiple mutated minigenes including the mutation of interest. For each of these models, we generated predictions for the single mutation effect. The prediction accuracy was assessed by calculating the difference between model and direct single mutation measurements for a certain mutation occurrence. The standard deviation of the difference between model and data was used as a measure for the model error. We find that a mutation occurrence of 3 leads to an error level equal to two WT standard deviations (calculated based on inclusion levels of all WT minigenes in the first replicate). For higher mutation occurrences, the prediction accuracy does not improve further (**Figure S3C**).

The final modelling step was to identify splicing-effective mutations. For this purpose, we adopted an approach analogous to empirical *P* values, i.e., we compared predicted single mutation effects to empirical isoform frequency distributions in the WT. Isoform frequencies were measured for 195 and 194 WT minigenes in the two replicates. For each isoform and replicate, we chose the 2.5% and 97.5% quantiles of the respective empirical WT frequency distribution as cutoffs (corresponding to a two-sided 5% cutoff). A mutation was considered to have an effect on a splice isoform if, for both replicates, the frequencies predicted by the model were beyond the respective cutoffs and if the effects were in the same direction.

Splice site characterisation

Splice site usage for a given position represents the frequency of the isoforms using a given splice site in a particular minigene divided by the sum of all isoform frequencies for the same minigene. For **Figure 4A**, we used the maximum usage of a particular splice site across all minigenes. The strength of putative splice sites along the minigene was calculated using MaxEnt scores {Yeo, 2004 #5} in sliding windows of 9 nt or 23 nt to evaluate the corresponding sequences as potential 5' or 3' splice sites, respectively. The procedure was repeated for all individual point mutations to assess their potential to create cryptic splice sites. For the

calculations we used the Python implementation of MaxEnt (maxentpy, v0.0.1, https://github.com/kepbod/maxentpy). We filtered the output by keeping only windows that contained a GU or AG dinucleotide in the positions 4-5 (5' splice site) or 19-20 (3' splice site), respectively.

We compared the effects of single point mutations in our library to predictions by the state-of-the-art deep learning algorithm SpliceAI {Jaganathan, 2019 #6}. We ran SpliceAI (v1.3.1) with the default parameters plus masking (-M1), using GENCODE {Frankish, 2019 #52} (v31) annotation for the human genome version hg38 as a reference. Given that SpliceAI results are reported in terms of a probability of gain or loss of a particular splice site, we assigned the gained splice sites in our cryptic isoforms by comparison to the canonical exon 2 inclusion isoform, such that if a new splice site appears in the cryptic isoform, it is considered as 'gained' with respect to the 'lost' WT splice site. All splice sites in a cryptic isoform were given the same prevalence score, i.e., the prevalence score of the mutation-isoform pair. To compare the SpliceAI scores for a given splice site gain with our prevalence score (**Figure 4F**), we considered the mutations that (i) share the same gain-loss pair of positions in both assays, and (ii) are predicted by SpliceAI to gain of a new splice site (i.e., a cryptic site where score_gain > score_loss) upon a given mutation.

RBP binding site predictions

For the prediction of RBP binding motifs, we used the web versions of the oRNAment (http://rnabiology.ircm.qc.ca/oRNAment) {Benoit Bouvrette, 2020 #8} and ATtRACT (https://attract.cnic.es/) {Giudice, 2016 #7} databases to query the minigene sequence for presence of RBP motifs (**Figure S5**). From the obtained predictions, we collapsed overlapping binding sites from the same tool and RBP.

We used DeepRiPe {Ghanbari, 2020 #9} to predict the potential impact of single point mutations on RBP binding. To this end, we downloaded the trained models for PAR-CLIP and ENCODE eCLIP data on 159 RBPs available in the Github repository (https://github.com/ohlerlab/DeepRiPe). We scored each mutation (annotated with regards to the hg38 reference genome) across the individual RBP models and preserved every mutation for which the model score changed by at least 0.25 compared to the WT sequence. The scoring functions are based on the iPython notebooks provided by DeepRiPe: https://colab.research.google.com/drive/18yeqRE7KmOjfbUaLAfJ6rMBjAulYo-Uc?usp=sharing

For the definition of significant RBP binding sites, we used the following strategy. For binding sites predicted by oRNAment and ATtRACT, we first checked their overlap separately for each isoform. If a binding site overlapped in at least one position with a splicing-effective mutation with respect to this particular isoform, we defined this binding site as an isoform-specific significant binding site. All binding sites that were significant for at least one isoform were collapsed into the complete list of significant binding sites, yielding a total of 315 significant binding sites for 74 RBPs. In the case of DeepRiPe, a mutation with a delta score > 0.25 for a given RBP model was required to overlap with a splicing-effective mutation for a particular isoform (our screen) to be considered an isoform-specific significant RBP-changing mutation. In a similar manner, all isoform-specific mutations for any isoform were collapsed into a complete list of significant RBP-changing mutations, yielding a total of 222 significant mutations that affected the binding of 58 RBPs.

iCLIP data processing

iCLIP libraries were sequenced on an Illumina NextSeq 500 sequencing machine as 92 nt single-end reads including a 6 nt sample barcode as well as 5+4 nt unique molecular identifiers (UMIs). Basic quality controls were done with FastQC (v0.11.8) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and reads were filtered based on sequencing qualities (Phred score) in the barcode region using the FASTX-Toolkit (v0.0.14) (http://hannonlab.cshl.edu/fastx toolkit/) and seqtk (v1.3) (https://github.com/lh3/seqtk/). Reads were de-multiplexed based on the experimental barcode, which is found on positions 6 to 11 of the reads, using Flexbar {Roehr, 2017 #53} (v3.4.0). Afterwards, barcode regions and adapter sequences were trimmed from read ends using Flexbar. Here, a minimal overlap of 1 nt of read and adapter was required, UMIs were added to the read names and reads shorter than 15 nt were removed from further analysis. Downstream analysis was done as described in Chapters 3.4 and 4.1 of Busch et al. {Busch, 2020 #54}. Genome assembly and annotation of GENCODE {Frankish, 2019 #52} v31 were used during mapping.

Data availability

All the sequencing data is available as a SuperSeries collection in the Gene Expression Omnibus (GEO) under the accession number GSE182894. The collection consists of the PacBio DNA-seq libraries (GSE182891), the Illumina RNA-seq libraries (GSE182892) and the PTBP1 iCLIP2 libraries in NALM-6 cells (GSE182893).

Scripts used to process the files are accessible under the GitHub repository located at: https://github.com/mcortes-lopez/CD19_splicing_mutagenesis.

Competing interests

A.T.-T. has an interest in intellectual property “Discovery of CD19 Spliced Isoforms Resistant to CART-19”. This interest does not meet the definition of a reviewable interest under Children’s Hospital of Philadelphia’s (CHOP’s) conflict of interest policy and is therefore not a financial conflict of interest. Furthermore, this intellectual property has not been licensed or otherwise commercialised to date. However, should this technology be commercialised in the future, A.T.-T. would be entitled to a share of royalties earned by CHOP per its patent policy.

The other authors have no competing interests.

Acknowledgements

The authors would like to thank the members of the participating labs for support and discussion. We gratefully acknowledge the Institute of Molecular Biology Core Facilities for their support, especially the Genomics Core Facility and the use of its NextSeq 500 (funded by the Deutsche Forschungsgemeinschaft [DFG, German Research Foundation] INST 247/870-1 FUGG) and the Bioinformatics Core Facilities. We gratefully acknowledge the PacBio SMRT sequencing platform at MPI-CBG Dresden. xx TARGET ALL?

Author contributions

M.C.-L. performed most bioinformatics analyses. L.S. performed the *CD19* minigene experiments as well as the massively parallel *CD19* splicing reporter assay. L.S. and B.S performed shRNA-mediated RBP knockdown experiments and corresponding splicing assays. M.E. and S.L. designed the mathematical modelling and prevalence score approach, and M.E. performed the analyses. F.K. contributed to quantification of mutation effects. A.O., M.C.-L., L.S. and J.K. performed PTB iCLIP experiments. A.B. performed iCLIP and RNA-seq data processing as well as splice isoform quantification. M.Q.-V. and M.T.D., performed TARGET ALL data analysis under supervision of Y.B. and A.T.-T.. Study was designed by M.C.-L., L.S., M.E., K.Z., S.L. and J.K. with help from C.P., J.F. and all co-authors. K.Z., S.L. and J.K. supervised most of the bioinformatics analyses, mathematical modelling, and experimental work, respectively. M.C.-L, L.S., M.E., C.P., K.Z., S.L., and J.K. wrote the manuscript with help and comments from all co-authors.

Funding

This work was funded by the Naturwissenschaftlich-Medizinische Forschungszentrum (NMFZ) to J.F., J.K. and C.P. and the Deutsche Forschungsgemeinschaft (DFG) to K.Z., J.K. and S.L. (ZA 881/2-3 to K.Z., KO 4566/4-3 to J.K., and LE 3473/2-3 to S.L.). K.Z. was also supported by the Deutsche Forschungsgemeinschaft (SFB902 B13). Xx
This work was supported by the grant from the National Institutes of Health (U01 CA232563 to A.T.-T. and Y.B.), St. Baldrick’s-Stand Up to Cancer (SU2C-AACR-DT-27-17 to A.T.-T.) and the V Foundation for Cancer Research (T2018-014 to A.T.-T.).

Figure captions

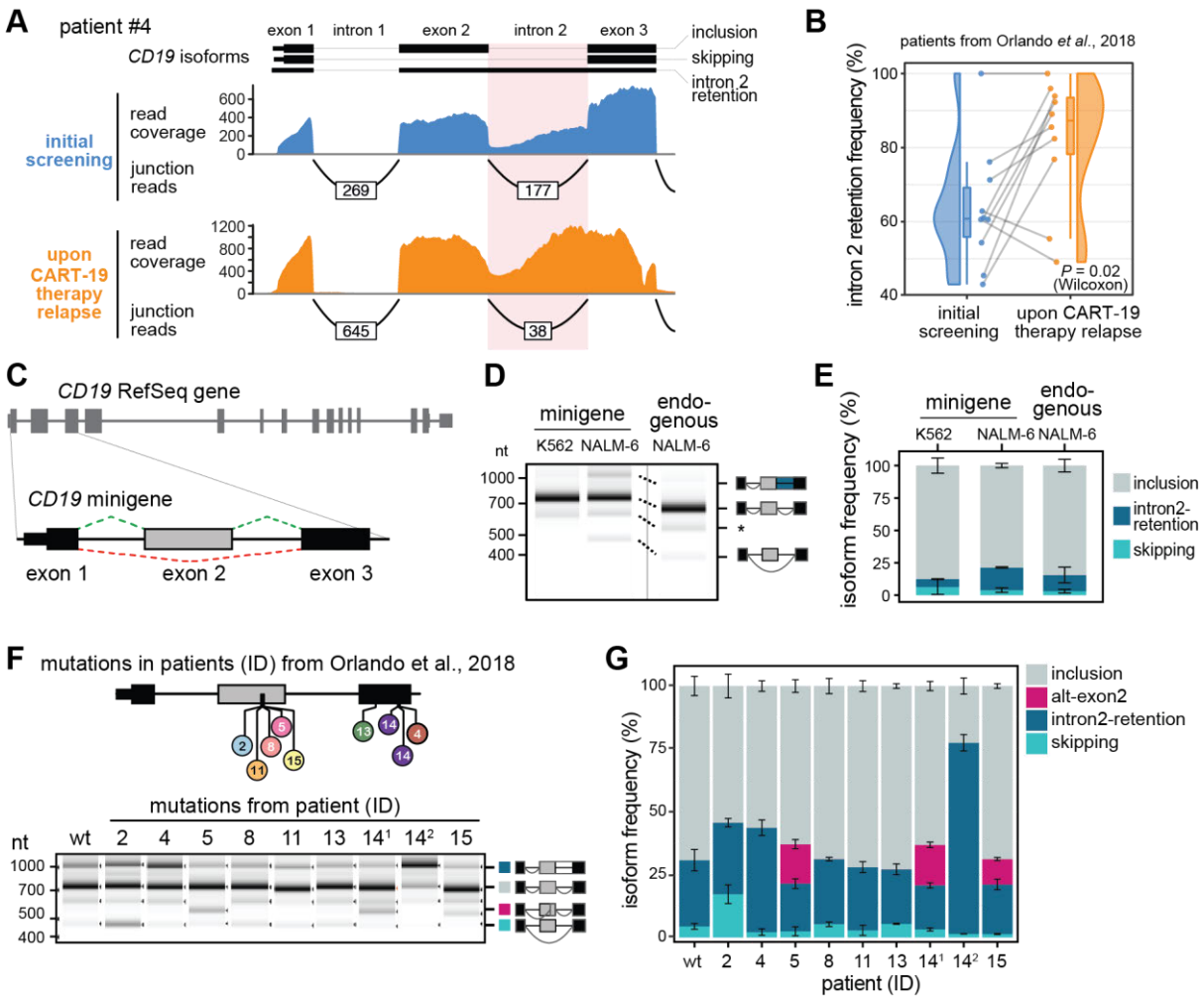


Figure 1. Mutations from B-ALL patients cause *CD19* mis-splicing.

(A) Patient #4 shows increased *CD19* intron 2 retention after CART-19 therapy relapse, evidenced by reduced junction-spanning reads and increased intron coverage. Re-analysed RNA-seq data from Orlando et al. {Orlando, 2018 #10}. Selected isoforms (GENCODE) are shown above.

(B) Intron 2 retention increases in B-ALL patients after CART-19 therapy relapse. Intron 2 retention frequency (as % of all isoforms) is shown for 10 patients with matched RNA-seq data at screening and after relapse. *P* value = 0.02, paired Wilcoxon signed-rank test.

(C) The *CD19* minigene spans exons 1-3 and the intervening introns from the *CD19* gene.

(D, E) The minigene generates the same isoforms as the endogenous *CD19* gene in NALM-6 cells. Gel-like representation (D) and quantification (E) of semi-quantitative RT-PCR showing detected isoforms intron2-retention (blue), inclusion (grey) and skipping (turquoise) for the WT minigene in NALM-6 and K562 cells as control. Isoforms of *CD19* gene in NALM-6 cells are shown for comparison. Asterisk indicates a previously reported RT-PCR artefact {Schulz, 2021 #34} (see methods). Error bars indicate standard deviation of mean (s.d.m.), n = 3 replicates.

(F, G) Patient mutations cause splicing changes in the *CD19* minigene. Top: Location of the tested mutations. Numbers refer to patient IDs as reported in Orlando et al. {Orlando, 2018 #10}. 14.1 and 14.2 correspond to distinct mutations from patient #14. Gel-like representation (F) and quantification (G) of semi-quantitative RT-PCR as in (D) and (E). Additional isoform

alt-exon2 (purple) includes a truncated version of exon 2. Error bars indicate s.d.m., n = 3 replicates.

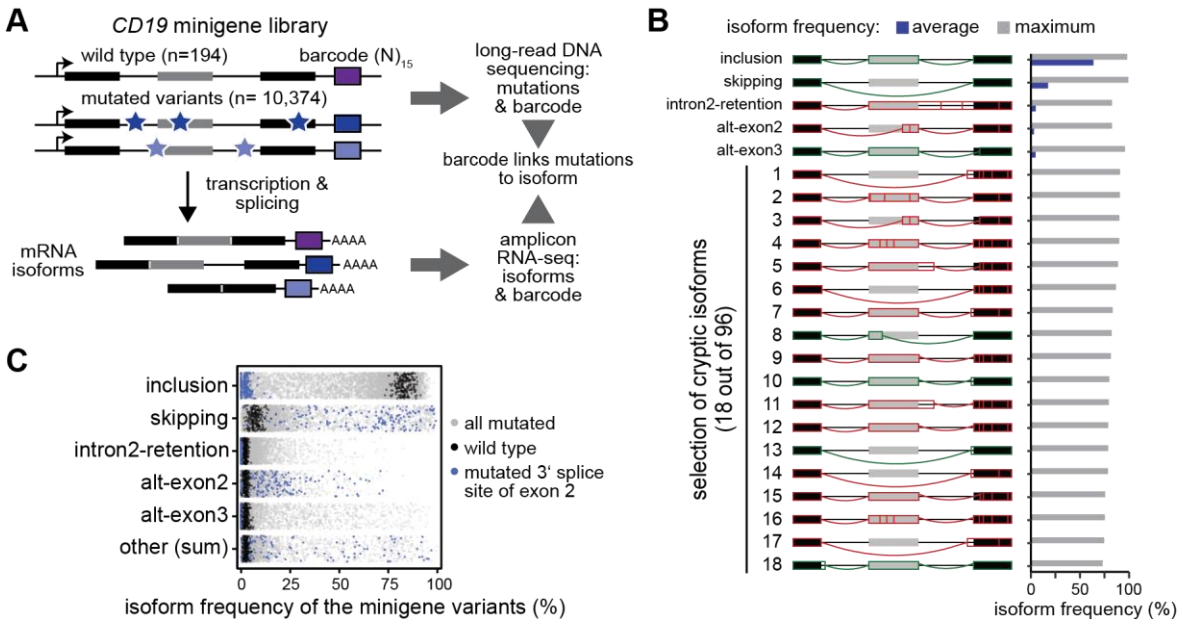


Figure 2. High-throughput mutagenesis identifies splicing-effective mutations and cryptic isoforms in the *CD19* minigene.

(A) High-throughput detection of splicing-effective mutations and cryptic isoforms. Mutagenic PCR creates mutated minigene variants (top) that upon transfection into NALM-6 cells give rise to alternatively spliced transcripts (bottom). Mutations (stars) and corresponding splicing products are characterised by DNA and RNA sequencing, respectively, and linked by a unique 15-nt barcode sequence in each minigene (coloured boxes). Black and grey boxes depict constitutive and alternative exons, respectively.

(B) A large number of *CD19* splice isoforms arise in the minigene library. *CD19* splice isoforms with highest maximal isoform frequency across all 9,321 minigene variants. Schematic representation (left) of 5 major and 18 cryptic isoforms depicts exons 1-3 (boxes) and introns (horizontal lines) with splice junctions for each isoform (arches). Bar graph (right) shows average and maximal isoform frequency across all minigenes. Cryptic isoforms are sorted by maximal isoform frequency (**Table S2**).

