

# Title: RNA splicing analysis in genomic medicine

**Authors:** Htoo Wai<sup>1</sup>, Andrew G. L. Douglas<sup>1,2</sup>, Diana Baralle<sup>1,2</sup>

1. Human Development and Health, Faculty of Medicine, University of Southampton, UK
2. Wessex Clinical Genetics Service, University Hospital Southampton NHS Foundation Trust, Southampton, UK

## Corresponding author:

Professor Diana Baralle, Human Development and Health, Faculty of Medicine, University of Southampton, Duthie Building, Southampton General Hospital, Tremona Road, Southampton, UK, SO16 6YD

[d.baralle@soton.ac.uk](mailto:d.baralle@soton.ac.uk)

**Tel:** +44 (0)23 8120 6162

## Abstract

High-throughput next-generation sequencing technologies have led to a rapid increase in the number of sequence variants identified in clinical practice via diagnostic genetic tests. Current bioinformatic analysis pipelines fail to take adequate account of the possible splicing effects of such variants, particularly where variants fall outwith canonical splice site sequences, and consequently the pathogenicity of such variants may often be missed. The regulation of splicing is highly complex and as a result, *in silico* prediction tools lack sufficient sensitivity and specificity for reliable use. Variants of all kinds can be linked to aberrant splicing in disease and the need for correct identification and diagnosis grows ever more crucial as novel splice-switching antisense oligonucleotide therapies start to enter clinical usage. RT-PCR provides a useful targeted assay of the splicing effects of identified variants, while minigene assays, massive parallel reporter assays and animal models can also be used for more detailed study of a particular splicing system, given enough time and resources. However, RNA-sequencing (RNA-seq) has the potential to be used as a rapid diagnostic tool in genomic medicine. By utilising data science approaches and machine learning, it may prove possible to finally understand and interpret the 'splicing code' and apply this knowledge in human disease diagnostics.

**Keywords:** Splicing, Sequence variants, RNA-sequencing, Machine learning, Clinical diagnosis

## Introduction

Aberrant splicing detection is the next gateway to diagnostic uplift in genomic medicine. The introduction of next-generation DNA sequencing technology has allowed a recent rapid expansion of alternative splicing knowledge. Up to an estimated 62% of all disease-causing point mutations are thought to affect RNA splicing (López-Bigas *et al.*, 2005). Whilst such a high figure may at first glance sound surprising or even alarming, it serves to highlight two pertinent realities facing those at the clinical 'coal face' of genomic diagnostics. First of all, it belies the underlying complexity of the splicing process itself and how much our knowledge of this mechanism and its regulation lags behind the true state of things. Secondly, it suggests that current diagnostic genetic testing of clinical samples may in fact be missing a significant proportion of potentially diagnosable cases, since RNA analysis is not a routine part of the diagnostic pipeline.

A single gene with multiple exons can produce different mRNAs using alternative splicing and in eukaryotic species splicing has been shown to be an important player for protein diversity and function (Figure 1). The consequences of such alternative splicing on a resulting protein depend upon the nature and function of the coding sequence region that is alternatively spliced and also on the resulting reading frame of the spliced mRNA. A spliced-out exon may, for example, code for a functionally important domain or a cellular localisation sequence, while an alternative spliced-in exon may incorporate a different domain that leads to alteration of a protein's function or localisation. However, should alternative splicing result in the reading frame becoming disrupted, the spliced mRNA will in most cases be subject to cellular degradation via nonsense-mediated decay (NMD) (Lloyd, 2018). This mechanism will generally lead to reduced gene expression and can therefore be a cause of haploinsufficiency.

Exon-intron boundaries have a distinctive but limited degree of sequence conservation which are recognised by the spliceosome. Introns usually have GU at their 5' end and AG at their 3' end (Breathnach *et al.*, 1978). In mammalian genomes, over 98% of splice sites utilise GU-AG as the splice donor and acceptor sites and less than 1% have GC-AG (Burset, 2000). In addition, a number of exonic and intronic sequence elements, named exonic/intronic splicing enhancers (ESEs and ISEs) and exonic/intronic splicing silencers (ESSs and ISSs), influence the final splicing outcome. These are recognised and bound by various RNA-binding proteins such as serine-arginine-rich (SR) proteins and heterogeneous ribonucleoproteins (hnRNPs) that are *trans*-acting splicing factors. The spliceosome complex itself is composed of small nuclear RNAs (snRNAs) plus various proteins (forming snRNPs) and accessory factors (Dvinge, 2018). In addition to snRNPs, long non-coding RNAs have also been reported to be involved in splicing (Romero-Barrios *et al.*, 2018).

Due to the complexity of splicing, and despite much research, its complete regulatory mechanism is still to be fully elucidated. This lack of knowledge hampers our abilities to predict the effect of a sequence variant of unknown significance on splicing. Many splicing prediction tools, programs and algorithms have been developed based on available biological research data. However, none of them has deciphered the complete 'splicing code' and none predict splicing effects with 100% accuracy. As has been reviewed elsewhere (Baralle and Baralle, 2018), to splice or not to splice is determined by not only the splice site consensus

sequences but also by exonic/intronic enhancer/silencer elements as well as cellular and tissue-specific factors.

The ultimate goal of genomic medicine in the high-throughput sequencing era is to exploit knowledge of sequencing variants among patients for genomic diagnosis and personalised therapeutics. This will be aided by learning how to interpret genomic DNA sequence variation not only in terms of protein function but also with regards to splicing function. RNA sequencing (RNA-seq) will be a useful diagnostic technique to detect these splice variants as well as for determining RNA expression levels. In this review, we will discuss splicing defects in disease, technological advances in predicting splicing and the use of RNA-seq and new bioinformatics techniques as diagnostic tools.

## Splicing in disease

It is well documented that splicing abnormalities play a major role in human diseases (Douglas and Wood, 2011). Since approximately 95% of human multi-exon genes are alternatively spliced through incompletely understood regulatory mechanisms (Pan *et al.*, 2008), almost any type of sequence variation in the genome, including both single nucleotide polymorphisms (SNPs) and copy number variants (CNVs), has the potential to affect splicing. Estimates of how often this occurs vary between different genes. Studies of neurofibromatosis type 1 (*NF1*) and ataxia-telangiectasia (*ATM*) genes, found 50% of sequence variants led to aberrant alternative splicing (Teraoka *et al.*, 1999; Ars *et al.*, 2000) and even the lowest estimates, which take into account only variants affecting splice sites, comprise a significant 15% of pathogenic mutations (Krawczak, Reiss and Cooper, 1992). However, it is also important to consider the fact that variants may affect splicing beyond the bounds of consensus splice sites. Newer studies estimate that 27% of developmental disorder-related splicing mutations are not within canonical splice sites (Lord *et al.*, 2018), providing at least part of the explanation for the thus far 'missing' mutations in genetic disease.