(C) Inclusion isoform dominates in WT minigenes, whereas mutated variants show broad spread in all major isoforms. Frequencies of five major isoforms in replicate 1 for all wild type (black; n = 195) and mutated (grey; n = 9,476) minigenes in the library. Minigene variants harbouring a mutation in the 3' splice site of exon 2 (n = 174) are highlighted in blue. 'Other' refers to the sum of 96 cryptic isoforms.

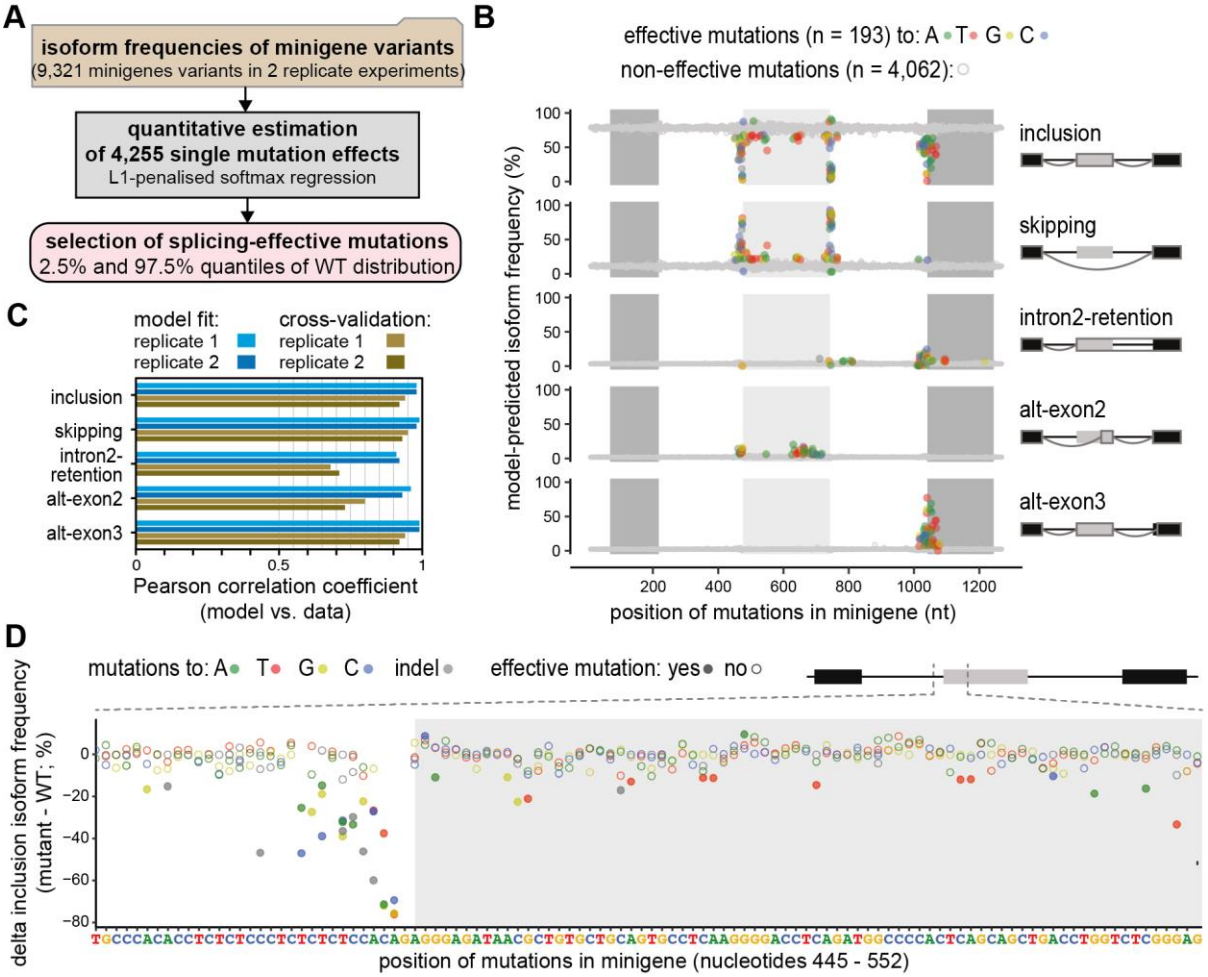


Figure 3. Quantitative modelling predicts single mutation effects on splice isoforms.
(A) Multinomial logistic regression workflow for the quantification and selection of single mutation effects. Based on the experimentally measured frequencies of five major isoforms in 9,321 minigene variants (top box), a softmax regression model was formulated to estimate 4,255 single mutation effects from the data (middle box) using L1 penalisation to prevent overfitting. Splicing-effective mutations were selected for each isoform based on comparison with the respective empirical WT frequency distribution using the 2.5% and 97.5% quantiles as cutoff.
(B) Splicing-effective mutations accumulate in distinct regions around exons 2 and 3. Landscape of model-predicted single mutation effects on five major isoforms (indicated on the right). Predicted isoform frequencies are plotted as a function of the position of a mutation. Colours indicate the nucleotide substitution of splicing-effective point mutations (see legend), whereas non-effective mutations are grey.
(C) The model performs well in fitting and 10-fold cross-validation. Bars show Pearson correlation coefficients between model and data for two replicates and each of the five isoforms across all combined mutation minigenes considered in model training and validation, respectively. See **Figure S3A** for corresponding scatter plots.
(D) Zoom-in shows the model-predicted delta inclusion isoform frequency (frequency for a point mutation - frequency in WT) for nucleotides 445-552 of the minigene. The type of

nucleotide substitution is shown for all mutations, with splicing-effective mutations highlighted as filled circles.

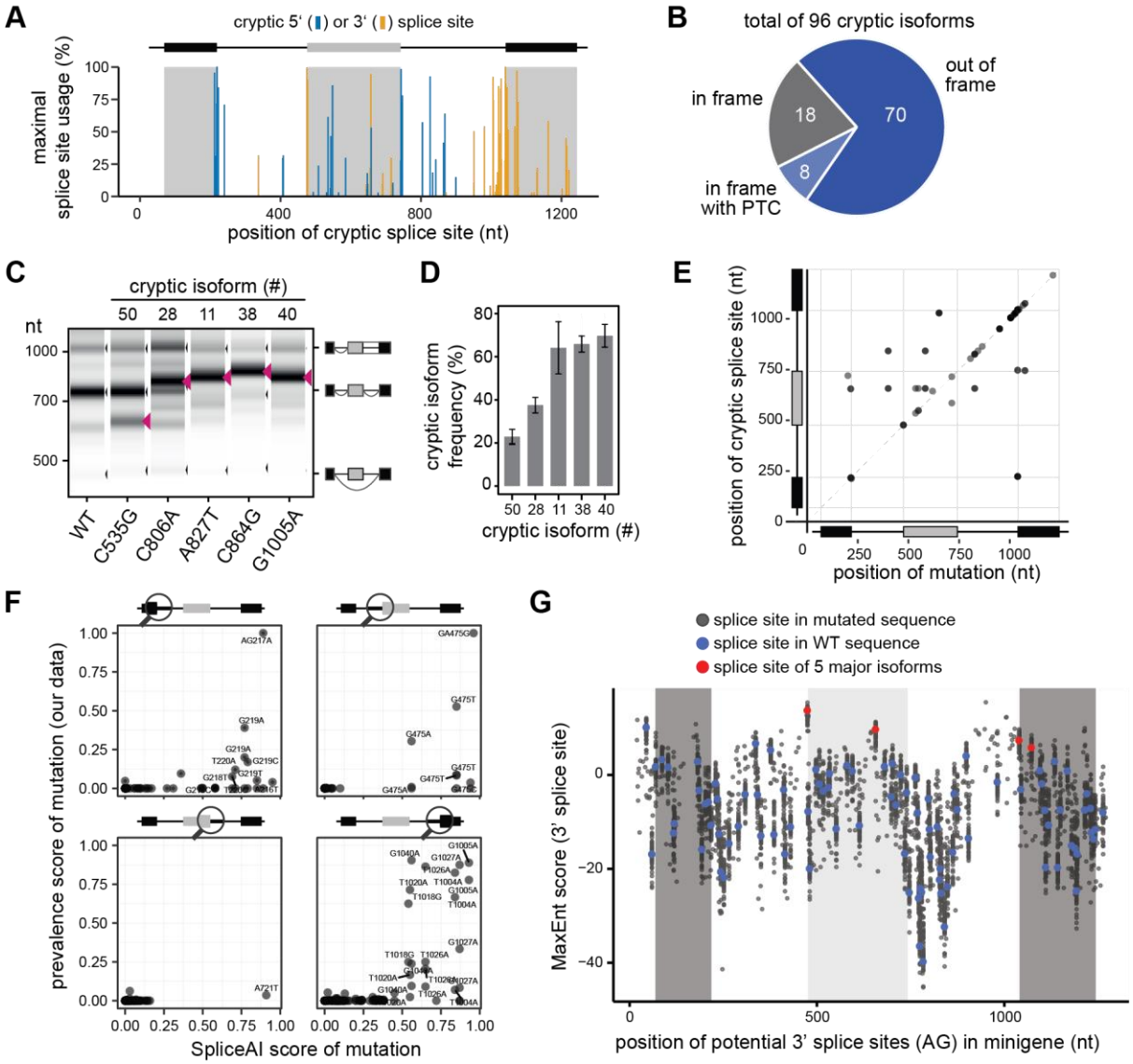


Figure 4. *CD19* mutations frequently activate cryptic splice sites.

(A) Alternative splicing of *CD19* minigene variants involves 71 cryptic splice sites. Splice site usage was calculated for each minigene variant by dividing the sum of junction reads involving a particular splice site by the total number of reads. The maximum usage across all minigenes is plotted against the corresponding position to the cryptic splice sites.

(B) Cryptic isoforms code for non-functional CD19 proteins. Out of 96 cryptic isoforms, 8 run into a premature termination codon (PTC) and 70 are out-of-frame, thus potentially encoding non-functional CD19 protein variants. The remaining 18 remain in frame, but are shortened or extended relative to the reference inclusion isoform.

(C, D) Experimental validation of five point mutations that are associated with distinct cryptic isoforms. Targeted point mutations were introduced into the *CD19* minigene, and splicing outcomes were determined by semi-quantitative RT-PCR. Predicted cryptic isoforms are indicated by red arrowheads. Gel-like representation (C), with major isoforms indicated on the right, and quantification (D). Error bars indicate s.d.m., n = 3 replicates.

(E) Mutations leading to cryptic isoforms are often located within or near cryptic splice sites. For 31 cryptic isoforms that are highly associated with a mutation (prevalence score > 0.25;

y-axis), the position of this mutation (x-axis) was related to the position of the used cryptic splice site (y-axis).

(F) SpliceAI successfully predicts single mutations leading to the generation of cryptic isoforms. SpliceAI was used to predict changes in splice junctions based on pre-mRNA sequence for all possible *CD19* minigene single mutants. SpliceAI scores of 0 and 1 reflect 0% or 100% probability to gain a cryptic splice site in response to a mutation, respectively (see Methods). Scatter plots compare the SpliceAI score against the prevalence score from our data which quantifies the association of a mutation with a cryptic isoform. Separate panels are shown for each region around a canonical splice site (circle in schematic minigene representation). For xx mutations, SpliceAI predicts activation of the same cryptic splice site that is also activated in our experimental data (shown here), while for xx mutations it predicts an effect on a different cryptic splice site (**Figure S4B**).

(G) Exon 3 harbours a weak 3' splice site and is preceded by a high number of potentially competing cryptic 3’ splice sites, which often reach similar strength upon mutation. Dotplot shows splice site strengths (MaxEnt score) for putative 3’ splice sites (AG dinucleotides) in the *CD19* minigenes. MaxEnt score was calculated in a 23-nt sliding window for the WT sequence (red and blue dots) and hypothetical mutant minigenes, in which all possible single point mutations were introduced (grey dots). The 3’ splice sites used in the five major isoforms are highlighted in red.

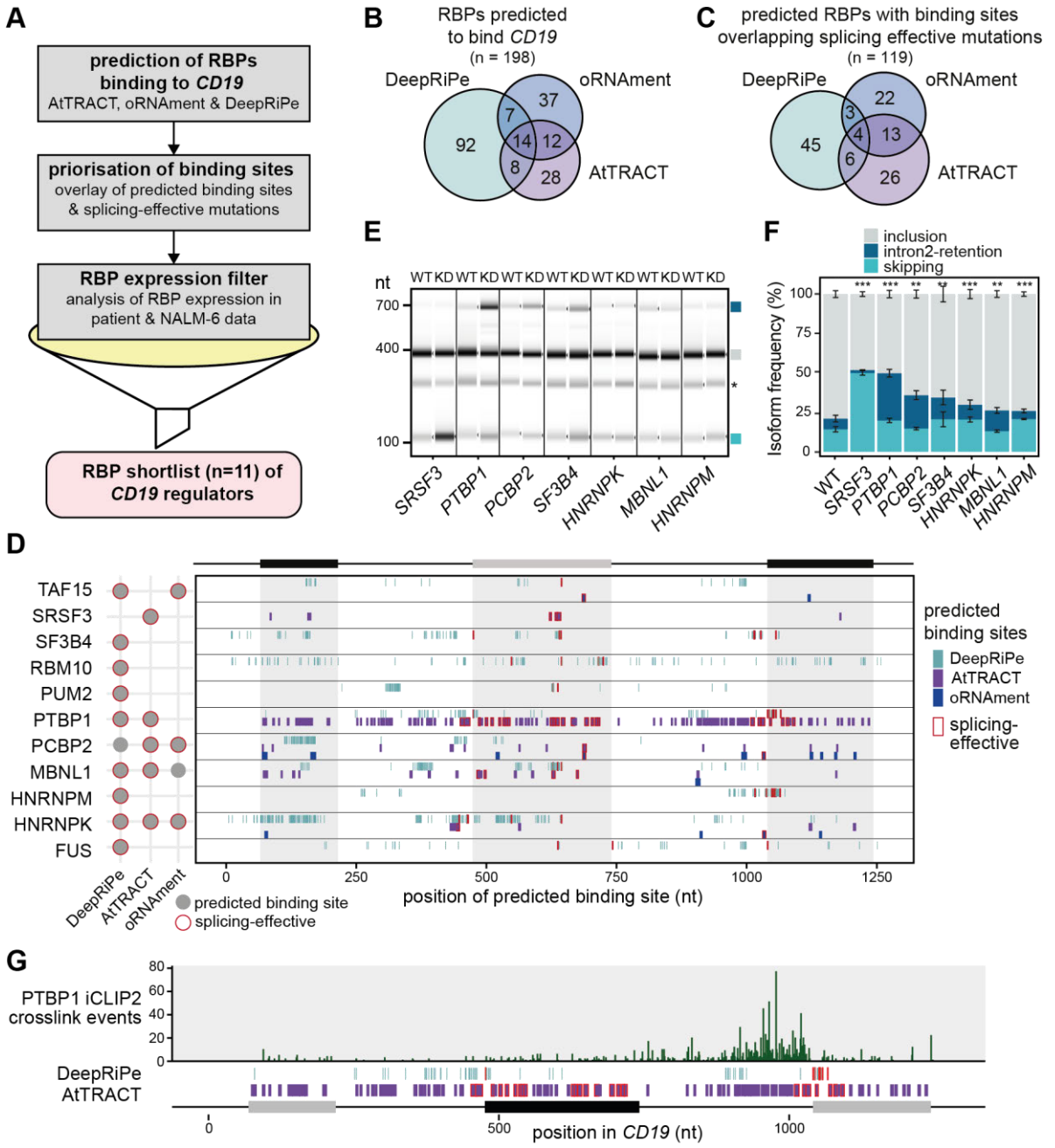


Figure 5. *In silico* predictions identify RBP regulators of *CD19* alternative splicing.
(A) Pipeline for the identification of potential RBP regulators of *CD19* splicing. Starting with *in silico* predictions, we obtained 198 candidate RBPs with predicted binding motifs (ATtRACT/oRNAment) or predicted differential binding upon mutation (DeepRiPe). These were prioritised by overlapping with the splicing-effective mutations from our screen. Additionally, based on publicly available RNA-seq data, we required a minimum mean expression in RNA-seq data from B-ALL patients {Gu, 2019 #46} and NALM-6 cells {Barretina, 2012 #68}. Together with literature information, we shortlisted 11 candidate RBPs for knockdown (KD) experiments, including SRSF3 as a positive control.
(B, C) *In silico* analyses predict dozens of RBPs binding to *CD19*. Venn diagrams show overlap of RBPs in initial predictions (B) and after overlay with splicing-effective mutations (C).

(D) The 11 candidate RBPs are predicted to bind throughout the *CD19* minigene region. For each RBP, the binding sites predicted by ATtRACT and oRNAment and disrupting mutations predicted by DeepRiPe, are indicated (see legend). Sites overlapping with splicing-effective mutations are framed in red. The schematic summary (left) shows that all 11 candidate RBPs have at least one predicted site that overlaps with a splicing-effective mutation. A full list of predicted binding sites (ATtRACT/oRNAment) and differential binding mutations (DeepRiPe) is provided in **Table S5**.

(E, F) Seven RBP KDs significantly change *CD19* splicing. Gel-like representation (E) and quantification (F) of semi-quantitative RT-PCR showing detected isoforms exon 2 inclusion (grey), intron 2 retention (blue) and skipping (turquoise) from the endogenous *CD19* gene in KD and control NALM-6 cells. Asterisk indicates a previously reported RT-PCR artefact {Schulz, 2021 #34} (see methods). Error bars indicate s.d.m., n = 3 replicates. ** *P* value < 0.01, *** *P* value < 0.001, Student's *t*-test. Measurements for all 11 KD experiments are shown in **Figure S6B, C**.

(G) PTBP1 shows extensive binding to *CD19* intron 2. Bar diagram shows the number of PTBP1 iCLIP crosslink events from NALM-6 cells on each nucleotide in endogenous *CD19* exons 1-3. Predicted PTBP1 binding motifs (ATtRACT) and mutations predicted to alter PTBP1 binding (DeepRiPe) are shown below (see legend in panel D). Nucleotide positions are given relative to minigene sequence.

High-throughput mutagenesis identifies mutations and RNA-binding proteins controlling *CD19* splicing and CART-19 therapy resistance

Mariela Cortés-López[1#], Laura Schulz[1#], Mihaela Enculescu[1#], Claudia Paret[2], Bea Spiekermann[1], Anke Busch[1], Anna Orekhova[1], Fridolin Kielisch[1], Mathieu Quesnel-Vallières[3], Manuel Torres-Diz[4], Jörg Faber[2], Yoseph Barash[3], Andrei Thomas-Tikhonenko[4 5], Kathi Zarnack[6*], Stefan Legewie[1 7*], and Julian König[1*]

[1] Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz, Germany. [2] Department of Pediatric Hematology/Oncology, Center for Pediatric and Adolescent Medicine, University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz, Germany & University Cancer Center (UCT), University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz & German Cancer Consortium (DKTK), site Frankfurt/Mainz, Germany, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. [3] Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, US and Department of Biochemistry and Biophysics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, US. [4] Division of Cancer Pathobiology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, US. [5] Department of Pathology & Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, US. [6] Buchmann Institute for Molecular Life Sciences (BMLS) and Faculty Biological Sciences, Goethe University Frankfurt, Max-von-Laue-Str. 15, 60438 Frankfurt, Germany. [7] Department of Systems Biology and Stuttgart Research Center for Systems Biology (SRCSB), University of Stuttgart, Stuttgart, Germany.

# These authors contributed equally.
* Corresponding authors: Kathi Zarnack (kathi.zarnack@bmls.de), Stefan Legewie (legewie@iig.uni-stuttgart.de), Julian König (j.koenig@imb-mainz.de)

SUPPLEMENTARY MATERIAL

Table of content:

Supplementary Figures

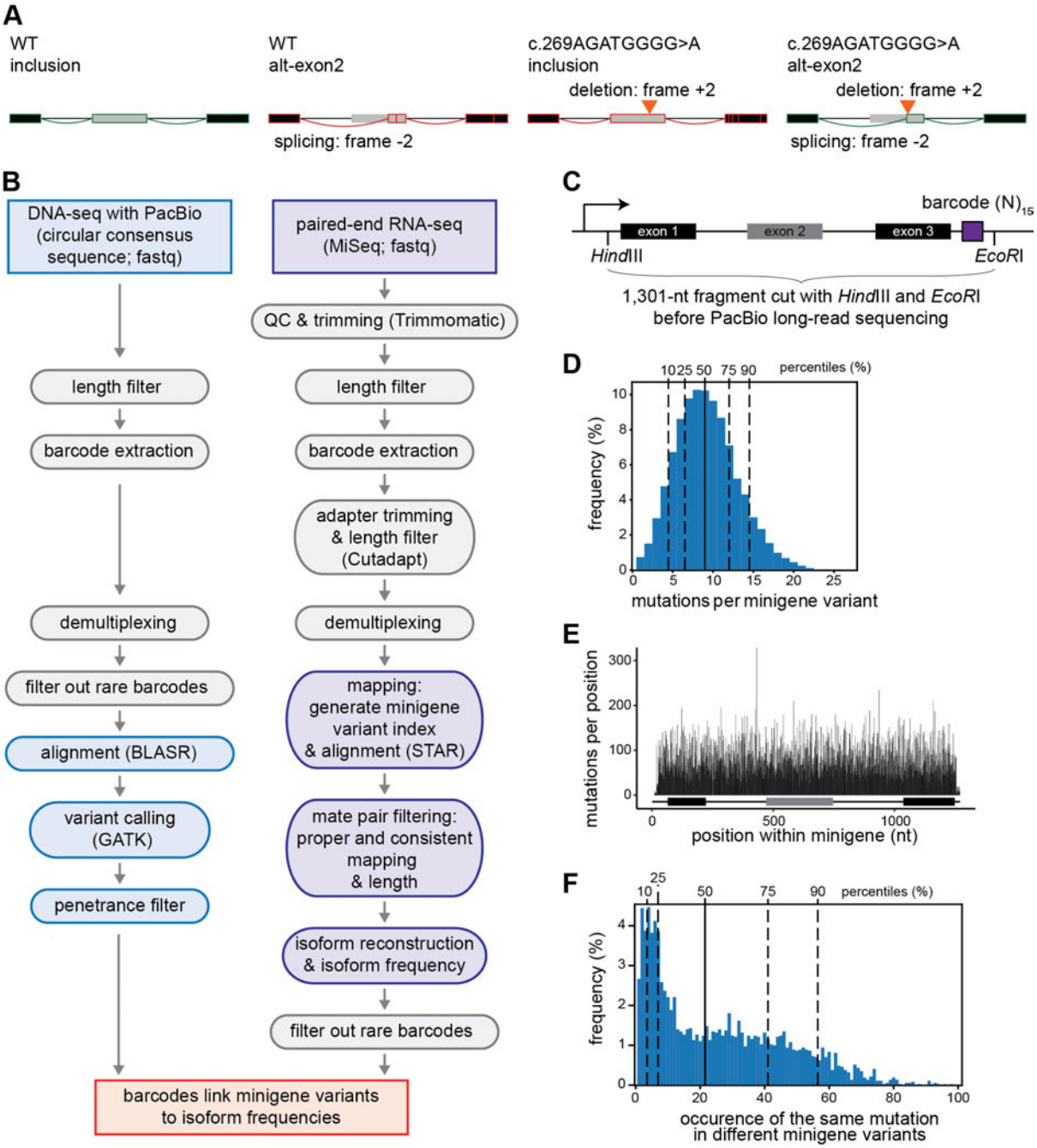


Figure S1. Long-read sequencing identifies the introduced mutations. (A) The deletion c.269AGATGGGG>A from patient #5 in Orlando et al. [1] introduces a frameshift (+2) that is compensated by the activation of an out-of-frame cryptic splice site (-2). Shown are the major isoforms inclusion and alt-exon2 and their coding potential in the absence (left) or presence (right) of the deletion (orange arrowhead). Schematic representation of depicts exons 1-3 (boxes) and introns (horizontal lines) with splice junctions for each isoform (arches). Colour indicates coding potential (green, coding; red, non-coding). **(B)** Analysis pipeline for the targeted DNA-seq and RNA-seq data. Left: Long-read DNA-seq data (PacBio, Pacific Bioscience) in the form of circular consensus sequences (CSS) were filtered by length (1,150-1,500 nt). 15-nt barcodes were extracted and demultiplexed, keeping only minigenes supported by at least 4 CSS. Alignment to the minigene reference was performed with BLASR [2] and variants were called using GATK HaplotypeCaller [3]. Mutations in the minigene were

filtered by the “penetrance score” (allele frequency, AF), discarding all the barcodes with more than 25% variants of low penetrance (AF < 0.8). Right: Short-read RNA-seq data (Illumina) were trimmed based on quality using Trimmomatic [4] and filtered by length (305 nt for read 1, 157 nt for read 2), and adapters were trimmed using Cutadapt [5] and 15-nt barcodes were extracted and demultiplexed, keeping only minigenes supported by at least 100 read pairs. Alignment to the specific mutated version of the minigene was performed using STAR [6]. Isoform reconstruction and isoform frequency estimation was done using custom scripts (see Methods). Only minigenes with 100 or more read pairs usable for isoform reconstruction were kept. **(C)** Structure of the *CD19* minigene fragment for long-read sequencing (PacBio) to identify introduced mutations. The minigene covers exons 1-3 with the intervening introns, followed by a 15-nt barcode. The fragment for PacBio sequencing is defined by the restriction sites for *Hind*III upstream of exon 1 and *EcoR*I downstream of the barcode sequence. **(D)** 91.6% of the minigene variants carry five or more mutations. Histogram shows number of mutations per minigene for 10,295 mutated minigene variants. **(E)** 4,255 distinct mutations are spread along the *CD19* minigene, with an average of 21 mutations per position. Barplot shows the sum of mutations per position in the minigene. **(F)** 81.9% of the mutations occur in at least three minigenes, which is sufficient for a reliable estimation of single mutation effects (**Figure S3C**). Histogram shows the frequencies of the same mutations in different minigene variants.

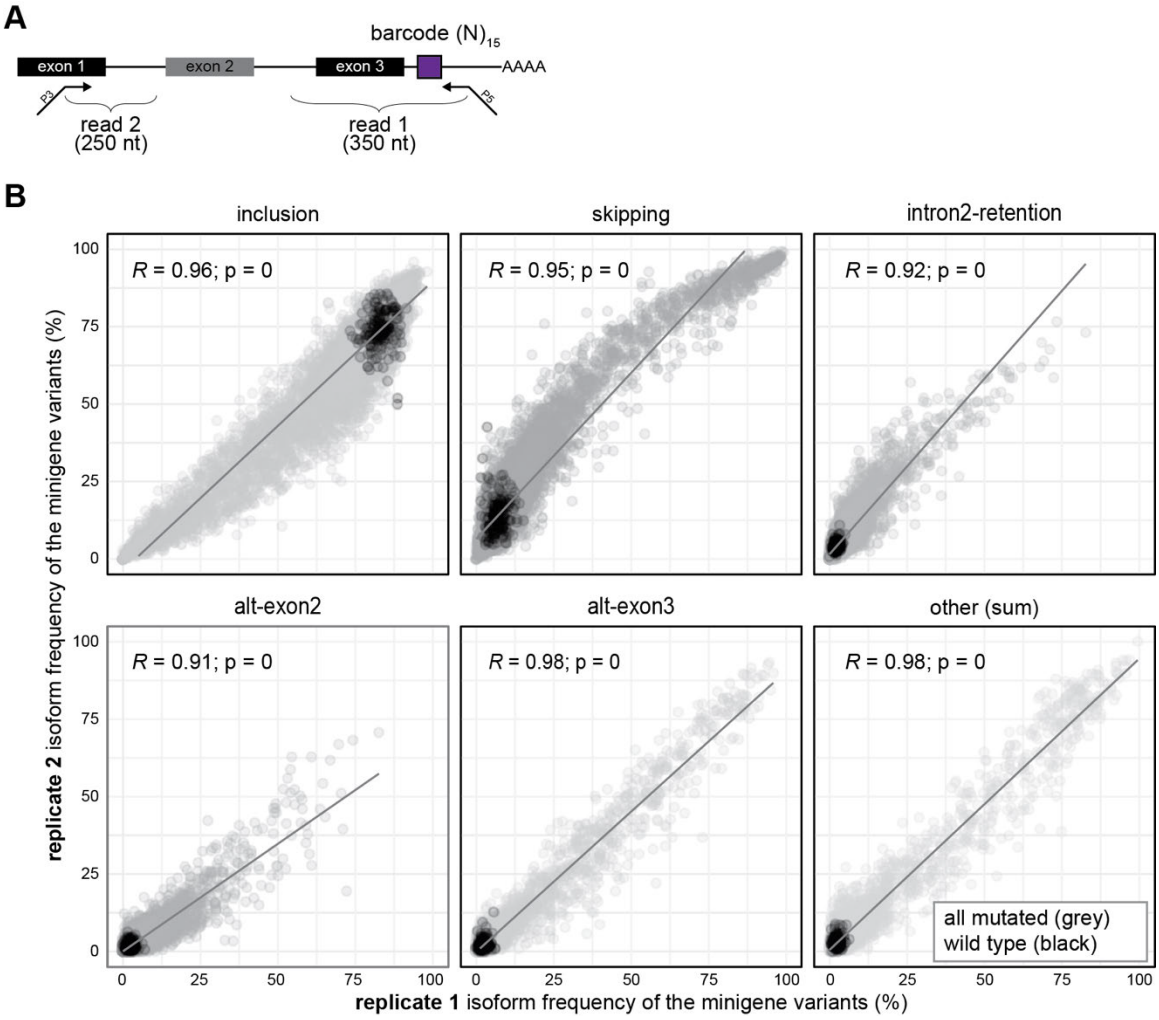


Figure S2. Isoform measurements from targeted RNA-seq results are consistent between replicates. (A) Description of the short-read RNA-seq strategy (Illumina) to capture the splicing products in the *CD19* minigene. Read 2 (250 nt) extends beyond exon 1, i.e. covering the exon 1/exon 2 junction, while read 1 (350 nt) includes the 15-nt barcode and extends beyond exon 3. **(B)** The isoform measurements correlate well between replicates. Scatterplots compare isoform frequencies for five major isoforms as well as the sum of 96 cryptic isoforms between replicate 1 and 2. Each dot represents a particular minigene captured in both replicates. WT and mutated minigenes appear in black and grey, respectively. Pearson correlation coefficients (*R*) and associated *P* values are given.

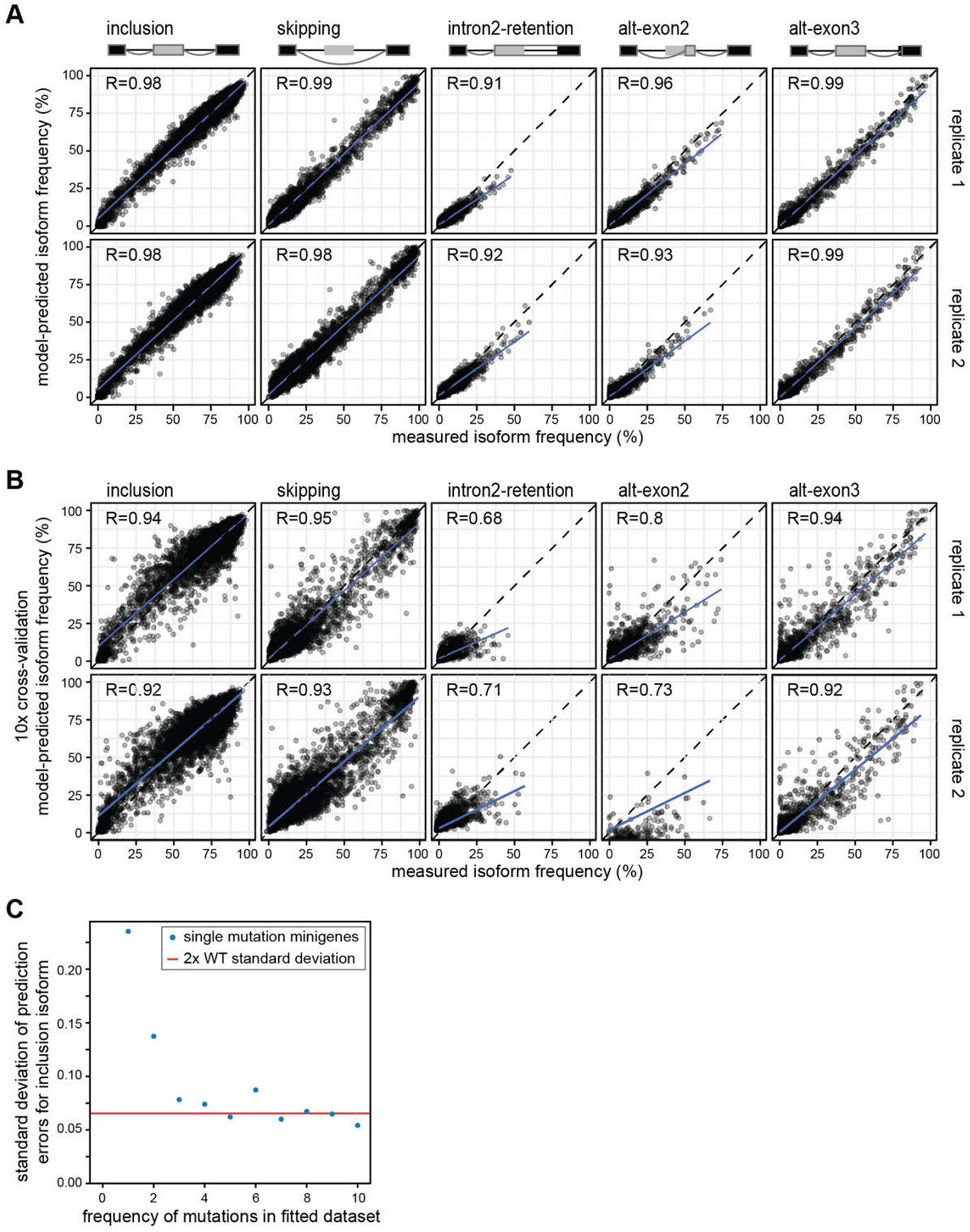


Figure S3. The softmax regression model performs well for training and test data. (A) Regression model fits measured combined mutation effects (i.e., minigene measurements) with high accuracy. Scatterplots show frequencies of the five major isoforms in the measurements (x-axis) against the model fit (y-axis) for two biological replicates and 9,321 minigene variants used in model training. Pearson correlation coefficients (*R*) are shown for each scatter plot. **(B)** Cross-validation confirms the predictive power of the model for minigenes not used in training. The minigene library was randomly split into ten equally sized subsets. During 10-fold cross-validation, the softmax regression model was fitted to all data excluding one subset. Scatterplots compare model-predicted splicing outcome for left-out

subsets to corresponding experimental data for all major splice isoforms and are an overlay of the results of all cross-validation runs. Representation as in (A). **(C)** The model correctly infers single mutation effects. Seven single-mutation minigenes in which inclusion is significantly changed were left-out separately from softmax regression fitting and their effects were predicted based on the fit to the remaining minigene data. This procedure was repeated while additionally excluding random permutations of other minigenes containing the mutation. The standard deviation of the prediction error (y-axis) is plotted against the number of minigenes used in model training (x-axis). The inference power of the model reaches two standard deviations of the WT minigenes (horizontal line) if more than two minigenes containing the mutation are considered in model training. See Methods for details.