### *How splicing goes wrong in disease*

Owing to its inherent complexity, the splicing process can be perturbed in multiple different ways and at different levels. At the most basic level, a canonical splice site mutation may interrupt an individual splicing event within a specific gene transcript. Variants in either splice donor or acceptor sequences are commonly associated with exon skipping of their closest exon (Buratti and Baralle, 2012). For example, exon 7 of the *COL5A1* collagen gene is found to be skipped in patients with Ehlers-Danlos syndrome who have a sequence change in the splice acceptor site of intron 6 (c.925-2A>G) (Symoens *et al.*, 2011) (Figure 2A). Similarly, a splice donor sequence variant in intron 3 in the major intrinsic protein gene (*MIP*) causes exon 3 skipping found in patients with congenital cataract (Zeng *et al.*, 2013) (Figure 2C).

In addition, simple sequence variants in splice donor/acceptor sequences can cause multi-exon skipping, particularly if the regulation of the neighbouring exons is linked. For example, exon 12 and 13 are skipped in fibroblast *OXT1* transcripts with an intronic sequence variant

near the splice donor site of intron 13 (c.1245+5G>A) (Hori *et al.*, 2013) (Figure 2B). Similar double-exon skipping (exon 11 and exon 12a of *NF1*) is found with a splice donor sequence variant of *NF1* intron 12a (Fang *et al.*, 2001) (Figure 2D). However, it is well reported that whole exon skipping is not always the effect of mutations in the consensus splice donor/acceptor sequences. For example, a new cryptic splice donor site is activated within exon 45 of *DMD* when the splice donor site of intron 45 is altered, resulting in partial loss of exon 45 (Habara *et al.*, 2009) (Figure 2E). There is also evidence that variants deep within exons need to be considered with regards to their effects on the splicing process. These sequence changes may disrupt enhancer or silencer elements and thus affect the delicate balance of splice factors involved. An example of this includes a single nucleotide variant, well within exon 2 of *COL2A1* (c.196G>A) resulting in exon 2 skipping (McAlinden *et al.*, 2008) (Figure 2H).

Intronic sequence variants may cause splicing defects in a number of ways. They may, for example, create a *de novo* splice site, transforming part of an intron into a new exon (Vaz-Drago, Custódio and Carmo-Fonseca, 2017). This was first described with a deep intronic sequence variant in *NF1* intron 30 (c.293-279A>G) (Raponi, Upadhyaya and Baralle, 2006). Similar cryptic exon formation has been reported as a result of a deep intronic sequence variant within intron 12 of *CFTR* (Sanz *et al.*, 2017) (Figure 2F). In some cases, a single deep intronic sequence change produces more than one abnormal splice variant. For example, AR c.2450-118A>G in intron 6 of the androgen receptor gene produces a splice variant with a new cryptic exon and another one with intron inclusion (Känsäkoski *et al.*, 2016) (Figure 2G).

Sequence variants at the branch point or in the polypyrimidine tract (Van De Water *et al.*, 2004; Raponi, Upadhyaya and Baralle, 2006; Raponi *et al.*, 2008; Aoyama *et al.*, 2017) (Figure 2J) of an intron may also disrupt specific splice events and can cause skipping of the adjacent exon (Khan *et al.*, 2010) (Figure 2I). Similarly, mutations within important regulatory elements such as ESEs, ESSs, ISEs and ISSs can abrogate the normal splicing events they control by affecting the binding of specific splicing factors to these motifs. Such a mechanism is thought to explain why the evolutionarily duplicated gene *SMN2* is not able to fully compensate for deficiency of its paralogue *SMN1* in cases of spinal muscular atrophy (SMA), despite the two genomic sequences differing at only a handful of single nucleotide sites, including just one synonymous coding sequence change (Douglas and Wood, 2013).

As well as mutations within *cis*-acting regulatory motifs, abnormalities of their *trans*-acting RNA binding protein partners can also lead to secondary splicing defects. Splicing factors themselves may have mutations, such as in the case of *TARDBP*-related amyotrophic lateral sclerosis (ALS) or frontotemporal dementia (FTD), where depletion of TDP-43 protein has been shown to cause widespread retention of very long introns within its target transcripts (Polymenidou *et al.*, 2011). Alternatively, splicing factors may become sequestered and thereby downregulated by binding to toxic RNA species, with consequent downstream splicing abnormalities. A well-known example of this is myotonic dystrophy type 1, where transcripts containing a non-coding CUG trinucleotide expansion within the *DMPK* gene sequester the splicing factor MBNL1, with subsequent abnormal splicing of the muscle chloride channel gene *CLCN1* (Charlet-B *et al.*, 2002; Lin *et al.*, 2006). A similar sequestration

mechanism may occur in ALS/FTD caused by a hexanucleotide GGGGCC expansion in the *C9orf72* gene, which has been shown to bind multiple splicing factors including hnRNP-H and SRSF1, with detectable downstream splicing abnormalities in the central nervous system (Reddy *et al.*, 2013; Cooper-Knock *et al.*, 2014, 2015; Conlon *et al.*, 2016).

Long non-coding RNAs (lncRNAs) are increasingly being found by high-throughout RNA sequencing techniques and have been shown to be important for normal cellular functions and disease (Dey, Mueller and Dutta, 2014). Of note, lncRNAs are generally subject to the same splicing mechanisms as coding RNAs and a large number of such lncRNAs have annotated splicing events with intron/exon boundaries, frequently generating multiple splice isoforms. Indeed the non-coding exons of lncRNAs are found to be GC rich in a manner similar to coding RNA exons (Haerty and Ponting, 2015). Altered lncRNA splicing can be associated with disease, for example where the alternatively spliced isoform of lncRNA-PXN-AS1 appears to promote hepatocellular carcinoma growth (Yuan *et al.*, 2017). lncRNAs are also involved in the splicing process itself (Romero-Barrios *et al.*, 2018). One prominent role they appear to play in splicing is through the recruitment of splicing factors such as SR proteins and regulation of their phosphorylation (Tripathi *et al.*, 2010; Cooper *et al.*, 2014). In another example, the lncRNA BC200, which is upregulated in breast cancer, recruits splicing factor hnRNP A2/B1, which suppresses the alternative splicing of Bcl-x into the otherwise pro-apoptotic factor Bcl-xS, thus suggesting an oncogenic role for BC200 (Singh *et al.*, 2016). However, for the most part it remains unclear what genetic disorders are directly related to lncRNA variants.

#### *Therapeutics that manipulate splicing*

As knowledge of the splicing mechanisms involved in disease has increased, so too has our knowledge of how to therapeutically manipulate RNA splicing in a beneficial manner. Once the key sequence features governing a given splice event have been identified, it becomes possible to design and synthesise antisense oligonucleotide (ASO) compounds that are complementary to these sequences and can thus bind to their targets and interfere with their usage (Douglas and Wood, 2013; Rinaldi and Wood, 2018). Such ASO drugs are short chemically modified analogues of nucleic acid, whose chemical properties allow retention of base-pairing ability with specific target sequences but whose modifications can enhance stability, activity, cellular delivery and pharmacokinetics and dynamics.

Splice-switching ASO therapies are now starting to enter the clinical arena, perhaps best exemplified by the drug nusinersen for SMA (Finkel *et al.*, 2016). Nusinersen is a 2'-O-methoxyethyl phosphorothioate ASO that binds to an ISS sequence within pre-mRNA transcripts of *SMN2*, restoring the inclusion of exon 7 in mature mRNA. This rescue of splicing leads to generation of functional SMN protein, with a profound effect on survival and motor function in this heretofore lethal neuromuscular disease of infancy. Splice-switching ASOs have also been developed for the muscle-wasting disease Duchenne muscular dystrophy (DMD) and the drug eteplirsen has been granted clinical approval in the USA (Mendell *et al.*, 2016). Eteplirsen is a phosphorodiamidate morpholino ASO that binds to and masks an ESE sequence within exon 51 of *DMD* pre-mRNA. The consequent skipping of exon 51 has the effect (in DMD patients with amenable frameshift exon deletion mutations) of restoring the

mRNA reading frame, leading to an internally shortened but mechanically functional dystrophin protein.

Therapeutic splicing manipulation is also being developed through use of the CRISPR-Cas9 gene editing system. Cryptic splice site mutations in *CEP290* causing severe retinal dystrophy (Leber congenital amaurosis 10) can be edited out in cell models, restoring normal splicing patterns (Ruan *et al.*, 2017). In a dog model of DMD with a deletion of *DMD* exon 50, it has also been shown that CRISPR-Cas9-mediated disruption of an ESE within *DMD* exon 51 can result in both restoration of mRNA reading frame through introduction of indels as well as induced skipping of exon 51, with resulting production of significantly restored levels of dystrophin protein (Amoasii *et al.*, 2018). Similarly, it has also been shown that splice-corrected human myoblasts from DMD patients lacking exon 44 can be generated using CRISPR-Cas9 gene-editing at the iPS cell stage so as to disrupt the splice acceptor site of exon 45 and induce frame-correcting skipping of that exon (Ifuku *et al.*, 2018). Such gene-editing techniques always raise concerns about possible off-target effects that could disrupt important genes. However, approaches to reduce this possibility, such as through the use of precisely targeted base-editing enzymes that do not cause double-strand DNA breaks, are being developed to offer the potential for broadly applicable platforms for splice-site editing (Gapinske *et al.*, 2018).

#### *Alternative splicing signature as a disease biomarker*

Another important role that splicing analysis may play in disease is its use in the discovery and monitoring of disease biomarkers. Since the transcriptome-wide regulation of splicing is a finely balanced process, specific disease signatures are likely to manifest as unique and recognisable patterns of splicing dysregulation. This idea has been shown to be feasible in microsatellite expansion disorders such as myotonic dystrophy and discovery of splicing biomarkers for other diverse disorders may in time prove to be similarly tractable (Sznajder *et al.*, 2018). Should this be the case, it would not only allow for clinically useful measures of disease severity and treatment response but would also provide a powerful means of making diagnoses in the first place. Furthermore, functional genomic splicing signatures of this kind could play critical roles in helping clarify the pathogenicity of the many genomic variants of uncertain significance discovered through genomic testing.

#### ***In vitro/in vivo assays for validation of alternative splicing***

##### ***RT-PCR***

The mainstay of diagnostic splicing analysis for the past several decades has been reverse transcription polymerase chain reaction (RT-PCR). The high sensitivity, sequence specificity and robust reproducibility of RT-PCR make it an ideal tool for the analysis of specific splice events on a quick and relatively cheap basis. However, the targeted nature of PCR generally precludes the ability of comprehensively screening a gene for all possible splicing abnormalities. Unexpected splicing mutations in a gene may therefore be missed by RT-PCR,

as too will pathogenic splicing mutations in alternative disease genes not covered by the assay.

#### *Minigene assays and MPRA*

Despite numerous splicing prediction tools being available, predicted splicing events still need to be validated using wet-lab experiments. A commonly used cell-based *in vitro* approach for alternative splicing studies is the minigene assay (Baralle *et al.*, 2003; Singh and Cooper, 2006). In this assay, the genomic DNA sequence of interest, which must include at least one exon and ideally more than one exon, flanked by upstream and downstream introns, is synthesised by PCR before cloning into a minigene plasmid vector. The insertion site of the vector is sandwiched between two intrinsic exons, one next to a promoter while the other one has a poly-A tail. After cloning the sequence of interest into the plasmid, the construct is transfected into appropriate cells. Splicing variations in the transcripts can be determined by RT-PCR (Figure 3A).

Massively parallel reporter assay (MPRA) is an additional high-throughput, powerful technique for analysing sequence variants affecting alternative splicing. The assay determines the effects of sequence variants on splicing at the single nucleotide level (Rosenberg *et al.*, 2015)(Soemedi *et al.*, 2017). In this assay, a large number of sequence variants of interest are cloned into minigene reporters before transfection into appropriate cell cultures. Thereafter, the transcriptomes can be analysed for alternative splicing using RNA-seq (Figure 3B).