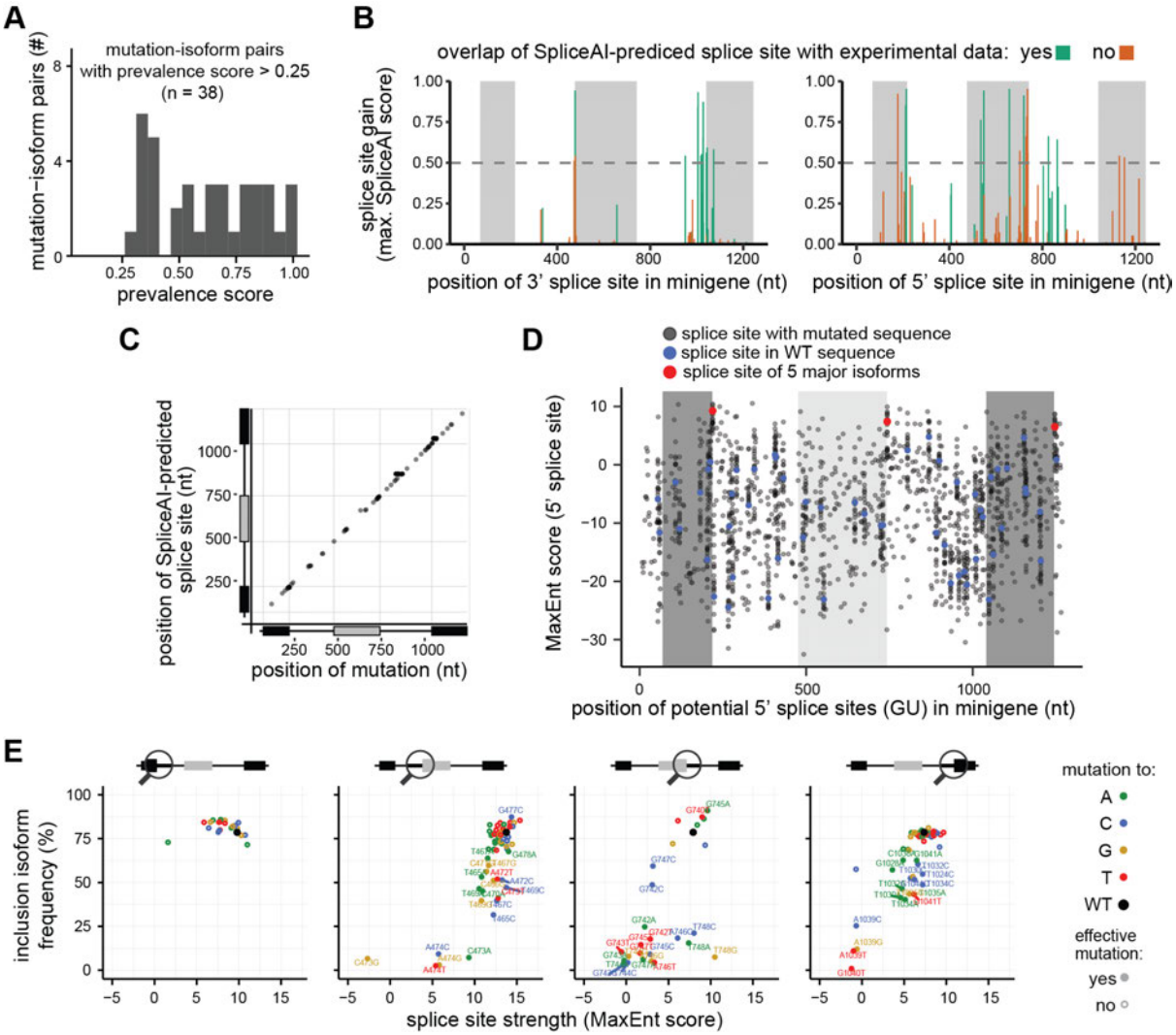


Figure S4. Multiple mutations give rise to distinct cryptic isoforms. (A) Multiple mutations are associated with a specific cryptic isoform. Histogram shows distribution of prevalence scores for 38 mutation-isoform pairs in which a specific mutation is associated with a distinct cryptic isoform (prevalence score > 0.25). A prevalence score of 1 indicates perfect correspondence between mutation and isoform. **(B)** SpliceAI [7] predictions for gained cryptic splice sites overlap with experimental data. Barplot shows the maximum SpliceAI score ("acceptor gain") for all the mutations that increase the probability of a given cryptic splice site to be used (38 mutations with Splice AI score [gain] > 0.5, including 15 and 23 gained 3' [left] and 5' splice sites [right]). Dotted horizontal line represents the recommended minimum threshold for a SpliceAI prediction (SpliceAI score > 0.2) [7]. Predicted gained splice sites that also appear in our experimental data are shown in green. **(C)** The mutation effects predicted by SpliceAI exclusively occur in close range, such that all SpliceAI-predicted effective mutations reside on average within 6 nt from the cryptic splice site generated. Scatterplot shows location of the gained cryptic splice sites with respect to the mutations. Only the splice site with the highest score for each mutation is considered. **(D)** The 5' splice sites of the main isoforms (red) are stronger than most other 5' splice sites in the *CD19* minigene sequence. Dotplot shows splice site strengths (MaxEnt score) [8] for putative 5' splice sites in WT (blue) and mutated (grey) minigenes in a 9-nt sliding window containing a GU dinucleotide at positions 4-5. 5' splice sites used in the five major isoforms are shown in red. **(E)** Mutation effects at 3' and 5' splice sites of *CD19* exons 2 and 3 are consistent with predicted splice site strengths. Mutations are coloured according to the changed nucleotides. Scores for WT sequence are coloured in black. Splicing-effective mutations (according to our results) are shown as filled circles and labelled, while non-effective mutations are shown as open circles.

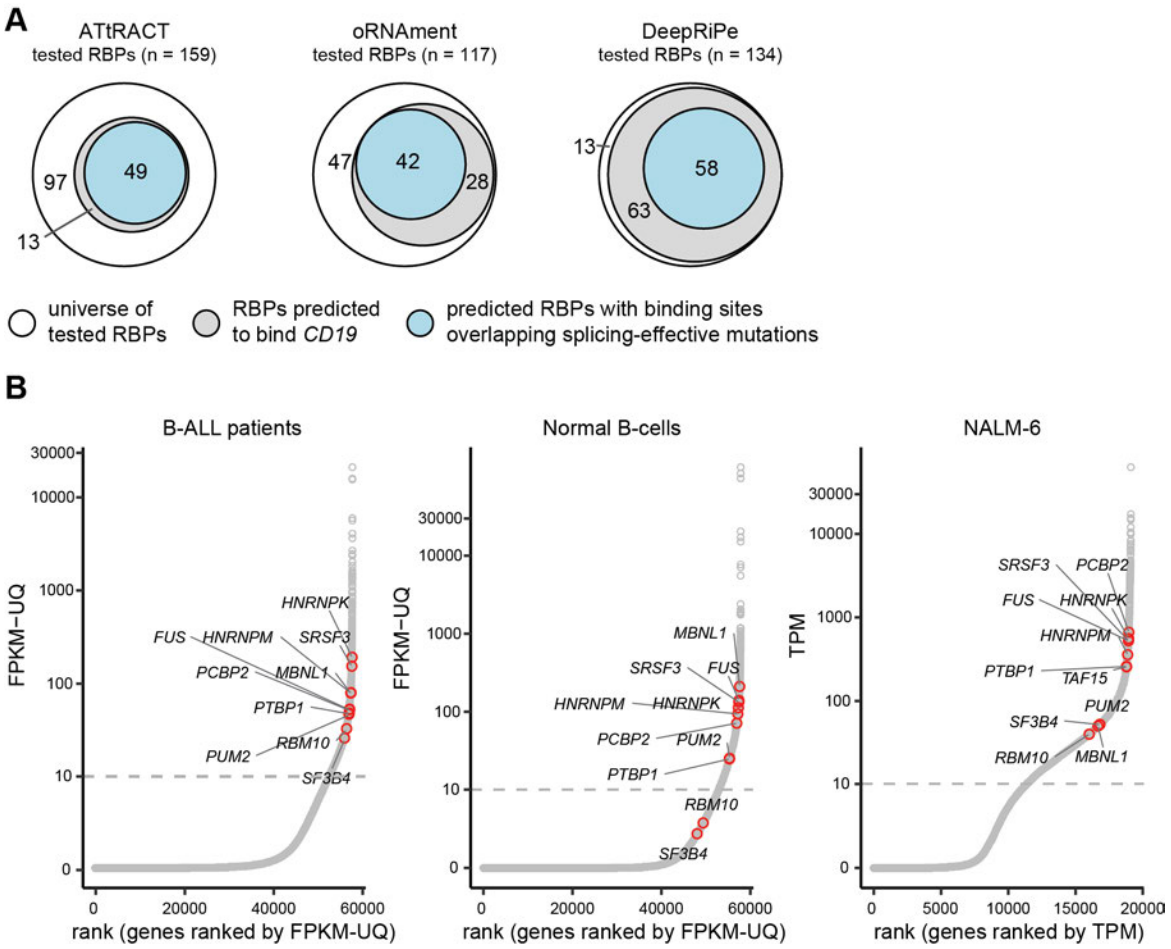


Figure S5. *In silico* RBP binding site predictions suggest dozens of candidate regulators of *CD19* alternative splicing. (A) *In silico* predictions of RBP binding sites were performed with ATtRACT [9] and oRNAment [10] as well as of point mutations affecting RBP binding using DeepRiPe [11]. For each prediction tool, the total number of available RBPs (white circles) is split up into those that are predicted to bind *CD19* (grey circles) and whose predicted binding sites overlap with splicing-effective mutations from our data (blue circles). Numbers refer to exclusive RBPs in each area. **(B)** Predicted RBPs were filtered based on their expression observed in B-ALL patients reported in [12]. Plot shows ranked expression values for all detected genes in samples from B-ALL patients, normal B-cells [13] and NALM-6 cells [14]. Highlighted in red are the RBP candidate genes (n = 11) tested in knockdown experiments. TPM, transcripts per million. FPKM-UQ, fragments per kilobase of transcript per million mapped reads upper quartile, a modified RNA-seq normalisation method (https://docs.gdc.cancer.gov/Encyclopedia/pages/HTSeq-FPKM-UQ/).

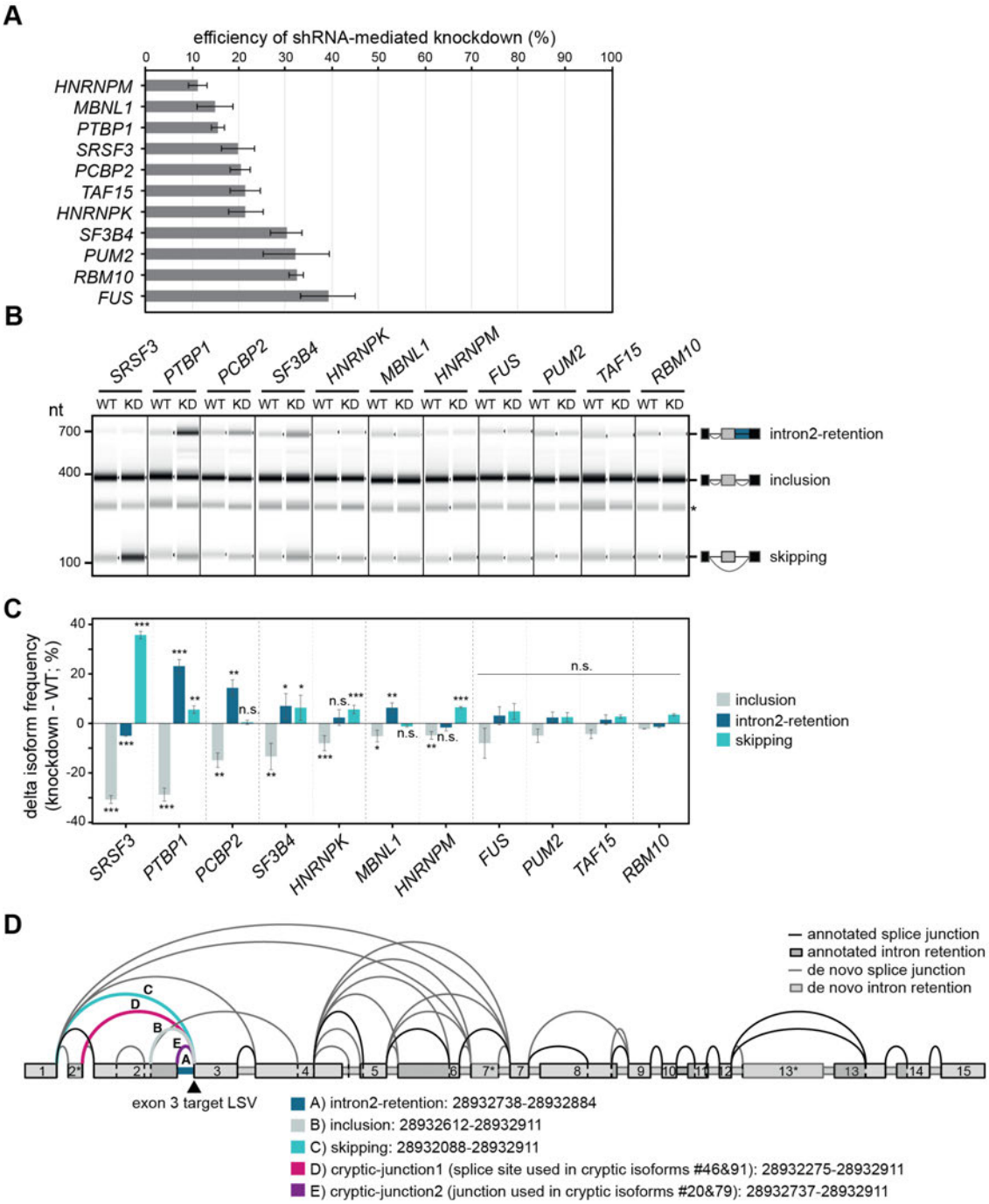


Figure S6. Knockdown experiments show significant effects on endogenous *CD19* splicing for seven candidate RBPs. (A) All tested RBPs are efficiently depleted upon shRNA knockdown (KD). Barplot shows mean qPCR measurements of remaining transcripts (relative to WT) for 11 candidate RBPs. Error bars indicate standard deviation of the mean (s.d.m.), n = 3 replicates. **(B, C)** Seven RBP knockdowns significantly affect *CD19* alternative splicing. Semiquantitative RT-PCR was performed to detect isoforms generated from exons 1-3 of the endogenous *CD19* gene. Gel-like representation (B), with major isoforms indicated on the right, and quantification (C), as difference in isoform frequency compared to WT, are shown. Error bars indicate s.d.m., n = 3 replicates. * *P* value < 0.05, ** *P* value < 0.01, *** *P* value < 0.01, n.s., not significant, Student's *t*-test. **(D)** *CD19* shows extensive mis-splicing in B-ALL patients. Splice junctions were quantified with MAJIQ [15] for 222 B-ALL patients from the Therapeutically Applicable Research To Generate Effective Treatments (TARGET) program (https://ocg.cancer.gov/programs/target). Splicegraph shows all splice junctions with a usage

level (percent selected index, PSI) of at least 5% in any patient. Junctions and target exon of the local splicing variation (LSV) shown in **Figure 5H, I** are highlighted.

Supplementary Data

Data S1. Single mutation effects on the major isoforms from the *CD19* minigene in NALM-6 cells. For each isoform, the y-axis shows the isoform frequency (mean of two biological replicates) resulting from each individual mutation in a given position along the y-axis. Each dot represents one mutation, with colours indicating the inserted nucleotide (green, mutation to A; blue, to C; yellow, to G; red, to T). Splicing-effective mutations are shown as filled circles and non-effective mutations as open circles. Dashed lines indicate the median isoform frequency of the WT minigenes (black) ± 2 standard deviations (grey). The shown isoforms are *CD19* exon 2 inclusion, skipping, intron2-retention, alt-exon2 and alt-exon3 as well as the sum of 96 cryptic isoforms ("other").

Supplementary Tables

Table S1. Mutations from relapsed B-ALL patients reported in Orlando et al. that were tested in the *CD19* minigene splicing reporter. Patient IDs are given as reported in Orlando et al. [1]. Note that for patient #14, two separate minigene variants were tested (#14.1 and #14.2), and that #14.2 is a combination of two adjacent mutations reported in patient #14, namely c.509A>AGTGG and c.510GCCTC>GTGGGGGAG.

patient ID	mutation	genomic coordinate (hg38)	position in minigene	reference allele (REF)	alternative allele (ALT)
#2	c.259G>GGGGGC	chr16:28932516	646	G	GGGGGC
#4	c.517TGTCTCCCACCG>T	chr16:28933072	1202	TGTCTCCCACCG	T
#5	c.269AGATGGGG>A	chr16:28932526	656	AGATGGGG	A
#8	c.265CA>C	chr16:28932522	652	CA	C
#11	c.264TCAACAGATGGGGGGCTTCTACCTGTGC>T	chr16:28932521	651	TCAACAGATGGGGGGCTTCTACCTGTGC	T
#13	c.421T>TC	chr16:28932976	1106	T	TC
#14.1	c.297GGGGC>G	chr16:28932554	684	GGGGC	G
#14.2	c.510AGCCTC>AGTGGGGGAG	chr16:28933065	1195	AGCCTC	AGTGGGGGAG
#15	c.271ATGGGGGGCTTCTACCTGTGCCAGCCGGGGCCC>AAGACGT	chr16:28932528	658	ATGGGGGGCTTCTACCTGTGCCAGCCGGGGCCC	AAGACGT

Table S2. Quantification of splicing isoforms for all minigene variants in the library. For each minigene variant, the 15-nt barcode sequence is shown together with the contained mutations, with multiple mutations separated by commas. The total number of reads per minigene variant and their distribution among the 101 isoforms are given for RNA-seq replicates 1 and 2 from NALM-6 cells. Isoform notation (219 475) indicates a splice junction that removed the region from nucleotides 219 to 475. The five major isoforms are *CD19* exon 2 inclusion (219 475)(743 1040), skipping (219 1040), intron2-retention (219 475), alt-exon2 (219 657)(743 1040) and alt-exon3 (219 475)(743 1073). In total, we detected splicing isoforms for 9,671 minigene variants in replicate 1 and for 9,372 minigene variants in replicate 2, including 9,321 minigene variants that were present in both replicates.