#### *Animal models for splicing analysis*

Using the advantages of transparent model organisms and colourful reporter genes, *in vivo* minigene assays have been developed using *C. elegans* (Wani and Kuroyanagi, 2017). This fluorescent *in vivo* minigene reporter allows the visualisation of tissue-specific alternative splicing patterns in a model organism. Although classic minigene assays have been applied in a vertebrate model organism, the zebrafish (Barboric *et al.*, 2009; Markmiller *et al.*, 2014), fluorescent minigene assays are yet to be tested in it. The potential success of these fluorescent assays in the transparent zebrafish could provide a better understanding of alternative tissue-specific splicing mechanisms *in vivo*.

#### ***In silico* approaches for predicting aberrant splicing**

Alongside improvements in wet-lab based techniques for assaying alternative splicing, dry-lab based computational tools, databases and machine learning techniques are being developed to expand our knowledge of alternative splicing and with a view to applying this knowledge for patient benefit.

#### *Currently available tools*

There are three basic types of tools available based on various different approaches such as consensus sequence analysis and statistical modeling. SplicePredictor (Brendel, Xing and Zhu, 2004), SplicePort (Dogan *et al.*, 2007), GENSCAN (Burge and Karlin, 1997), GeneSplicer (Pertea, Lin and Salzberg, 2001), Spliceman (Lim and Fairbrother, 2012), Human Splicing

Finder (Desmet *et al.*, 2009) and similar other tools have been developed based on consensus splicing donor and acceptor sequences at exon-intron junctions. Tools such as Branch Site Analyser (Kol, Lev-Maor and Ast, 2005), SVM-BP Finder (Corvelo *et al.*, 2010) and IntSplice (Shibata *et al.*, 2016) are able to predict alternative splicing events based on the branch sites and polypyrimidine tracts of introns. Moreover, tools such as EX-SKIP and HOT-SKIP are designed to predict splicing events based on exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs) (Raponi *et al.*, 2011). All these prediction tools rely heavily on degenerate sequence motifs but do not consider the tissue type or cell-specific splicing patterns, non-coding RNAs or splicing factor changes.

#### *Current diagnostic practice and ACMG guidelines*

Current American College of Medical Genetics (ACMG) guidelines for assessing a genomic sequence variant with regards to its effect on splicing recommend the use of multiple *in silico* tools for splicing prediction (Richards *et al.*, 2015). This is to take account of the significant room for error inherent in such tools (Jian, Boerwinkle and Liu, 2014). One reason for this problem is that while most tools generate some sort of splicing “score” to indicate the likelihood of a splicing effect, there are no agreed thresholds for the interpretation of such scores. By combining the outputs of more than one tool, there is some hope that more accurate predictions can be made. A study comparing 272 *BRCA1* and *BRCA2* variants both *in vitro* and *in silico* found that a combined MaxEntScan cutoff value of 15% and a 5% cutoff value for use of a position weight matrix model achieved an overall sensitivity of 96% and a specificity of 83% (Houdayer *et al.*, 2012). More recently, a study combining two or more *in silico* tools (HSF, SSF-like and MES) was able to show 99.44% sensitivity in detecting disruption of splice donor sites and 92.63% sensitivity for disruption of splice acceptor sites in breast/ovarian cancer genes (Moles-Fernández *et al.*, 2018). However, such *in silico* evidence (even if sourced from multiple tools) can also only be applied as one single combined piece of supporting evidence in the assessment of a single variant, so as not to attribute undue weight to such predictions and to take account of some tools using overlapping algorithms.

#### **Alternative splicing RNA-seq analysis**

Over the past decade or so, massively parallel sequencing of RNA (RNA-seq) has become a widely used modality for studying the transcriptome, not only with regard to differential gene expression but increasingly also for alternative splicing and differential isoform expression, as well as for the study of lncRNAs. Following on from RT-PCR’s ability to prove the relevance of disrupted splicing function in disease, RNA-seq has now also been shown to be a complementary diagnostic application for Mendelian genetic disorders (Cummings *et al.*, 2017). Moreover, patients’ RNA-seq analysis can potentially yield unique patterns that are linked to specific diseases, differentiating them from control RNA-seq samples of the same tissue, which could thereby potentially lead to the development of RNA-seq biomarkers of disease.

#### *Data generation options for RNA-seq*



When considering RNA-seq for splicing analysis, it is critical to design the optimal sequencing parameters that will allow extraction of the maximum splicing information from a sample. This starts with suitable sample collection (from a relevant tissue or cell type) and the extraction of high-quality RNA. Poor-quality RNA with low RNA integrity number (RIN) will generate bias when sequenced, favouring for example coverage of the 3' region of transcripts that have been poly-A-selected over the degraded 5' ends. The next consideration is the type of library preparation required. Should total RNA be sequenced, with the inclusion of long non-coding RNAs? This may potentially reduce selection bias but generally comes at the expense of reduced interpretable read-depth of mRNA, requires removal of rRNA and increases sequencing 'noise'. Would poly-A-selected mRNA be more suitable? This may be appropriate if non-coding RNAs are not of interest, since a higher proportion of sequenced reads will map to annotated transcripts, facilitating interpretation. Should small RNAs such as miRNAs be required for analysis, separate library preparation types will generally be required.

In addition, one must choose what sequencing platform to use and what sequencing parameters to employ. While useful gene expression data can be obtained from single-end reads as short as 50 base-pairs, analysis of splicing relies on mapping reads that span exon-exon junctions and this really requires reads of at least 100 bp or more, ideally paired-end (Chhangawala *et al.*, 2015). Read depth is another limiting factor in terms of detecting abnormal splicing events, particularly if a gene is expressed at only low levels in a given tissue. Thus, in general the rule is very much 'the more the better' when it comes to numbers of reads per sample. For short-read mRNA sequencing, around 70 million reads per sample appears to achieve a reasonable balance between adequate coverage and depth without costs becoming prohibitive.

There are, of course, also long-read RNA-seq approaches that are becoming increasingly available in the form of nanopore sequencing and PacBio technologies (Weirather *et al.*, 2017). Such approaches have the advantage of generating full-length reads of individual whole transcripts, making splice isoform variants readily identifiable. Nanopore technology relies on detection of changes in electrical current generated as a single-stranded nucleic acid molecule passes through a protein nanopore. Different patterns of current change are generated by different combinations of bases as they transit the pore, which thereby allows base-calling to take place. An array of nanopores is embedded within a flow cell membrane and this allows parallel sequencing of multiple single molecules during the course of a sequencing run. In addition, since no moving mechanical parts are required for the sequencing process, nanopore sequencers can be very small devices indeed, making them highly portable. Oxford Nanopore Technologies currently offer RNA-seq capability via PCR amplicon-based cDNA sequencing, direct cDNA sequencing without PCR, and also direct RNA sequencing, which potentially removes any bias introduced through reverse transcription. Direct RNA sequencing also has the potential to detect RNA modifications, since these affect how the current is altered as bases move through the nanopore. Nanopore RNA-seq approaches have been shown capable of deconvoluting large numbers of individual splice isoforms (Bolisetty, Rajadinakaran and Graveley, 2015). Detecting allele-specific isoform usage has also been demonstrated (Seki *et al.*, 2018). One issue that remains is the limited

accuracy of base-calling using the nanopore method, which at best may be up to 92.3% for cDNA sequencing, though <90% for direct RNA sequencing. However, methods are in development that should improve such accuracy and may allow applications such as single-cell nanopore transcriptomics (Volden *et al.*, 2018).

PacBio sequencing employs polymerase enzymes individually tethered to the base of zero-mode wave guides, nanophotonic confinement structures that allow excitation and emission of light confined to the very small volume within which the polymerase sits. Single-molecule real-time (SMRT) sequencing within each well proceeds through incorporation of fluorescently labelled nucleotides by the tethered polymerase and the incorporation event can then be detected through epifluorescence. This type of approach applied to herpes viruses resulted in identification of novel transcripts and isoforms, extending the known number of isoforms by at least around 100% (Tombácz *et al.*, 2018). Similarly, SMRT sequencing identified over one thousand novel zebrafish isoforms in comparison to what had previously been annotated through short-read sequencing (Nudelman *et al.*, 2018).

Attempts to accurately quantify and define alternatively spliced isoforms from short-read data have been shown to be of limited accuracy, with analysis tools such as CEM, Cufflinks, iRECKON, RSEM and SLIDE frequently producing high numbers of false positive results (Angelini, Canditiis and Feis, 2014). High GC content can also lead to the over-representation of such fragments in short-read data (Dabney and Meyer, 2012). With the increasing accuracy of nanopore sequencing and the increasing throughput available on PacBio platforms, it is likely that long-read RNA-seq will continue to gain in popularity and will eventually in time become the preferred method for transcriptomic analysis of differential isoforms and alternative splicing.

Another consideration in data generation is the addition of RNA spike-in control sequences (Devonshire, Elaswarapu and Foy, 2010; Lee *et al.*, 2016). Such spike-ins can help in controlling for levels of expression between samples as well as allowing quality assessment of library preparation and sequencing. Related to this, it is widely known that RNA-seq can be highly susceptible to confounding effects introduced by inter-sample and inter-run variability. It is therefore advisable to seek to process and run all samples together in parallel so as to minimise these effects. Having said this, the introduction of RNA-seq into a diagnostic setting will likely make this approach unviable, since although samples from patients may be batched for convenience, it would be prohibitively expensive and time-consuming to fully sequence multiple control samples with each run. Another potential concern is the lack of replicates available for clinical samples. Many downstream RNA-seq data analysis pipelines rely on the presence of replicates for their statistical analysis. In this regard, diagnostic splicing analysis may prove to be resilient, since in many cases a genetic diagnosis will depend only on the identification of a novel splice event in a patient rather than on comparing relative increases or decreases in exon usage.

#### *Alignment and splice site identification (short-read sequencing)*

Once data are generated, they must be analysed appropriately in order to extract meaning. Following quality control and filtering of raw reads, alignment mapping to the genome must

be undertaken by a splice-aware alignment algorithm. In very simple terms, a “splice-unaware” aligner will only map reads that continuously align along their length to the provided reference sequence (e.g. the human genome) and will not make allowances for reads that only partially align (e.g. on account of a splice junction spanning two exons). Junction-spanning reads would therefore tend to be discarded by a splice-unaware aligner. Splice-aware aligners, on the other hand, do make these allowances and can therefore be used to map spliced reads, either to a known transcriptome or to a whole genome. Commonly used splice-aware aligners include TopHat2, GSNAP, Olego, STAR and HISAT2, among others (Wu and Nacu, 2010; Dobin *et al.*, 2013; Kim *et al.*, 2013; Wu *et al.*, 2013; Kim, Langmead and Salzberg, 2015). Mapping is generally a computationally intense process and for full transcriptome analysis this is greatly facilitated by access to a high-performance computing cluster. Once mapping has been completed and BAM files have been generated, subsequent analysis can in many cases be achieved on a personal computer. Of note, it is also possible to map reads to known transcriptome datasets, rather than to a genome. This may prove faster but comes with the cost that unannotated novel transcripts, isoforms and splicing events may not be detected.

Newly developed software such as Kallisto (Bray *et al.*, 2016) and Salmon (Patro *et al.*, 2017) provide faster ways to map and quantify RNA-seq data without using high capacity performance computers and this may prove useful in a diagnostic setting. Kallisto (Bray *et al.*, 2016) software uses pseudo-alignment and offers a very fast method of raw data analysis, analysing 30 million unaligned paired-end RNA-seq reads in less than ten minutes even using a standard laptop computer. However, these are not ideal when analysing multiple isoforms from a single gene as they use very short reads (Xie *et al.*, 2014). Importantly, it must also be emphasised that Kallisto and Salmon are not programs designed for splicing analysis but rather are ultra-rapid read quantitation tools for RNA-seq data.

Following read mapping, a quick and straightforward way to visualise splicing data at specific genomic loci is to view the aligned BAM files via an interface such as the Integrated Genome Viewer (IGV), which can illustrate the numbers of mapped reads spanning individual splice junctions using the Sashimi plot function (Katz *et al.*, 2010). However, direct visualisation is not a high-throughput method and additional software is required to perform more detailed analysis.

### *Differential splicing analysis and isoform quantification*

When RNA-seq was introduced using high-throughput massively parallel sequencing technology (Wang, Gerstein and Snyder, 2009), splicing knowledge expanded exponentially, not only in the ability to discover new novel transcripts but also in terms of measuring the abundance of mRNAs (Mortazavi *et al.*, 2008; Pan *et al.*, 2008; Wang *et al.*, 2008). In addition, RNA-seq is also a useful tool to analyse alternative splicing events at single cell resolution (Song *et al.*, 2017) (Figure 3). Current widely used bead-based RNA capture methods for single cell short-read RNA-seq analysis (such as Drop-seq and 10X Genomics) tend to only yield sequence data from the 3' end of transcripts, limiting its utility for splicing analysis. However, with the correct experimental design, long-read nanopore sequencing can be used to identify isoforms from individual cells (Byrne *et al.*, 2017; Seki *et al.*, 2018; Volden *et al.*, 2018).