< provided as Excel file >

Table S3. List of detected isoforms from the *CD19* minigene. A total of 101 isoforms reached a relative frequency of at least 5% in at least one minigene variant, including the five major isoforms inclusion, skipping, intron2-retention, alt-exon2 and alt-exon3 (>5% in WT) as well as 96 cryptic isoforms. For each isoform, the assigned name or number is shown together with the isoform specification. Isoform notation (219 475) indicates a splice junction that removed the region from nucleotides 219 to 475. With respect to the predicted impact on the encoded CD19 protein, the number of premature stop codons (PTCs), the frame (in-frame or out-of-frame) and the resulting coding potential (coding or non-coding) are reported. With respect to an isoform's relative abundance, the average isoform frequency in the library and the maximal isoform frequency in an individual minigene are given. For the 38 cryptic isoforms that are associated with a specific mutation (prevalence score > 0.25), the respective mutations are provided together with their prevalence score and genomic coordinate (hg38). Notation G475T indicates that G in position 475 was mutated to T.

< provided as Excel file >

Table S4. Single mutation effects predicted by the mathematical model. Worksheet "Mutation effects" provides the model estimates of splice isoform frequencies (in %) and average delta frequency (compared to WT) in replicates (rep) 1 and 2 in response to individual mutations (single nucleotide variants, SNV; insertions or deletions, INDEL) in NALM-6 cells. Notation G475T indicates that G in position 475 was mutated to T. Individual entries are given for each affected isoform. Isoform notation (219 475) indicates a splice junction that removed the region from nucleotides 219 to 475. The five major isoforms are *CD19* exon 2 inclusion (219 475)(743 1040), skipping (219 1040), intron2-retention (219 475), alt-exon2 (219 657)(743 1040) and alt-exon3 (219 475)(743 1073). Worksheet "WT statistics" provides the mean, standard deviation (sd) and median of measured splice isoform frequencies (in %) for the five major isoforms as well as the sum of 96 cryptic isoforms ("other"). Isoform frequencies were measured for 195 and 194 WT minigenes in the two replicates.

< provided as Excel file >

Table S5. Overlapping single nucleotide variants (SNVs) and cancer-related mutations. Worksheet "Annotated variants" contains the SNVs (from ENSEMBL [16] v104, gnomAD [17] v3.1 and ClinVar [18] accessed 09/2021) and cancer-related variants (obtained from COSMIC [19] v94) that overlap with splicing-effective mutations and mutations with a prevalence score > 0.25 in our screen. Notation A950G indicates that A in position 950 was mutated to G. For variants present in the database dbSNP [20], the respective ID is also included. REF and ALT refer to the reference and alternative allele.

< provided as Excel file >

Table S6. Predicted RBP binding sites in the region of the *CD19* minigene. Worksheet "Binding sites" reports *in silico* predictions by ATtRACT [9] and oRNAment [10], providing the source tool, start and end and width (relative to the *CD19* minigene), predicted RNA-binding protein (RBP) and whether the binding site overlaps with splicing-effective mutations from our screen (see Methods). Worksheet "DeepRiPe mutations" reports all mutations predicted by DeepRiPe [11] to change RBP binding (i.e., with a delta score > 0.25), including RBP, mutation, DeepRiPe score and set as well as whether the mutation overlaps with a splicing-effective mutation from our screen and if so, for which isoform. Set refers to the DeepRiPe model that was trained for a given RBP using PAR-CLIP or ENCODE eCLIP data from HepG2 or K562 cells (see [11] for details).

< provided as Excel file >

Table S7. Oligonucleotides used to clone the different shRNA sequence carrying vectors in this study. Oligonucleotides were purchased from Integrated DNA Technologies.

shRNA_FUS	TGCTGTTGACAGTGAGCGCACAGGATAATTCAGACAACAATAG TGAAGCCACAGATGTATTGTTGTCTGAATTATCCTGTTTGCCTA CTGCCTCGGA
shRNA_HNRNPK	TGCTGTTGACAGTGAGCGACGAGTTGAGGCTGTTGATTCATAG TGAAGCCACAGATGTATGAATCAACAGCCTCAACTCGCTGCCT ACTGCCTCGGA
shRNA_HNRNPM	TGCTGTTGACAGTGAGCGAAGCAGACATTCTTGAAGATAATAGT GAAGCCACAGATGTATTATCTTCAAGAATGTCTGCTCTGCCTAC TGCCTCGGA
shRNA_MBNL1	TGCTGTTGACAGTGAGCGCCAGCACAATGATTGACACCAATAG TGAAGCCACAGATGTATTGGTGTCAATCATTGTGCTGTTGCCTA CTGCCTCGGA
shRNA_PCBP2	TGCTGTTGACAGTGAGCGCTCCATCATTGAGTGTGTCAAATAGT GAAGCCACAGATGTATTTGACACACTCAATGATGGATTGCCTAC TGCCTCGGA
shRNA_PTBP1	TGCTGTTGACAGTGAGCGCTAGCAAGATGATACAATGGTATAG TGAAGCCACAGATGTATACCATTGTATCATCTTGCTATTGCCTA CTGCCTCGGA
shRNA_PUM2	TGCTGTTGACAGTGAGCGCAACATAGTTGTTGACTGTTAATAGT GAAGCCACAGATGTATTAACAGTCAACAACTATGTTATGCCTAC TGCCTCGGA
shRNA_RBM10	TGCTGTTGACAGTGAGCGCCGGCAAGACCATCAATGTTGATAG TGAAGCCACAGATGTATCAACATTGATGGTCTTGCCGTTGCCTA CTGCCTCGGA
shRNA_SF3B4	TGCTGTTGACAGTGAGCGCTGCCTTCAAGAAGGACTCCAATAG TGAAGCCACAGATGTATTGGAGTCCTTCTTGAAGGCATTGCCTA CTGCCTCGGA
shRNA_SRSF3	TGCTGTTGACAGTGAGCGCTAAGATGTTTTAGCTGTTCAATAGT GAAGCCACAGATGTATTGAACAGCTAAAACATCTTAATGCCTAC TGCCTCGGA
shRNA_TAF15	TGCTGTTGACAGTGAGCGATCAGGCTATGATCAACATCAATAGT GAAGCCACAGATGTATTGATGTTGATCATAGCCTGACTGCCTAC TGCCTCGGA

Table S8. qPCR oligonucleotide pairs used in this study. Oligonucleotides were purchased from Sigma-Aldrich.

	Forward primer	Reverse primer
qPCR_FUS	AAGGCCTGGGTGAGAATGTT	GGCTGTCCCGTTTTCTTGTT
qPCR_HNRNPK	GCGAGTTGAGGCTGTTGATT	TCAGTGGAATGAGGACAGCA
qPCR_HNRNPM	GTCAAGGGGATGTGCTGTTG	TCCGCTCAGACTATGCTTGT
qPCR_MBNL1	CGGTTTGCTCATCCTGCTGA	TTTGCACTTTTCCCGAGAGC
qPCR_PCBP2	CCAGCTCTCCGGTCATCTTT	CTGGTGCAGCTTGGTCAAAT
qPCR_PTBP1	CGAGATGAACACGGAGGAGG	CTGGATGTAGATGGGCTGGC
qPCR_PUM2	TCAGCGTCCTCTTACTCCCA	CCAGTAGCAAGACCCTGACC
qPCR_RBM10	TGTTCCCGACGTCTCTACCT	TCTCCCCATCCCAGTACAGG
qPCR_SF3B4	GAACGACTTCTGGCAGCTCA	CACAGGATTGGGAGCAGAGG
qPCR_SRSF3	CCCGGCTTTGCTTTTGTTGA	TTCCACTCTTACACGGCAGC
qPCR_TAF15	GGTCACAGGGAGGAGGTAGA	CAGCATCTGTTCTGGGTCCA

Supplementary References

1. Orlando EJ, Han X, Tribouley C, Wood PA, Leary RJ, Riester M, Levine JE, Qayed M, Grupp SA, Boyer M, De Moerloose B, Nemecek ER, Bittencourt H, Hiramatsu H, Buechner J, Davies SM, Verneris MR, Nguyen K, Brogdon JL, Bitter H, Morrissey M, Pierog P, Pantano S, Engelman JA & Winckler W. Genetic mechanisms of target antigen loss in CAR19 therapy of acute lymphoblastic leukemia. *Nat Med* **24**, 1504-1506 (2018).
2. Chaisson MJ & Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
3. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M & DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
4. Bolger AM, Lohse M & Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
5. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12 (2011).
6. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M & Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
7. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, Chow ED, Kanterakis E, Gao H, Kia A, Batzoglou S, Sanders SJ & Farh KK. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548 e524 (2019).
8. Yeo G & Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377-394 (2004).
9. Giudice G, Sanchez-Cabo F, Torroja C & Lara-Pezzi E. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database (Oxford)* **2016** (2016).
10. Benoit Bouvrette LP, Bovaird S, Blanchette M & Lecuyer E. oRNAment: a database of putative RNA binding protein target sites in the transcriptomes of model species. *Nucleic Acids Res* **48**, D166-D173 (2020).
11. Ghanbari M & Ohler U. Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Res* **30**, 214-226 (2020).
12. Gu Z, Churchman ML, Roberts KG, Moore I, Zhou X, Nakitandwe J, Hagiwara K, Pelletier S, Gingras S, Berns H, Payne-Turner D, Hill A, Iacobucci I, Shi L, Pounds S, Cheng C, Pei D, Qu C, Newman S, Devidas M, Dai Y, Reshmi SC, Gastier-Foster J, Raetz EA, Borowitz MJ, Wood BL, Carroll WL, Zweidler-McKay PA, Rabin KR, Mattano LA, Maloney KW, Rambaldi A, Spinelli O, Radich JP, Minden MD, Rowe JM, Luger S, Litzow MR, Tallman MS, Racevskis J, Zhang Y, Bhatia R, Kohlschmidt J, Mrozek K, Bloomfield CD, Stock W, Kornblau S, Kantarjian HM, Konopleva M, Evans WE, Jeha S, Pui CH, Yang J, Paietta E, Downing JR, Relling MV, Zhang J, Loh ML, Hunger SP & Mullighan CG. PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nat Genet* **51**, 296-307 (2019).
13. Alexander TB, Gu Z, Iacobucci I, Dickerson K, Choi JK, Xu B, Payne-Turner D, Yoshihara H, Loh ML, Horan J, Buldini B, Basso G, Elitzur S, de Haas V, Zwaan CM, Yeoh A, Reinhardt D, Tomizawa D, Kiyokawa N, Lammens T, De Moerloose B, Catchpoole D, Hori H, Moorman A, Moore AS, Hrusak O, Meshinchi S, Orgel E, Devidas M, Borowitz M, Wood B, Heerema NA, Carrol A, Yang YL, Smith MA, Davidsen TM, Hermida LC, Gesuwan P, Marra MA, Ma Y, Mungall AJ, Moore RA, Jones SJM,

Valentine M, Janke LJ, Rubnitz JE, Pui CH, Ding L, Liu Y, Zhang J, Nichols KE, Downing JR, Cao X, Shi L, Pounds S, Newman S, Pei D, Guidry Auvil JM, Gerhard DS, Hunger SP, Inaba H & Mullighan CG. The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature* **562**, 373-379 (2018).

14. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jane-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P, Jr., de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R & Garraway LA. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).
15. Vaquero-Garcia J, Barrera A, Gazzara MR, González-Vallinas J, Lahens NF, Hogenesch JB, Lynch KW & Barash Y. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5**, e11752 (2016).
16. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Charkhchi M, Cummins C, Da Rin Fioretto L, Davidson C, Dodiya K, El Houdaigui B, Fatima R, Gall A, Garcia Giron C, Grego T, Guijarro-Clarke C, Haggerty L, Hemrom A, Hourlier T, Izuogu OG, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Gonzalez Martinez J, Marugan JC, Maurel T, McMahon AC, Mohanan S, Moore B, Muffato M, Oheh DN, Paraschas D, Parker A, Parton A, Prosovetskaia I, Sakthivel MP, Salam AIA, Schmitt BM, Schuilenburg H, Sheppard D, Steed E, Szpak M, Szuba M, Taylor K, Thormann A, Threadgold G, Walts B, Winterbottom A, Chakiachvili M, Chaubal A, De Silva N, Flint B, Frankish A, Hunt SE, GR II, Langridge N, Loveland JE, Martin FJ, Mudge JM, Morales J, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Cunningham F, Yates AD, Zerbino DR & Flicek P. Ensembl 2021. *Nucleic Acids Res* **49**, D884-D891 (2021).
17. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Genome Aggregation Database C, Neale BM, Daly MJ & MacArthur DG. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
18. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, Hoffman D, Jang W, Kaur K, Liu C, Lyoshin V, Maddipatla Z, Maiti R, Mitchell J, O'Leary N, Riley GR, Shi W, Zhou G, Schneider V, Maglott D, Holmes JB & Kattman BL. ClinVar: improvements to accessing data. *Nucleic Acids Res* **48**, D835-D844 (2020).
19. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ & Forbes SA. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947 (2019).
20. Sherry ST, Ward M & Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* **9**, 677-679 (1999).

3 Discussion

We performed a high-throughput mutagenesis screen on two minigenes. Using sequencing data of DNA and RNA from the minigenes in combination with computational modeling, we were able to determine which mutation triggers the formation of which splice isoform. In particular, for the *CD19* minigene, we identified nearly 100 previously unknown splice isoforms that emerge from cryptic splice sites and likely encode non-functional CD19 proteins. In addition, we used our mutagenesis data to learn more about *cis*-regulatory elements and *trans*-acting RBPs that control *MST1R* and *CD19* splicing. Through knockdown experiments, we confirm that HNRNPH is a master regulator of *MST1R* and several other RBPs such as PTBP1 play a role in *CD19* splicing. Last but not least, we also analyzed a putative splice isoform of *CD19*, "*CD19*ex2part", using a splice reporter as well as ONT long-read sequencing. These experiments have revealed that this alleged splice isoform is an artifact and does not play a role in B-ALL therapy resistance.

All three projects are closely interlinked. They all focus on specific cancer-related splicing events that are important for either disease progression or therapeutic success. These events can only be properly understood if all *cis*- and *trans*-regulatory factors in the corresponding cell types in which they occur are taken into account.

Another common feature is their experimental setup. All three projects are based on minigene variants. First, a few thousands of different minigene variants were generated to decipher the important *cis*-regulatory elements of a particular section of a gene via a high-throughput splicing reporter assay. In the case of the *RON* minigene, the minigenes were based on exons 10, 11, and 12. In the case of *CD19*, exons 1, 2, and 3 were considered. In addition to the high-throughput assay, some minigene variants were generated to analyze somatic mutations of cancer patients that occur within the minigene sequence. The deduced *cis*-regulatory landscape is also used to find *trans*-factors involved by using state-of-the-art RBP motif analysis. Knockdown experiments allowed to identify specific RBPs that play important roles.

In the case of the *CD19*ex2part project, no classical minigene was used: only one constitutive exon was involved in the splicing reporter plasmid. Nevertheless, the presence of an alleged exitron, an intron within an exon, allowed the plasmid to be used in a similar way as classical minigenes. The *CD19*ex2part reporter showed whether exitron splicing

occurred within exon 2 using two fluorescent reporter proteins, eGFP and mCherry, for detection.

Besides the use of minigenes, one second experimental layer was thoroughly used in all three projects: Sequencing – the heart of modern molecular biology. Almost all cutting-edge sequencing technologies, Illumina DNA- and RNA-sequencing, PacBio's SMRT (Single-Molecule, Real-Time) sequencing, and ONT (Oxford Nanopore Technology) cDNA- and dRNA-sequencing, were used throughout the studies – each of them offering their specific advantages and disadvantages.

The results of all three studies were numerous. Multiple mutations inducing skipping of *RON* exon 11, the pathological splice isoform that triggers cancer progression, were uncovered, which is consistent with the data of clinical patients. In addition, HNRNPH binding was shown to cooperatively regulate *RON* exon 11 splicing.

Further, in the *CD19* project, we identified 200 mutations that change *CD19* splicing. In addition, we identified around 100 novel cryptic splicing isoforms that most likely encode non-functional CD19 proteins, eventually predisposing patients to relapse. In addition to *cis*-regulatory factors, we also demonstrated the importance of *trans*-regulators such as SRSF3 and PTBP1 in the production of therapy-relevant isoforms. This study points to potential prognostic markers for CART-19 resistance.

In contrast to classical splice sites, the *CD19*ex2part splice isoform did not exhibit constitutive splice sites but rather a short repetitive sequence at both sites. Interestingly, we were able to show that, contrary to earlier assumptions, *CD19*ex2part is not a true splicing isoform and indeed does not exist at all. It is the result of defective reverse transcription of RNA. Following this observation, we found many more examples of such cases throughout the human genome. To properly validate unexpected splice junctions, the use of direct RNA sequencing or at least the use of modern corrective bioinformatics tools is required.

3.1 Decoding splicing decisions using minigene assays

The main goal within the splicing community remains to decipher a general splicing code that can be used for all cell types and developmental stages. However, deciphering the entire spliceosomal network remains an incredibly difficult task. As for alternative splicing, nature even intended splicing to be a process that can change dramatically under certain

conditions. It cannot be compared to the simple tabular structure of the genetic code (Wang and Burge, 2008).

There are many levels at which splicing can be influenced and altered. The state of the RNA is important. Splicing can be influenced by RNA sequence, modifications, structure, or location (speckles) (McManus *et al.*, 2011; Spector *et al.*, 2011; Lewis *et al.*, 2017).

As mentioned earlier, the number of proteins involved in splicing is immense (~150) (Zhou *et al.*, 2002). To date, it is hardly possible to get an exact list of all proteins involved in the process. How would it be possible to also determine their exact function in each splicing event? When we think about different cell types, the absence or presence of RBPs or even their altered concentration are important for splicing. For example, we know that even minor changes in RBP levels (e.g., HNRNPH) can trigger an entire splicing switch that may determine the fate of the cell (LeFave *et al.*, 2011; Braun *et al.*, 2018).

Last but not least, mutations within splicing-regulative proteins also play a major role (Sveen *et al.*, 2016). These mutations can alter either protein stability, protein modifications, or binding potential toward RNA targets or other network interactors. Importantly, the alteration of only one interactor can already lead to a cascade of interaction changes within the network, since each factor is interdependent up to a certain point.