Developing an accurate picture of isoform usage at single cell level is likely to be critical for our understanding of complex biological systems such as tumours and immune cell diversity.

Alongside improvements in high-throughput sequencing technologies, software to analyse the raw data of RNA-seq has also been developed and improved. Multiple software packages have been developed that can compare alternative splicing and differential exon usage between samples. However, it is important to note that many such programs are only able to quantify the use of known, annotated splice junctions (or at least junctions that are otherwise pre-specified by the user) and may therefore not be able to identify or account for cryptic splicing events.

Open-source software, such as Cufflinks, allows alignment of short reads to the reference genome to detect alternative splicing (Trapnell *et al.*, 2010). Cufflinks was one of the first RNA-seq analysis algorithms to use short reads (75bp) generated from cDNA, which are then mapped onto a reference genome. About 27% of reported transcripts were previously unannotated and a wide range of splice isoform switches were apparent during different time point analysis. Analytical software such as MISO (mixtures of isoforms)(Katz *et al.*, 2010), SpliceTrap (Wu *et al.*, 2011), and rMATS (Shen *et al.*, 2014) are specifically designed to analyse alternative splicing. The MISO method shows that paired-end reads which are 300 base pairs or more are preferable for the detection of splice isoforms (Katz *et al.*, 2010). SpliceTrap is similar tool which focuses on full length transcript isoforms by quantifying exon inclusion level of every single exon using paired-end RNA-seq data (Wu *et al.*, 2011). A software program called SUPPA is designed for full-length transcript quantitation without using alignment methods, resulting in shorter assembly time (Alamancos *et al.*, 2015).

DEXSeq is a popular program for the analysis of differential exon usage between groups of samples (Anders, Reyes and Huber, 2012). However, it requires replicate samples for its statistical analysis and relies on annotated splice events. SpliceSeq is another program that utilises annotated splice graph alignment of RNA-seq data to quantify and visualise alternative splicing events through an interactive user interface (Ryan *et al.*, 2012). JunctionSeq builds on the DEXSeq methodology by assessing both differential exon and splice junction usage between samples with additional functionality allowing analysis of unannotated junctions and splicing visualisation (Hartley and Mullikin, 2016). SGSeq is a versatile software package that can also quantify splice junction usage but in addition makes direct predictions of novel isoforms based on detection of novel splice junctions in aligned RNA-seq data (Goldstein *et al.*, 2016). Very recently, another program called Whippet has been developed that uses contiguous splice graphs to model transcriptome structure and, through the use of k-mer indexing around splice site boundaries, is able to very rapidly map spliced reads and to efficiently quantify event-level alternative splicing utilising computational resources available on a laptop computer (Sterne-Weiler *et al.*, 2018).

Detailed guidelines for choosing which specific software packages to use for different purposes are reviewed elsewhere (Conesa *et al.*, 2016). However, the rapid rate of development of novel RNA-seq software tools means that up-to-date advice and guidance on the use and applicability of such tools is often best found via online resources, including the many bioinformatics discussion forums that are active on the worldwide web.

## **Deciphering the splicing code via artificial intelligence/machine learning**

Advances in computing over the past few decades have led to the development of machine learning algorithms that can be trained to recognise patterns in data through a system akin to experiential learning. These “deep learning” algorithms use data fed back from correct and incorrect inferences and decisions to refine their ability to achieve a predefined goal. This type of artificial intelligence (AI) approach effectively frees such algorithms from the constraints of having to rely upon human-devised rules of decision-making and allows instead the discovery of novel methods and solutions to a problem that may not have been obvious to (or possible for) a human observer.

Machine learning holds promise for disentangling the complex splicing prediction process and has already been applied in various genomics studies including splice site identification and the classification of splice-altering variants (Libbrecht and Noble, 2015; Xiong *et al.*, 2015). The idea of machine learning in splicing is to accurately predict a splicing event from any given sequence using rules that have been learnt from previous experiences. An early example was ExonScan which looks for potential splice donor and acceptor sites in any given sequence using maximum entropy splice site models, before presenting a candidate exon which has the highest scores (Wang *et al.*, 2004). In the same year, Sorek *et al.* developed a method to predict exon skipping without using ESTs (Sorek *et al.*, 2004). Support vector machine (Dror, Sorek and Shamir, 2005) and Acescan (Yeo *et al.*, 2005) are similar machine learning-based algorithms to predict alternative splicing. In 2010, Barash *et al.* presented a more sophisticated method to predict alternative splicing (Barash *et al.*, 2010), where the distinctive feature was integration of tissue/cell-specific splicing, not previously considered. The results produced new classes of splicing patterns as well as mutation-verified regulatory sequences. Hidden variables were added to this method using a Bayesian approach to improve splicing predictions (Xiong, Barash and Frey, 2011). A Galaxy-based web tool called AVISPA which predicts tissue-specific splicing patterns as well as regulatory element associations, is also based on this method (Barash *et al.*, 2013).

Deep learning uses multilayer data processing with multilevel abstraction (LeCun, Bengio and Hinton, 2015). Unlike previous models, Leung *et al.* demonstrated a deep learning approach to understanding the tissue-regulated splicing code by employing a deep neural network algorithm applied to RNA-seq-based tissue-specific splicing patterns (Leung *et al.*, 2014), which outperformed previous Bayesian neural network methods. It has also been shown that machine learning can now predict if a sequence variant is likely to interrupt splicing in a tissue-specific manner, even if the variant is outside the protein-coding region (Xiong *et al.*, 2015; Jha, Gazzara and Barash, 2017). This method could prove particularly useful for predicting the effect of an unknown variant with regards to its pathogenicity and clinical significance. In their publication, Xiong *et al.* report that the method correctly predicted up to 94% of splice disruptions in disease-related variants (Xiong *et al.*, 2015). This is a fast-changing and promising field, a very recent example of which is COSSMO (Competitive Splicing Site Model), a novel model using deep learning to predict competitive effects on splice site selection based on sequence alone (Bretschneider *et al.*, 2018).