These examples indicate how far we need to look to understand even a single specific splicing event - not to mention that there are about 19,000 protein-coding genes in the human genome (Ezkurdia *et al.*, 2014). Each of them consists on average of 11 exons and 10 introns (Piovesan *et al.*, 2016), so we can theoretically study several 100,000 splicing events at this depth.

Therefore, one approach that most RNA scientists take to achieve the goal of finding a global splicing code in the future is the "bottom-up" strategy. First, all *cis*-regulatory elements are identified. Then, thorough analysis is used to identify rules by which they are active and interact with other factors. Subsequently, integrating more and more of these rules into algorithms could be a path to decipher a global splicing code (Wang and Burge, 2008).

Minigenes are a way to understand the splicing of at least one section of a gene in great detail. They serve as the smallest possible structure for studying splicing. As the name implies, they are a simplified version of a gene. The first use of a minigene dates back almost 40 years (Kornblihtt *et al.*, 1984). Minigenes consist of an expression plasmid containing the exon of interest flanked by upstream and downstream constitutive exons. They can

help answer questions about existing *cis*-elements as well as *trans*-elements, especially when combined with knockdown or knockout experiments (Cooper, 2005; Singh and Cooper, 2006). Minigenes can be used to find new rules that apply under certain conditions. These rules can then be integrated into the algorithm, making it a little more efficient each time. The more individual examples come together, the more generally valid and precise the result becomes (Wang and Burge, 2008).

In the past, many high-throughput minigene studies have examined multiple minigene events at once (Adamson *et al.*, 2017, Mikl *et al.*, 2020). The construct remains the same, but parts of it, such as the middle exon, are completely swapped between minigene variants. This type of experiment is designed to predict the splicing behavior of a sequence as a whole. Single-nucleotide changes are not specifically studied. The study of multiple splicing events is useful, for example, to understand different clinical variants compared to the wild type. However, they cannot replace the in-depth analysis of a single splicing event to obtain a complete picture of the splicing-regulatory landscape of the minigene.

Also, most past studies have examined only the exonic regions of minigenes (Julien *et al.*, 2016; Ke *et al.*, 2019; Souček *et al.*, 2019). Introns tend to be quite long compared to exons, which causes technical problems in downstream sequencing applications. Therefore, scientists have previously neglected intronic regions in mutagenesis or even used other, shorter intronic sequences from other genetic regions. Unlike these other studies, the minigene assays in this work uses the complete introns, which represents a more natural and therefore more trustworthy structure for minigene experiments. In addition, in our analysis, both complete introns were mutated as extensively as the exons themselves. In this way, our studies allowed systematic quantification of mutational effects. We were able to show that, in addition to the exons, a large number of positions in the introns also influence splicing regulation, as has been also seen in another recent study (Conboy, 2021).

Another crucial factor for the strength of splicing-effective mutations was discussed by Baeza-Centurion *et al.* (2019). They showed that the effects of mutations on splicing do not scale monotonically. Minigenes that exhibit a more similar distribution between inclusion and skipping from baseline allow for stronger splicing changes than constructs in which one isoform is clearly prominent. These predictions also apply to our studies. For the *MST1R* minigene, where the inclusion to skipping ratio is rather balanced, we discovered several positions that strongly disrupt exon 11 inclusion upon mutation. However, in the case of the *CD19* minigene, where the inclusion frequency is dominant in the wild type, we did not find many mutations that severely disrupt this ratio, except for those in canonical

splicing signals. Nevertheless, we identified many mutations that have a strong effect on the formation and activation of multiple cryptic splice sites. This finding suggests that the impact of mutation-mediated splicing changes should go beyond the analysis of classical exon inclusion and exon skipping levels.

Of course, minigene studies also have shortcomings and limitations. One of them is that splicing is very tissue-specific (Saha *et al.*, 2017). The effects one measures in one cell line are not necessarily transferable to another cellular context. In particular, tissue types such as the brain or testis remain extremely regulation-specific (Naro *et al.*, 2021). Another challenge is also the diversity of technical setups of minigene experiments. Due to numerous differences, results are difficult to compare and integrate. However, to date, minigenes remain an indispensable experimental tool to understand splicing decisions in disease-related situations, to study both *cis*- and *trans*-elements of RNA, and to validate *in silico* predictions (Baralle and Baralle, 2005; Fraile-Bethencourt *et al.* 2018; Park *et al.*, 2018).

3.2 Relevance of direct long-read sequencing in mRNA variant calling

When using fragmented cDNA for sequencing, there are at least three steps where information can get lost: reverse transcription, fragmentation, and PCR amplification. During cDNA synthesis, reverse transcriptase can introduce errors. The accuracy of reverse transcription is highly dependent on the type of enzyme used, as they can differ in their fidelity and preferred sequence content. It also depends on the temperature at which the synthesis takes place, as the secondary structures in the RNA become looser and easier to open with increasing temperatures. Therefore, thermo-stable reverse transcriptases tend to be a safer approach than regular ones but nevertheless not completely safe (Qin *et al.*, 2016).

If one doubts the accuracy of their results obtained from reverse transcription either because the splice sites used are not constitutive or because, on a control gel, the splicing isoforms shown (at the cDNA level) vary inexplicably between experiments, further validations are necessary. Northern Blotting is an example of a fairly inexpensive method to directly estimate RNA sizes (He and Green, 2013). However, it only works if the isoforms in question differ in size enough to be adequately separated on the blot.

Fragmentation is a simple matter compared to the reverse transcriptase step: the more you fragment, meaning the shorter the reads, the more context is lost. Most short-read sequencing deals with reads of a length of up to 250 bp (Stark *et al.*, 2019). Considering that an average exon is 309 nucleotides long (Piovesan *et al.*, 2016), one can easily deduce that most reads will only contain few splice junctions. Reads can come from different mRNA molecules, so that the assignment and combination of different splice junction reads to distinct isoforms is not always possible. It is difficult and to some extent impossible to derive the actual splice isoforms from all theoretically possible isoforms when only short-read data are available.

Last but not least, PCR amplification can also lead to biases. Also in this case, the type of enzyme is of great importance. Even very confidential polymerases still cause errors (McInerney *et al.*, 2014). It is also well known that GC-rich sequences tend to cause problems in amplification steps (Benita *et al.*, 2003). The PCR step for enrichment, for example in Illumina library preparation, is the main source of base composition bias in fragmented libraries (Aird *et al.*, 2011). There is also a risk of over-amplification of the entire library. Therefore, the cycle numbers must be tared before the actual experiment (Wong *et al.*, 2013). In addition, other DNA polymerase amplification steps downstream on the instrument, such as cluster amplification or sequencing-by-synthesis, also introduce errors (Aird *et al.*, 2011).

A method that works for all isoform sizes and that excludes cDNA synthesis as well as any PCR amplification steps is direct long-read RNA sequencing (dRNA-seq) from Oxford Nanopore Technologies (ONT). It is the method of choice when there is uncertainty about RNA composition and isoforms. As we successfully demonstrated in Schulz *et al.* (2021), it avoids errors that occur during cDNA synthesis. Moreover, it is also capable of displaying the complete mRNA isoforms without loss of information and thus without the risk of mapping errors. Interestingly, it can even detect RNA modifications (Xu and Seki, 2020). Collectively, ONT dRNA-seq can solve many problems that can occur with conventional short-read sequencing (Marinov, 2017; Pyatnitskiy *et al.*, 2021).

3.3 Clinical relevance of *CD19* cryptic splicing isoforms

Using the high-throughput splice reporter assay, we found ~100 different *CD19* splicing isoforms by analyzing only the sequence between exon 1 and 3. In addition to the

constitutive splice sites, 71 additional cryptic splice sites were used to form all alternative *CD19* isoforms. Most of these splice isoforms contain open reading frame shifts, or, if still in frame, have introduced a premature stop codon (PTC). This means that if they were expressed in cells, they would theoretically be degraded by non-sense-mediated decay because the PTCs are located more than 50 nucleotides upstream from the last exon-exon junction (Nagy *et al.*, 1998). The higher the relative amount of this cryptic splicing isoform out of all the isoforms of *CD19* generated, the lower the overall expression of the *CD19* gene could be. Although mRNA levels do not always correlate linearly with gene expression, they generally correlate well across genes and also significantly within genes – although the exact correlation values are inconsistent between studies (Buccitelli *et al.*, 2020).

In the context of this thesis, CD19 protein levels were not investigated. At this point, we can only make assumptions about the actual protein levels resulting from the different isoforms. Wilhelm *et al.* (2014) reported that the correlation between predicted protein levels and actual protein levels was approximately 0.9 when considering only mRNA levels. On the other hand, this observation was challenged a few years later by Fortelny *et al.* (2017), who found a correlation of only 0.21. One possible way to test this would be to generate a *CD19* knock-out cell line, e.g. using NALM-6 cells. The cell line could be transfected with either a plasmid encoding *CD19* wild type or another plasmid encoding an alternative or cryptic *CD19* isoform, both in an open reading frame. CD19 protein levels could be compared between the two samples using a Western blot. Furthermore, both transfected cell samples could also be analyzed via flow cytometry using an antibody against CD19 similar to experiments performed in Sotillo *et al.* (2015). The flow cytometry analysis would also test cell surface abundance and availability of CD19 molecules for therapy.

Another interesting aspect that persists is the actual occurrence of these cryptic or alternative splice isoforms in patients. We already know that exon 2 skipping and intron 2 retention occur in patients and increase especially after CART-19 therapy (Sotillo *et al.*, 2015; Asnani *et al.*, 2020; preliminary Cortés-López *et al.*). But what about all the other 98 isoforms? Unfortunately, data availability is the limiting factor in most CART-19 relapse studies. Not only are these datasets sometimes withheld or only available to certain individuals (e.g. due to personal rights of the patients), but also detection of these splice isoforms remains a challenge, as only sophisticated splice analysis tools are able to reliably detect them. Hopefully, these challenges can be overcome in the future.

All of these derived mutations could be useful to study in patients even before treatment with CART-19. Many investigators assume that the mutations causing the loss of CD19 on the surface exist already in subclones prior to treatment (Fischer *et al.*, 2017; Li and Chen, 2019). Deep sequencing or single-cell sequencing at the *CD19* locus could be a way to detect problematic mutations early - before the corresponding cells eventually undergo positive selection, which can lead to tumor escape (Rabilloud *et al.*, 2021).

3.4 *Trans*-acting splicing factor PTBP-1 regulates a therapy-relevant isoform

We have shown that PTBP1 is an important regulator of endogenous *CD19* splicing. Knockdown of PTBP1 mainly increases the retention levels of intron 2, inducing the strongest changes among all newly identified RBPs. Only SRSF3, which is already known, shows stronger effects on *CD19* splicing.

Polypyrimidine tract-binding protein (PTBP) is part of the HNRNP subfamily, also called HNRNP1 (Busch and Hertel, 2011). It prefers to bind to polypyrimidine-rich segments of the RNA. There are three PTBP family members: PTBP1, PTBP2, and PTBP3. All three PTBP paralogs have more than 70% of their sequence at protein level in common. PTBP1 is a ubiquitously expressed protein known primarily to act as a suppressor of alternative splicing. (Robinson *et al.*, 2006; Zhu *et al.*, 2020). Usually, its action is connected to the occurrence of exon skipping. However, if it binds to exonic or/and flanking intronic sequences, it can also cause exon inclusion (Zhu *et al.*, 2020).

In recent years, the role of PTBP1 in cancer has become an increasingly relevant topic. Some viable therapeutic approaches related to PTBP1 in cancer have even been developed (Bredel *et al.*, 2015). It is known to be involved in processes such as migration, metastasis, proliferation, and carcinogenesis in various cancers (Zhu *et al.*, 2020). PTBP1 is upregulated in B lymphocytes that underwent B cell positive selection compared to naïve B cells (Monzón-Casanova *et al.*, 2018). It is also needed in B cell receptor-mediated antibody production (Sasanuma *et al.*, 2019). It also takes over many other functions in the immune system. For example, it regulates different ligands and receptors (such as CD40L, CD46, CD5, and IL2) and is also important during T cell activation (Domingues *et al.*, 2016).

However, concerning blood cancer, PTBP1 is not yet known to play a significant role. Nevertheless, PTBP1 is an important splicing factor that could theoretically cause further problems down the line in any cancer. Defective expression of splicing factors has become another hallmark of cancer (Ladomery, 2013). Therefore, it is not unlikely that PTBP1 is also deregulated in B cell lymphoblastic leukemia - not necessarily in a global manner, but possibly only in individual subclonal populations that might outgrow when under selection pressure.

In addition to PTBP1 and SRSF3, we also found several other splicing factors that had a significant effect on *CD19* exon 2 splicing. Most of them had only minor effects on *CD19* splicing. However, the actual shRNA-mediated knockdown of these RBPs was measured only by qPCR. Although the results indicated that the knockdown efficiency was always higher than 60%, some of the RBP candidates still had about 30% of the initial mRNA level left. Also, we cannot draw any conclusions about the actual decrease in the amount of protein resulting from the knockdown (see also chapter 3.3). We do not know what factors might have prevented a stronger shRNA-mediated knockdown. In *in vivo* cancer cells, of course, this may be different and a stronger knockdown effect may be possible, likely causing more severe splicing changes.

3.5 Evaluation of aberrant splicing in possible therapy applications

The development of therapies targeting splicing is particularly difficult because target specificity must be ensured. Thousands of mRNA molecules are coordinated simultaneously in the cell. Altering the splicing of a particular molecule without causing off-target effects remains a challenge. However, if this challenge can be overcome, it can be life-saving for people suffering from genetic diseases caused by a specific incorrect splicing event in an important gene. The two hereditary diseases spinal muscular atrophy and Duchenne muscular dystrophy are now treated with so-called splice-switching oligonucleotides (SSO). Both therapies have been first tested in animal models, later undergone clinical trials and are now approved by the FDA (Porensky *et al.*, 2012; McGreevy *et al.*, 2015; Havens *et al.*, 2016).

SSOs can be used in diseases caused by splicing defects to restore the healthy splice variant. SSOs are single-stranded RNA molecules, typically 15-25 nucleotides long, which bind in

a complementary manner to the mRNA of interest. Binding at a specific site within the mRNA prevents the binding of specific RBPs sterically. If this is a splice site, it cannot be used but it rather depends on the surrounding sequence if splicing is completely repressed or if an alternative splice site will be used. Also binding of such SSOs to other *cis*-elements can affect splicing. Depending on the nature of the hidden *cis*-element, splicing can be supported or prevented. As we recall (chapter 1.4.1), there are *cis*-elements that act as splicing enhancers and those that act as repressors (Bennett *et al.*, 2010). The use of SSOs to prevent CART-19 resistance may be possible. However, at the moment we do not have a suitable target. Our problem is that splicing does not take place as desired. If we had to avoid a splicing event, it would be much easier. Right now, we would have to look for *trans*-factors that cause problems when they are overexpressed. So far, we have only identified *trans*-factors that cause problems when they are depleted. If we also found a *trans*-factor that is problematic once its concentration is increased, we could block the corresponding *cis*-element on the RNA via SSOs to prevent their effect and sustain normal splicing. In this scenario, another option could also be to reduce the problematic *trans*-factor by corrupting its own splicing via SSOs.

Another way to modify splicing focuses on splicing factors directly. Small molecules of either synthetic or natural origin can be used to target specific splicing factors. They are mostly not able to affect a specific splicing event in a gene of choice, but can be used to disrupt splicing networks. This may be beneficial in diseases that are metabolically active (e.g. cancer) to affect cellular homeostasis (Havens *et al.*, 2013; Montes *et al.*, 2019). Of course, this method is only possible if it is known that presence or overexpression of a particular *trans*-regulator is the main problem - which is not the case so far in our *CD19* splicing event.

Last but not least, genome editing using CRISPR-Cas9 is a sophisticated method to change splicing. This can correct certain mutations in the DNA that are crucial for erroneous splicing in a pathological gene or in a disrupted splicing factor. In the meantime, however, attempts have also been made to edit RNA directly, e.g. in the case of microsatellite repeat expansions, which can cause many diseases when they become part of the RNA (Batra *et al.*, 2017). CRISPR-Cas9 could also be useful for the studied CART-19 event. Since we have found many mutations through our mutagenesis screen that cause severe splicing changes and sometimes even complete splicing into a corrupt isoform, deep-sequencing at the *CD19* position could be used to detect dangerous mutations that are present only in subpopulations of the tumor. This also applies to mutations that are deleterious for *CD19*

expression irrespective of splicing (Orlando *et al.*, 2018; Zhang *et al.*, 2020). As mentioned earlier, these subpopulations are at risk of being positively selected by therapy. So, these mutations could be eradicated by CRISPR approaches prior to therapy or during treatment, before they have a chance to proliferate. Of course, targeting exactly the affected subpopulation and not omitting a single cell can be a major challenge. Also, off-targets, especially for the remaining healthy blood cells (or other cells) in the body, are not insignificant.

3.6 Further ways to overcome the loss of CD19 in CAR-T cell therapy

In addition to trying to improve B cell recognition by modifying and targeting only *CD19*, one can also begin to improve chimeric antigen receptors and increase their target range. Transformations of the CAR structure can take place at the ectodomain, transmembrane domain, and endodomain (Huang *et al.*, 2020).

The ectodomain is responsible for target specificity. It is an art to find a suitable target, because it must be very specific for the tumor or at least for the cell type affected by the tumor. Otherwise, unwanted binding reactions on-target but off-tumor will occur. B cells have been a very successful target for immunotherapy because they possess unique surface markers. Apart from CD19, they also express CD20 and CD22 on their surface, also in most cases of B-ALL (Raponi *et al.*, 2011). These two markers have been introduced as additional targets for CAR-T cell therapy as post-relapse $CD19^-$ patients were usually still positive for CD20 and CD22.