Despite the rapid rise of AI, the challenge of understanding the “splicing code” presents machine learning with a difficult problem that may or may not be solvable. As we have seen, regulation of splicing and the use of specific splice sites depends not only on key RNA sequence motifs and their surrounding sequence contexts but also on epigenetic factors controlling gene expression, *trans*-acting levels of RNA-binding proteins and other cell- and tissue-specific environmental factors, many of which will differ to some extent from person to person. This inherent variability and the stochastic nature of biological processes can therefore make them somewhat refractory to reliable predictive modelling, since many such variables are unknowable in practical terms. Nevertheless, the computational power of AI-based methods applied to large multi-omic datasets makes it likely they will succeed in identifying at least some significant novel and biologically relevant connections between splice-regulating elements.

## Conclusion

The fields of medical genetics, RNA biology and data science are experiencing unprecedented explosions of knowledge and understanding, driven largely by the NGS revolution and by parallel continued advances in both computational hardware and software. With such change comes associated challenges: for medical genetics the challenge of variant interpretation for clinical diagnosis, for RNA biology the challenge of understanding the functions of the multiple coding, non-coding and small RNA species, and for data science the challenge of extracting meaning from vast datasets. However, this same change also brings opportunities in all these fields in ways that should be synergistic and mutually beneficial.

### *The utility of RNA in diagnostics*

The current paradigm of medical genetic diagnosis only considers testing of RNA as something of an afterthought, for example to confirm or refute the effect of an apparent splice site mutation. Indeed, little consideration is given to variants (whether exonic or intronic) that could alter splicing outside of annotated canonical splice sites and synonymous coding variants are immediately discounted by virtually all standard variant filtering pipelines, irrespective of their potential effects on critical splice enhancer or silencer sequences. Current clinically used algorithms for predicting the splicing effects of variants are wholly inadequate for the reliable filtering of this complexity and it is therefore without doubt that a significant proportion of genetic diagnoses are being missed because of cryptic RNA abnormalities of this kind.

In order to address this diagnostic gap, a shift in the paradigm of genetic investigation may be required. Namely, rather than turning to RNA analysis as a follow-up test of “last resort”, it may prove more effective to consider either targeted or transcriptome-wide RNA analysis concurrently with DNA-based investigations. This will be relevant not just for the diagnosis of rare genetic conditions but also for diagnosis and biomarker-based monitoring of polygenic and complex medical conditions that bear distinctive transcriptomic signatures. It may even be that whole-transcriptome RNA-seq will prove to be an effective first-line genetic screening test, abnormalities in which could be followed up by DNA-based confirmation. Since RNA-seq returns not only expression level and splicing data but also variant level sequencing data, coding variants can in fact be called as part of the analysis. The majority (at least 80%) of human genomic coding sequences are found to be expressed at some level in blood and with sufficient read depth it should therefore be possible to generate sufficient RNA-seq data for this type of analysis (Liew, 2006).

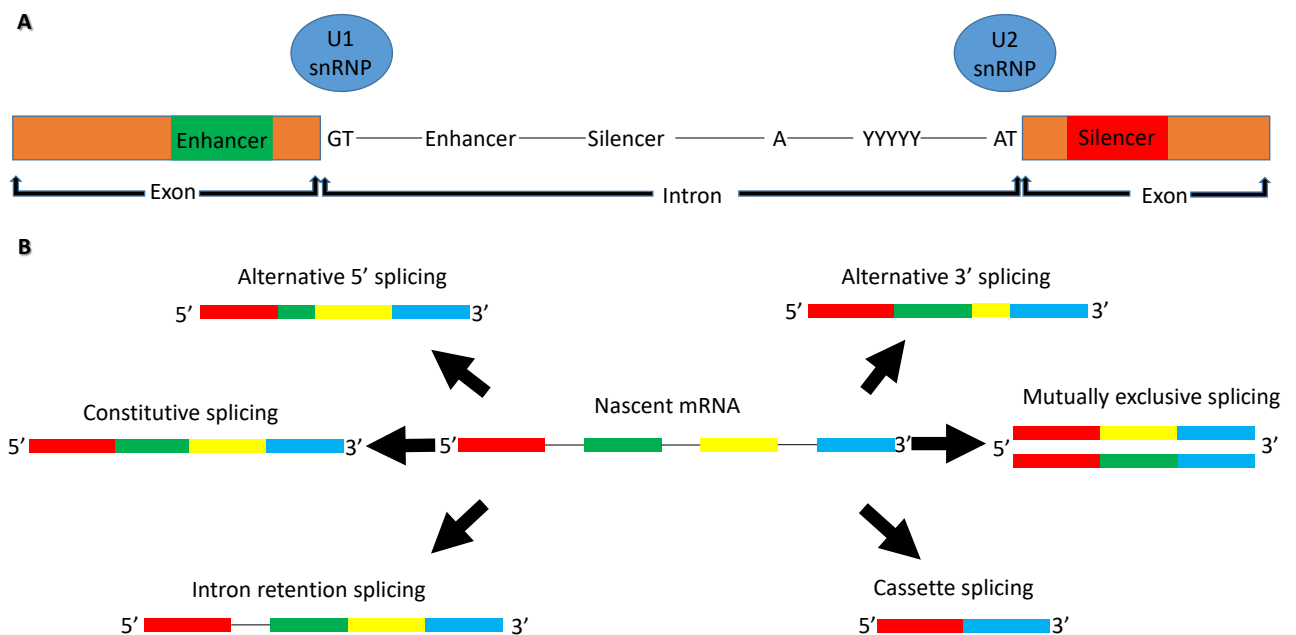
### *Personalised functional genomics and the future diagnostics-therapeutics pipeline*

Up until now, genetic diagnoses have all too often been associated with a lack of disease-modifying therapeutic options and their individual rarity has hindered investment in the development of such treatments. However, as mentioned previously, novel RNA-based therapies targeting splicing have now started to enter clinical usage. Such ASO-based RNA therapeutics are attractive on account of their versatility, specificity and titratability and are likely to play an increasing role in the treatment of both common and rare medical conditions. A key attribute of ASO technology is its potential for personalisation, since an individual