In case of CD22, Fry *et al.* (2018) generated a new CD22-targeted CAR. They demonstrated CD22-targeted CAR-T cells to have potent anti-leukemic activity without any evidence of off-targets. The success rate was similarly high as in patients treated with CART-19. Unfortunately, even with CD22-targeted therapies, decrease in antigen density or even loss in the tumor population is a major concern. Therefore, bispecific CARs are currently investigated in detail. In i*n vitro* as well as *in vivo* mice studies, these constructs have already shown activity against $CD19^+/CD22^+$, $CD19^+/CD22^-$ and $CD19^-/CD22^+$ B-ALL cells (Zanetti *et al.*, 2019). The first clinical trial showed inconsistent results. Unfortunately, even with bispecific CARs, relapse due to antigen-mediated escape effects of two antigens cannot be excluded, as shown by one out of six patients (Dai *et al.*, 2020).

Another reason to combine therapies initially rather than treating patients with one therapy at a time is that further lines of therapy are always associated with a greater risk to the patient. Therefore, Rennert *et al.* (2019) have generated monomeric CD19-anti-CD20 bridging proteins. In theory, these bridging proteins can be applied when residual tumor cells $CD19^-/CD20^+$ are detected in CART-19-treated patients. The remaining CAR-T cells in the body will then bind these proteins, be restimulated, and thereby switch their specificity towards CD20. The first human clinical trial enrolls relapsed $CD19^-$ patients that still have CAR-T cells in circulation. In addition to the approach of using bridging proteins, attempts are also being made to use CD20 antigen in bispecific CAR-T cells targeting CD19 and CD20, similar to those targeting CD19 and CD22 (Martyniszyn *et al.*, 2017). These bispecific CARs are also already in clinical trials, having shown high response rates in initial studies.

Aside from using additional antigens as targets, one method could be to identify other CD19 antibodies for use with the CAR variable single-chain antibody. Zhang *et al.* (2020) tested the 21D4-antibody in addition to the common FMC63 that is used in commercially available CAR-T cells therapies. The relapsed patient in this study escaped therapy using FMC63. A point mutation in *CD19* exon 3 led to tumor escape after initial remission. However, reinduction of CAR-T cells expressing 21D4-CARs resulted in further remission, indicating that these CAR-T cells recognized the mutant CD19 protein of that tumor cell subpopulation. In further experiments, they even showed that 21D4 recognized CD19 protein isoforms in which either exon 1, 2, or 3 was deleted at DNA level. This was not the case for FMC63-CARs. These results strongly suggest that CD19-directed CARs need to be tested for their ability to recognize mutant CD19 variants as well.

4 Conclusion and Outlook

In recent years, aberrant splicing has emerged as one of the newest hallmarks of cancer. Unfortunately, most of the molecular mechanisms that lead to this are not well understood. Aberrant splicing isoforms are nowadays not only associated with tumorigenesis or progression but can also interfere with cancer therapy directed against specific markers such as CD19.

This thesis not only gives insights into basic splicing regulatory networks using *MSTR1R* and also *CD19* as examples, but also draws attention to methodological challenges in measuring splicing variants. We have shown this in great detail using the putative splice isoform "*CD19*ex2part" and even found further similar examples in various datasets. These findings can be seminal for others who stumble upon such cases, as we also suggest strategies for refining isoform annotation.

Furthermore, considering CART-19 escape mechanisms, the results provide a comprehensive source of mutations associated with highly aberrant splicing of the *CD19* target molecule that are likely to lead to CART-19 relapse when present in patients. These mutations may represent the beginning of a list of prognostic markers predictive of CART-19 treatment success.

Abbreviations

ALL	Acute lymphoblastic leukemia
allo-SCT	Allogeneic stem cell transplantation
ASO	Antisense oligonucleotide
B-ALL	B cell acute lymphoblastic leukemia
BCR	B cell receptor
BiTE	Bispecific T cell engager molecules
bp	Base pair
CAR	Chimeric antigen receptor
CART-19	CAR-T cell therapy targeted against CD19
CD	Cluster of differentiation
cDNA	Complementary DNA
cDNA-seq	cDNA sequencing
CRISPR	Clustered regularly interspaced palindromic repeats
DNA	Desoxyribonucleic acid
dNTP	Deoxy nucleosid triphosphate
dRNA-seq	Direct RNA sequencing
eGFP	Enhanced green-fluorescent protein
ESE	Exonic splicing enhancer
ESS	Exonic splicing silencer
Fc	Fragment crystallizable region
FDA	U.S. Food and Drug Administration
gDNA	Genomic DNA
HLA	Human leukocyte antigen
hnRNP	Heterogeneous nuclear ribonucleoprotein
iCLIP	Individual-nucleotide resolution UV crosslinking and immunoprecipitation
IRE1	Inositol-requiring enzyme
ISE	Intronic splicing enhancer
ISS	Intronic splicing silencer
kD	Kilodalton
KD	Knockdown
mCherry	Family member of monomeric red-fluorescent proteins
MHC	Major histocompatibility complex
mRNA	Messenger RNA
MST1R	Macrophage-stimulating protein receptor
NGS	Next-generation sequencing
NMD	Nonsense-mediated decay
nt	Nucleotide
ONT	Oxford Nanopore Technologies
PCR	Polymerase chain reaction
pre-mRNA	precursor messenger RNA
PTBP1	Polypyrimidine tract binding protein 1
RBP	RNA-binding protein
RNA	Ribonucleid acid
RON	Recepteur d'origine nantais

RT-PCR	Reverse-transcription polymerase chain reaction
RT-qPCR	Reverse-transcription quantitative polymerase chain reaction
shRNA	Small hairpin RNA
SMRT-seq	Single-molecule real-time sequencing
SNP	Single-nucleotide polymorphism
snRNP	Small nuclear ribonucleoprotein
SR protein	Serine- and arginine-rich protein
SRSF3	Serine and arginine rich splicing factor 3
SSO	Splice-switching oligonucleotides
TCGA	The Cancer Genome Atlas
tRNA	Transfer RNA
UPR	Unfolded protein response
wt	Wild type

References

Adamson, S. I., Zhan, L., & Graveley, B. R. (2017). High-Throughput Identification of Genetic Variation Impact on pre-mRNA Splicing Efficiency. *bioRxiv,* 191122.

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., ... and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biology*, *12*(2), 1-14.

Aldoss, I., Forman, S. J., and Pullarkat, V. (2019). Acute lymphoblastic leukemia in the older adult. *Journal of Oncology Practice*, *15*(2), 67–75.

Andersen, M. H., Schrama, D., thor Straten, P., & Becker, J. C. (2006). Cytotoxic T cells. *Journal of Investigative Dermatology,* 126(1), 32-41.

Asnani, M., Hayer, K. E., Naqvi, A. S., Zheng, S., Yang, S. Y., Oldridge, D., ... & Thomas-Tikhonenko, A. (2020). Retention of CD19 intron 2 contributes to CART-19 resistance in leukemias with subclonal frameshift mutations in CD19. *Leukemia, 34*(4), 1202-1207.

Baeza-Centurion, P., Miñana, B., Schmiedel, J. M., Valcárcel, J., and Lehner, B. (2019). Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. *Cell*, *176*(3), 549-563.

Bagashev, A., Sotillo, E., Tang, C. H. A., Black, K. L., Perazzelli, J., Seeholzer, S. H., ... & Thomas-Tikhonenko, A. (2018). CD19 alterations emerging after CD19-directed immunotherapy cause retention of the misfolded protein in the endoplasmic reticulum. *Molecular and Cellular Biology, 38*(21), e00383-18.

Baralle, D., & Baralle, M. (2005). Splicing in action: assessing disease causing sequence changes. *Journal of medical genetics*, 42(10), 737-748.

Baralle, F. E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. *Nature Reviews. Molecular Cell Biology*, *18*(7), 437–451.

Baralle, M., Baralle, D., De Conti, L., Mattocks, C., Whittaker, J., Knezevich, A., and Baralle, F. E. (2003). Identification of a mutation that perturbs NF1 agene splicing using genomic DNA samples and a minigene assay. *Journal of medical genetics*, *40*(3), 220-222.

Batra, R., Nelles, D. A., Pirie, E., Blue, S. M., Marina, R. J., Wang, H., ... and Yeo, G. W. (2017). Elimination of toxic microsatellite repeat expansion RNA by RNA-targeting Cas9. *Cell,* 170(5), 899-912.

Benita, Y., Oosting, R. S., Lok, M. C., Wise, M. J., and Humphery-Smith, I. (2003). Regionalized GC content of template DNA as a predictor of PCR success. *Nucleic acids research, 31*(16), e99-e99.

Bennett, C. F., and Swayze, E. E. (2010). RNA targeting therapeutics: molecular mechanisms of antisense oligonucleotides as a therapeutic platform. *Annual review of pharmacology and toxicology*, 50, 259-293.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., ... and Roe, P. M. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature, 456*(7218), 53-59.

Bertrand, F. E., Vogtenhuber, C., Shah, N., and Lebien, T. W. (2001). Pro-B-cell to pre-B-cell development in B-lineage acute lymphoblastic leukemia expressing the MLL/AF4 fusion protein. *Blood*, *98*(12), 3398–3405.

Bhoj, V. G., Arhontoulis, D., Wertheim, G., Capobianchi, J., Callahan, C. A., Ellebrecht, C. T., ... & Milone, M. C. (2016). Persistence of long-lived plasma cells and humoral immunity in individuals responding to CD19-directed CAR T-cell therapy. *Blood, The Journal of the American Society of Hematology, 128*(3), 360-370.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., ... and Schloss, J. A. (2010). The potential and challenges of nanopore sequencing. *Nanoscience and technology: A collection of reviews from Nature Journals*, 261-268.

Bredel, M., Ferrarese, R., Thudi, N. K., Puliyappadamba, V. K., Bug, E., Trummell, H., ... and Carro, M. (2015). Targeting PTBP1 as a therapeutic strategy to reverse lineage-specific splicing of ANXA7 and ensuing EGFR activation in glioblastoma.

Braun, S., Enculescu, M., Setty, S. T., Cortés-López, M., de Almeida, B. P., Sutandy, F. X., ... and Zarnack, K. (2018). Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. *Nature communications*, *9*(1), 1-18.

Buccitelli, C., and Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10), 630-644.

Busch, A., and Hertel, K. J. (2012). Evolution of SR protein and hnRNP splicing regulatory factors. Wiley Interdisciplinary Reviews: *RNA*, 3(1), 1-12.

Conboy, J. G. (2021). Unannotated splicing regulatory elements in deep intron space. *Wiley Interdisciplinary Reviews: RNA, 12*(5), e1656.

Coolidge, C. J., Seely, R. J., & Patton, J. G. (1997). Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic acids research, 25*(4), 888-896.

Cooper, T. A. (2005). Use of minigene systems to dissect alternative splicing elements. *Methods*, *37*(4), 331-340.

Copelan, E. A. (2006). Hematopoietic stem-cell transplantation. *New England Journal of Medicine*, *354*(17), 1813-1826.

Dai, H., Wu, Z., Jia, H., Tong, C., Guo, Y., Ti, D., ... and Han, W. (2020). Bispecific CAR-T cells targeting both CD19 and CD22 for therapy of adults with relapsed or refractory B cell acute lymphoblastic leukemia. *Journal of hematology and oncology*, 13(1), 1-10.

De Conti, L., Baralle, M., & Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. *Wiley Interdisciplinary Reviews*: RNA, 4(1), 49-60.

Domingues, R. G., Lago-Baldaia, I., Pereira-Castro, I., Fachini, J. M., Oliveira, L., Drpic, D., ... & Moreira, A. (2016). CD5 expression is regulated during human T-cell activation by alternative polyadenylation, PTBP1, and miR-204. *European journal of immunology*, 46(6), 1490-1503.

Dvinge, H., Kim, E., Abdel-Wahab, O., & Bradley, R. K. (2016). RNA splicing factors as oncoproteins and tumour suppressors. *Nature Reviews Cancer, 16*(7), 413-430.

Escobar-Hoyos, L. F., Penson, A., Kannan, R., Cho, H., Pan, C. H., Singh, R. K., ... & Leach, S. D. (2020). Altered RNA splicing by mutant p53 activates oncogenic RAS signaling in pancreatic cancer. *Cancer cell, 38*(2), 198-211.

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*, 19(1), 96-102.

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., ... & Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human molecular genetics, 23*(22), 5866-5878.

Fischer, J., Paret, C., El Malki, K., Alt, F., Wingerter, A., Neu, M. A., ... and Faber, J. (2017). CD19 isoforms enabling resistance to CART-19 immunotherapy are expressed in B-ALL patients at initial diagnosis. *Journal of Immunotherapy (Hagerstown, Md.: 1997)*, *40*(5), 187.

Fortelny, N., Overall, C. M., Pavlidis, P., and Freue, G. V. C. (2017). Can we predict protein from mRNA levels?. *Nature*, *547*(7664), E19-E20.

Fraile-Bethencourt, E., Valenzuela-Palomo, A., Díez-Gómez, B., Acedo, A., and Velasco, E. A. (2018). Identification of eight spliceogenic variants in BRCA2 Exon 16 by minigene assays. *Frontiers in Genetics, 9*, 188.

Fry, T. J., Shah, N. N., Orentas, R. J., Stetler-Stevenson, M., Yuan, C. M., Ramakrishna, S., ... & Mackall, C. L. (2018). CD22-targeted CAR T cells induce remission in B-ALL that is naive or resistant to CD19-targeted CAR immunotherapy. *Nature medicine, 24*(1), 20-28.

Fu, X. D., and Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, *15*(10), 689-701.

Gao, K., Masuda, A., Matsuura, T., & Ohno, K. (2008). Human branch point consensus sequence is yUnAy. *Nucleic acids research, 36*(7), 2257-2267.

Gardner, R. A., Finney, O., Annesley, C., Brakke, H., Summers, C., Leger, K., ... and Jensen, M. C. (2017). Intent-to-treat leukemia remission by CD19 CAR T cells of defined formulation and dose in children and young adults. *Blood, The Journal of the American Society of Hematology*, *129*(25), 3322-3331.

Gassas, A., Ashraf, K., Zaidman, I., Ali, M., Krueger, J., Doyle, J., ... and Leucht, S. (2015). Hematopoietic stem cell transplantation in infants. *Pediatric blood and cancer*, *62*(3), 517-521.

Gökbuget, N., Dombret, H., Bonifacio, M., Reichle, A., Graux, C., Faul, C., ... Bargou, R. C. (2018). Blinatumomab for minimal residual disease in adults with B-cell precursor acute lymphoblastic leukemia. *Blood*, *131*(14), 1522–1531.

Green, M. R. (1986). Pre-mRNA splicing. *Annual review of genetics*, *20*(1), 671-708.

Hall, S. L., & Padgett, R. A. (1996). Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. *Science, 271*(5256), 1716-1718.

Hastings, M. L., & Krainer, A. R. (2001). Pre-mRNA splicing in the new millennium. *Current opinion in cell biology*, 13(3), 302-309.

Havens, M. A., and Hastings, M. L. (2016). Splice-switching antisense oligonucleotides as therapeutic drugs. Nucleic acids research, 44(14), 6549-6563.

Havens, M. A., Duelli, D. M., and Hastings, M. L. (2013). Targeting RNA splicing for disease therapy. Wiley Interdisciplinary Reviews: *RNA,* 4(3), 247-266.

Hay, K. A., Gauthier, J., Hirayama, A. V., Voutsinas, J. M., Wu, Q., Li, D., ... and Turtle, C. J. (2019). Factors associated with durable EFS in adult B-cell ALL patients achieving MRD-negative CR after CD19 CAR T-cell therapy. *Blood, The Journal of the American Society of Hematology*, *133*(15), 1652-1663.

He, S. L., and Green, R. (2013). Northern blotting. In *Methods in enzymology* (Vol. 530, pp. 75-87). Academic Press.

Holmila, R., Fouquet, C., Cadranel, J., Zalcman, G., and Soussi, T. (2003). Splice mutations in the p53 gene: case report and review of the literature. *Human mutation*, *21*(1), 101-102.

Huang, R., Li, X., He, Y., Zhu, W., Gao, L., Liu, Y., ... and Zhang, X. (2020). Recent advances in CAR-T cell engineering. *Journal of Hematology and Oncology*, 13(1), 1-19.

Illumina. For all you seq. Illumina https://emea. illumina.com/techniques/sequencing/ngs- library-prep/ library- prep-methods.html (2014). A tour de force that includes a graphical abstract, a brief description and primary references for most sequencing methods.

Inaba, H., and Mullighan, C. G. (2020). Pediatric acute lymphoblastic leukemia. *Haematologica*, *105*(11), 2524–2539.

Jabbour, E., O'Brien, S., Konopleva, M., and Kantarjian, H. (2015). New insights into the pathophysiology and therapy of adult acute lymphoblastic leukemia. *Cancer*, *121*(15), 2517–2528.

Jaganathan, K., Panagiotopoulou, S. K., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., ... & Farh, K. K. H. (2019). Predicting splicing from primary sequence with deep learning. *Cell, 176*(3), 535-548.

Jiang, X. R., Song, A., Bergelson, S., Arroll, T., Parekh, B., May, K., ... and Schenerman, M. (2011). Advances in the assessment and control of the effector functions of therapeutic antibodies. *Nature reviews Drug discovery*, *10*(2), 101-111.

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*(1), 1-8.

June, C. H., & Sadelain, M. (2018). Chimeric antigen receptor therapy. *New England Journal of Medicine*, 379(1), 64-73.

Ke, S., Anquetil, V., Zamalloa, J. R., Maity, A., Yang, A., Arias, M. A., ... and Chasin, L. A. (2018). Saturation mutagenesis reveals manifold determinants of exon definition. *Genome research*, *28*(1), 11-24.

Kershaw, M. H., Westwood, J. A., & Darcy, P. K. (2013). Gene-engineered T cells for cancer therapy. *Nature Reviews Cancer*, 13(8), 525-541.

King, H. A., and Gerber, A. P. (2016). Translatome profiling: methods for genome-scale analysis of mRNA translation. *Briefings in functional genomics*, *15*(1), 22-31.

Kornblihtt, A. R., Vibe-Pedersen, K., and Baralle, F. E. (1984). Human fibronectin: molecular cloning evidence for two mRNA species differing by an internal segment coding for a structural domain. *The EMBO journal*, *3*(1), 221-226.

Krecic, Annette M., and Maurice S. Swanson. "hnRNP complexes: composition, structure, and function." *Current opinion in cell biology* 11.3 (1999): 363-371.

Kumar, K. R., Cowley, M. J., & Davis, R. L. (2019, October). Next-generation sequencing and emerging technologies. *In Seminars in thrombosis and hemostasis* (Vol. 45, No. 07, pp. 661-673). Thieme Medical Publishers.

Kurzrock, R., Gutterman, J. U., and Talpaz, M. (1988). The molecular genetics of Philadelphia chromosome–positive leukemias. *New England Journal of Medicine*, *319*(15), 990-998.

Kwok, C. K., Tang, Y., Assmann, S. M., and Bevilacqua, P. C. (2015). The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends in biochemical sciences*, *40*(4), 221-232.

Ladomery, M. (2013). Aberrant alternative splicing is another hallmark of cancer. *International journal of cell biology*, 2013.

LeFave, C. V., Squatrito, M., Vorlova, S., Rocco, G. L., Brennan, C. W., Holland, E. C., ... and Cartegni, L. (2011). Splicing factor hnRNPH drives an oncogenic splicing switch in gliomas. *The EMBO journal*, *30*(19), 4084-4097.

Lewis, C. J., Pan, T., and Kalsotra, A. (2017). RNA modifications and structures cooperate to guide RNA–protein interactions. *Nature reviews Molecular cell biology*, *18*(3), 202-210.

Li, X., and Chen, W. (2019). Mechanisms of failure of chimeric antigen receptor T-cell therapy. *Current Opinion in Hematology*, *26*(6), 427.

Liu, D., Zhao, J., and Song, Y. (2019). Engineering switchable and programmable universal CARs for CAR T therapy. *Journal of Hematology and Oncology*, 12(1), 1-9.

Locatelli, F., Zugmaier, G., Mergen, N., Bader, P., Jeha, S., Schlegel, P. G., ... and Chen-Santel, C. (2020). Blinatumomab in children with relapsed or refractory B-precursor acute lymphoblastic leukemia (R/R-ALL): final results of 110 patients treated in an expanded access study (RIALTO). *Blood*, *136*, 24-25.

Lodish, H., Berk, A., Kaiser, C., Krieger, M., Bretscher, A., Matsudaira, P., ... and Zipursky, S., L.(2004). *Molecular Cell Biology. 5th ed*. ISBN 13: 9780716743668

Loffler, A., Kufer, P., Lutterbüse, R., Zettl, F., Daniel, P. T., Schwenkenbecher, J. M., ... and Bargou, R. C. (2000). A recombinant bispecific single-chain antibody, CD19× CD3, induces rapid and high lymphoma-directed cytotoxicity by unstimulated T lymphocytes. Blood, *The Journal of the American Society of Hematology*, 95(6), 2098-2103.

Long, J. C., and Caceres, J. F. (2009). The SR protein family of splicing factors: master regulators of gene expression. *Biochemical Journal*, *417*(1), 15-27.

Lugo, T. G., Pendergast, A. M., Muller, A. J., & Witte, O. N. (1990). Tyrosine kinase activity and transformation potency of bcr-abl oncogene products. *Science, 247*(4946), 1079-1082.

Lutterbuese, R., Raum, T., Kischel, R., Hoffmann, P., Mangold, S., Rattel, B., ... and Kufer, P. (2010). T cell-engaging BiTE antibodies specific for EGFR potently eliminate KRAS-and BRAF-mutated colorectal cancer cells. *Proceedings of the National Academy of Sciences*, *107*(28), 12605-12610.

Maquat, L. E. (2004). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nature reviews Molecular cell biology,* 5(2), 89-99.

Marinov, G. K. (2017). On the design and prospects of direct RNA sequencing. *Briefings in functional genomics*, *16*(6), 326-335.

Martinelli, G., Boissel, N., Chevallier, P., Ottmann, O., Gökbuget, N., Topp, M. S., ... Stein, A. (2017). Complete hematologic and molecular response in adult patients with relapsed/refractory philadelphia chromosome-positive B-precursor acute lymphoblastic leukemia following treatment with blinatumomab: Results from a phase II, single-arm, multicenter study. *Journal of Clinical Oncology*, *35*(16), 1795–1802.

Martyniszyn, A., Krahl, A. C., Andre, M. C., Hombach, A. A., and Abken, H. (2017). CD20-CD19 bispecific CAR T cells for the treatment of B-cell malignancies. *Human gene therapy*, 28(12), 1147-1157.

Matera, A. G., & Wang, Z. (2014). A day in the life of the spliceosome. *Nature reviews Molecular cell biology, 15*(2), 108-121.

Maude, S. L., Laetsch, T. W., Buechner, J., Rives, S., Boyer, M., Bittencourt, H., ... & Grupp, S. A. (2018). Tisagenlecleucel in children and young adults with B-cell lymphoblastic leukemia. *New England Journal of Medicine*, 378(5), 439-448.

Maude, S. L., Teachey, D. T., Rheingold, S. R., Shaw, P. A., Aplenc, R., Barrett, D. M., ... and Grupp, S. A. (2016). Sustained remissions with CD19-specific chimeric antigen receptor (CAR)-modified T cells in children with relapsed/refractory ALL.

McGreevy, J. W., Hakim, C. H., McIntosh, M. A., and Duan, D. (2015). Animal models of Duchenne muscular dystrophy: from basic mechanisms to gene therapy. *Disease models and mechanisms*, 8(3), 195-213.

McInerney, P., Adams, P., and Hadi, M. Z. (2014). Error rate comparison during polymerase chain reaction by DNA polymerase. *Molecular biology international*, *2014*.

McManus, C. J., and Graveley, B. R. (2011). RNA structure and the mechanisms of alternative splicing. *Current opinion in genetics and development*, *21*(4), 373-379.

Mikl, M., Pilpel, Y., & Segal, E. (2020). High-throughput interrogation of programmed ribosomal frameshifting in human cells. *Nature Communications, 11*(1), 3061.

Montes, M., Sanford, B. L., Comiskey, D. F., and Chandler, D. S. (2019). RNA splicing and disease: animal models to therapies. *Trends in Genetics*, 35(1), 68-87.

Monzón-Casanova, E., Screen, M., Díaz-Muñoz, M. D., Coulson, R. M., Bell, S. E., Lamers, G., ... and Turner, M. (2018). The RNA-binding protein PTBP1 is necessary for B cell selection in germinal centers. *Nature immunology*, 19(3), 267-278.

Moore, M. J., Query, C. C., and Sharp, P. A. (1993). Splicing of precursors to mRNAs by the spliceosome. *Cold Spring Harbor Monograph Series*, *24*, 303-303.

Oltean, S., & Bates, D. O. (2014). Hallmarks of alternative splicing in cancer. *Oncogene, 33*(46), 5311-5318.

Nadler, L. M., Anderson, K. C., Marti, G., Bates, M., Park, E., Daley, J. F., and Schlossman, S. F. (1983). B4, a human B lymphocyte-associated antigen expressed on normal, mitogen-activated, and malignant B lymphocytes. *The Journal of Immunology*, *131*(1), 244-250.

Nagy, E., and Maquat, L. E. (1998). A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends in biochemical sciences*, 23(6), 198-199.

Narayanan, S., and Shami, P. J. (2012). Treatment of acute lymphoblastic leukemia in adults. *Critical Reviews in Oncology/Hematology*, *81*(1), 94–102.

Naro, C., Cesari, E., and Sette, C. (2021). Splicing regulation in brain and testis: common themes for highly specialized organs. *Cell Cycle*, *20*(5-6), 480-489.

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*(1), 1-11.

Ning, B. T., Tang, Y. M., Chen, Y. H., Shen, H. Q., and Qian, B. Q. (2005). Comparison between CD19 and CD20 expression patterns on acute leukemic cells. *Zhongguo shi yan xue ye xue za zhi*, *13*(6), 943-947.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., ... and Hayashizaki, Y. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, *420*(6915), 563-573.

Orlando, E. J., Han, X., Tribouley, C., Wood, P. A., Leary, R. J., Riester, M., ... and Winckler, W. (2018). Genetic mechanisms of target antigen loss in CAR19 therapy of acute lymphoblastic leukemia. *Nature medicine*, 24(10), 1504-1506.

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2), 87-98.

Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The expanding landscape of alternative splicing variation in human populations. *The American Journal of Human Genetics*, *102*(1), 11-26.

Paul, S., Kantarjian, H., and Jabbour, E. J. (2016). Adult Acute Lymphoblastic Leukemia. *Mayo Clinic Proceedings*, *91*, 1645–1666.

Piovesan, A., Caracausi, M., Antonaros, F., Pelleri, M. C., and Vitale, L. (2016). GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. *Database*, *2016*.

Porensky, P. N., Mitrpant, C., McGovern, V. L., Bevan, A. K., Foust, K. D., Kaspar, B. K., ... and Burghes, A. H. (2012). A single administration of morpholino antisense oligomer rescues spinal muscular atrophy in mouse. *Human molecular genetics*, 21(7), 1625-1638.

Pyatnitskiy, M. A., Arzumanian, V. A., Radko, S. P., Ptitsyn, K. G., Vakhrushev, I. V., Poverennaya, E. V., and Ponomarenko, E. A. (2021). Oxford Nanopore MinION Direct RNA-Seq for Systems Biology. *Biology*, *10*(11), 1131.

Qian, X., Wang, J., Wang, M., Igelman, A. D., Jones, K. D., Li, Y., ... and Chen, R. (2021). Identification of deep-intronic splice mutations in a large cohort of patients with inherited retinal diseases. *Frontiers in genetics*, *12*.

Qin, Y., Yao, J., Wu, D. C., Nottingham, R. M., Mohr, S., Hunicke-Smith, S., and Lambowitz, A. M. (2016). High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. *Rna,* 22(1), 111-128.

Rabilloud, T., Potier, D., Pankaew, S., Nozais, M., Loosveld, M., and Payet-Bornet, D. (2021). Single-cell profiling identifies pre-existing CD19-negative subclones in a B-ALL patient with CD19-negative relapse after CAR-T therapy. *Nature communications*, *12*(1), 1-7.

Rambaldi, A., Huguet, F., Zak, P., Cannell, P., Tran, Q., Franklin, J., and Topp, M. S. (2020). Blinatumomab consolidation and maintenance therapy in adults with relapsed/refractory B-precursor acute lymphoblastic leukemia. *Blood Advances*, *4*(7), 1518–1525.

Rang, F. J., Kloosterman, W. P., and de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome biology*, *19*(1), 1-11.

Raponi, M., Kralovicova, J., Copson, E., Divina, P., Eccles, D., Johnson, P., ... and Vorechovsky, I. (2011). Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6. *Human mutation*, *32*(4), 436-444.

Raponi, S., Stefania De Propris, M., Intoppa, S., Laura Milani, M., Vitale, A., Elia, L., ... and Guarini, A. (2011). Flow cytometric study of potential target antigens (CD19, CD20, CD22, CD33) for antibody-based immunotherapy in acute lymphoblastic leukemia: analysis of 552 cases. *Leukemia and lymphoma*, 52(6), 1098-1107.

Rennert, P., Su, L., Dufort, F., Birt, A., Sanford, T., Wu, L., ... and Lobb, R. (2019). A novel CD19-anti-CD20 bridging protein prevents and reverses CD19-negative relapse from CAR19 T cell treatment in vivo. *Blood,* 134, 252.

Robinson, F., and Smith, C. W. (2006). A splicing repressor domain in polypyrimidine tract-binding protein. *Journal of Biological Chemistry*, 281(2), 800-806.

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*(3), 698-711.

Ruella, M., Barrett, D. M., Kenderian, S. S., Shestova, O., Hofmann, T. J., Perazzelli, J., ... & Gill, S. (2016). Dual CD19 and CD123 targeting prevents antigen-loss relapses after CD19-directed immunotherapies. *The Journal of clinical investigation, 126*(10), 3814-3826.

Ruf, P., and Lindhofer, H. (2001). Induction of a long-lasting antitumor immunity by a trifunctional bispecific antibody. *Blood, The Journal of the American Society of Hematology*, *98*(8), 2526-2534.

Saha, A., Kim, Y., Gewirtz, A. D., Jo, B., Gao, C., McDowell, I. C., ... and Wu, F. (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome research*, *27*(11), 1843-1858.

Sasanuma, H., Ozawa, M., and Yoshida, N. (2019). RNA-binding protein Ptbp1 is essential for BCR-mediated antibody production. *International immunology*, 31(3), 157-166.

Schaal, H. (2013). Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *Rna*, *19*(1), 96-102.

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*, *102*(14), 5002-5007.

Schulz, L., Torres-Diz, M., Cortés-López, M., Hayer, K. E., Asnani, M., Tasian, S. K., ... and Thomas-Tikhonenko, A. (2021). Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts. *Genome biology*, 22(1), 1-12.

Singh, A. K., and McGuirk, J. P. (2016). Allogeneic stem cell transplantation: a historical and scientific overview. *Cancer research*, *76*(22), 6445-6451.

Singh, G., and Cooper, T. A. (2006). Minigene reporter for identification and analysis of cis elements and trans factors affecting pre-mRNA splicing. *Biotechniques, 41*(2), 177-181.

Soukarieh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frébourg, T., ... & Martins, A. (2016). Exonic splicing mutations are more prevalent than currently estimated and can be predicted by using in silico tools. *PLoS genetics, 12*(1), e1005756.

Sotillo, E., Barrett, D. M., Black, K. L., Bagashev, A., Oldridge, D., Wu, G., ... and Thomas-Tikhonenko, A. (2015). Convergence of acquired mutations and alternative splicing of CD19 enables resistance to CART-19 immunotherapy. *Cancer discovery*, 5(12), 1282-1295.

Souček, P., Réblová, K., Kramárek, M., Radová, L., Grymová, T., Hujová, P., ... and Freiberger, T. (2019). High-throughput analysis revealed mutations' diverging effects on SMN1 exon 7 splicing. *RNA biology*, *16*(10), 1364-1376.

Spector, D. L., and Lamond, A. I. (2011). Nuclear speckles. *Cold Spring Harbor perspectives in biology*, *3*(2), a000646.

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., ... and Soreq, H. (2005). Function of alternative splicing. *Gene*, *344*, 1-20.

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics 2019 20:11*, *20*(11), 631–656.

Steiner, M., and Neri, D. (2011). Antibody-radionuclide conjugates for cancer therapy: historical considerations and new trends. *Clinical Cancer Research*, *17*(20), 6406-6416.

Stieglmaier, J., Benjamin, J., and Nagorsen, D. (2015). Utilizing the BiTE (bispecific T-cell engager) platform for immunotherapy of cancer. *Expert opinion on biological therapy*, *15*(8), 1093-1099.

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*(19), 2413-2427.

Tardaguila, M., De La Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., ... & Conesa, A. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome research, 28*(3), 396-411.

Terwilliger, T., and Abdul-Hay, M. (2017). Acute lymphoblastic leukemia: a comprehensive review and 2017 update. *Blood Cancer Journal*, *7*, 577.

Thomas, X., Boiron, J. M., Huguet, F., Dombret, H., Bradstock, K., Vey, N., ... and Fiere, D. (2004). Outcome of treatment in adults with acute lymphoblastic leukemia: analysis of the LALA-94 tri*al. Journal of clinical oncology*, *22*(20), 4075-4086.

Uprety, D., and Adjei, A. A. (2020). KRAS: From undruggable to a druggable Cancer Target. *Cancer Treatment Reviews*, *89*.

Venables, J. P. (2004). Aberrant and alternative splicing in cancer. *Cancer research, 64*(21), 7647-7654.

Voglino, G., Castello, S., Silengo, L., Stefanuto, G., Friard, O., Ferrara, G., & Fessia, L. (1997). An intronic deletion in TP53 gene causes exon 6 skipping in breast cancer. *European Journal of Cancer, 33*(9), 1479-1483.

Wang, K., Wei, G., and Liu, D. (2012). CD19: a biomarker for B cell development, lymphoma diagnosis and therapy. *Experimental hematology and oncology*, *1*(1), 1-7.

Wang, Z., and Burge, C. B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *Rna*, *14*(5), 802-813.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., ... and Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, *509*(7502), 582-587.

Will, C. L., and Lührmann, R. (2011). Spliceosome Structure and Function. *Cold Spring Harbor Perspectives in Biology*, *3*(7), a003707.

Wong, K. H., Jin, Y., and Moqtaderi, Z. (2013). Multiplex Illumina sequencing using DNA barcoding. *Current Protocols in Molecular Biology*, *101*(1), 7-11.

Xu, L., and Seki, M. (2020). Recent advances in the detection of base modifications using the Nanopore sequencer. *Journal of human genetics*, *65*(1), 25-33.

Yannakou, C. K., Came, N., Bajel, A. R., and Juneja, S. (2015). CD19 negative relapse in B-ALL treated with blinatumomab therapy: avoiding the trap. *Blood*, *126*(23), 4983.

Yeo, G., & Burge, C. B. (2003, April). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. In Proceedings of the seventh annual international conference on *Research in computational molecular biology* (pp. 322-331).

Zanetti, S. R., Velazco-Hernandez, T., Gutierrez-Agüera, F., Roca-Ho, H., Sánchez-Martínez, D., Petazzi, P., ... & Menéndez, P. (2019). CD19 and CD22-directed biespecific CAR for B-cell Acute Lymphoblastic Leukemia. *Klinische Pädiatrie, 231*(03), 1.

Zhang, Z., Chen, X., Tian, Y., Li, F., Zhao, X., Liu, J., ... and Zhang, Y. (2020). Point mutation in CD19 facilitates immune escape of B cell lymphoma from CAR-T cell therapy. *Journal for immunotherapy of cancer*, 8(2).

Zhou, Z., Licklider, L. J., Gygi, S. P., and Reed, R. (2002). Comprehensive proteomic analysis of the human spliceosome. Nature, 419(6903), 182-185.

Zhu, W., Zhou, B. L., Rong, L. J., Ye, L., Xu, H. J., Zhou, Y., ... & Ren, C. P. (2020). Roles of PTBP1 in alternative splicing, glycolysis, and oncogensis. *Journal of Zhejiang University-SCIENCE B*, 21(2), 122-136.