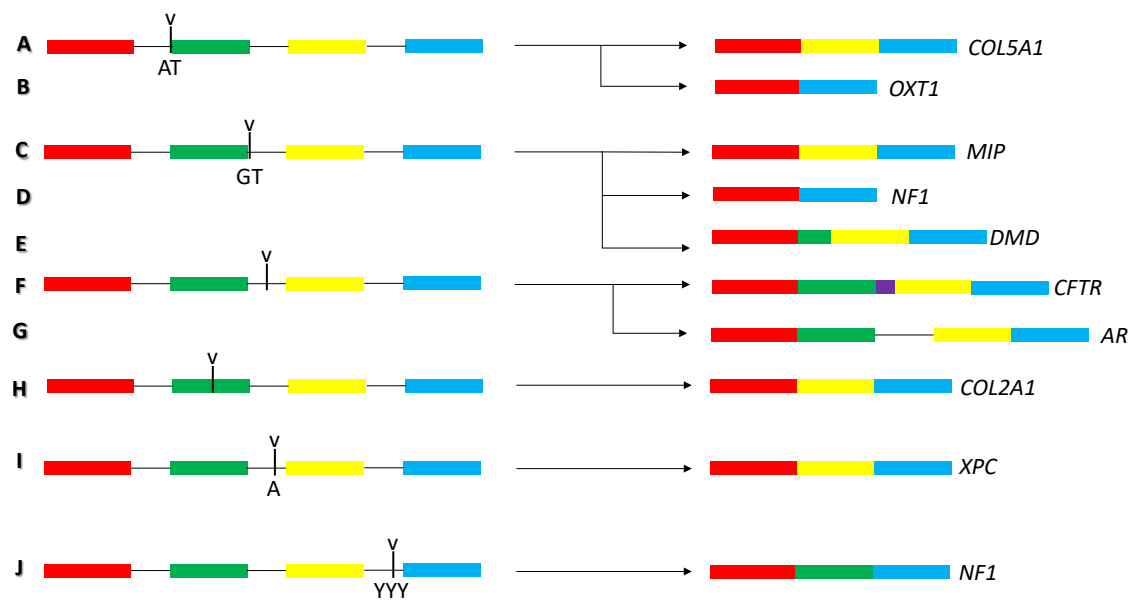
patient's splicing mutation could be uniquely targeted for correction given the correct sequence modifications. Whilst drug regulatory authority practices are not currently compatible with the idea of sequence-specific personalised drug development, the pressing need for such therapies in orphan diseases may act as a catalyst for the regulatory changes required to facilitate these truly personalised medicines.

If RNA-based therapeutics are to be part of the future of medicine, so too must be the routine analysis of RNA and of splicing in particular. Modern medicine increasingly relies upon correct molecular diagnosis for guiding clinical management decisions and the diagnosis of a pathogenic splicing mutation in a patient could in future allow bespoke splice-modulating therapies to be employed. Thus, it may now be time to consider routine collection of RNA samples alongside DNA samples, certainly in the case of suspected monogenic disorders and perhaps also in the setting of selected other disorders where transcriptomic profiling may aid diagnosis. More broadly, our best chance of achieving a holistic molecular understanding of an individual patient's disease is by leveraging the power of personalised multi-omic datasets (genomic, transcriptomic, methylomic, proteomic, metabolomic and beyond) to create a truly data-integrated approach to genomic medicine.



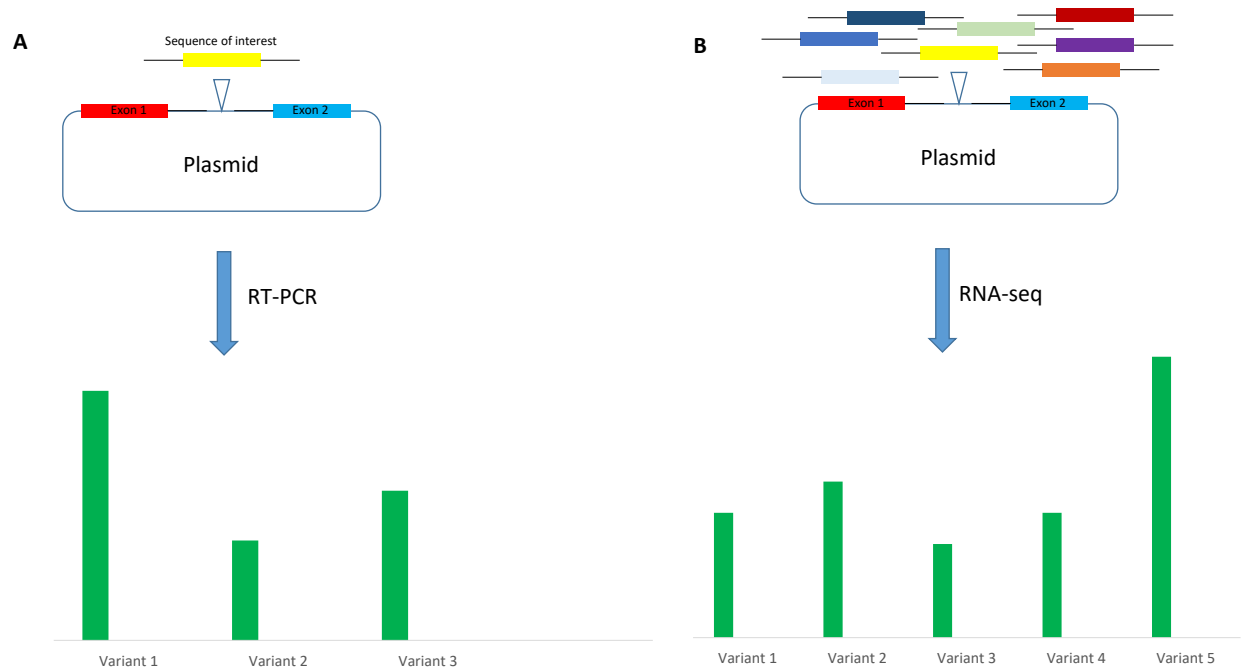


**Figure 1. Simplified diagram of an intron and five different types of alternative splicing.** (A) Classic intron characteristics consist of GT at the 5' splice donor site and AT at the 3' splice acceptor site. These are recognised by U1snRNP and U2snRNP respectively. The branch point is located approximately 40 nucleotides upstream of the splice acceptor and is always an adenine. The polypyrimidine tract is a chain of 15-20 pyrimidine bases near the splice acceptor site. In addition to these conserved sequences, both introns and exons have less well conserved sequences which regulate splicing as enhancers or silencers. (B) A nascent mRNA with four exons can produce six types of different mature mRNAs. In addition to simple constitutive splicing which simply joins all the exons after intron removal, there are five different types of alternative splicing patterns possible. These are: alternative 5' splicing, alternative 3' splicing, mutually exclusive splicing, cassette splicing and retained intron splicing.

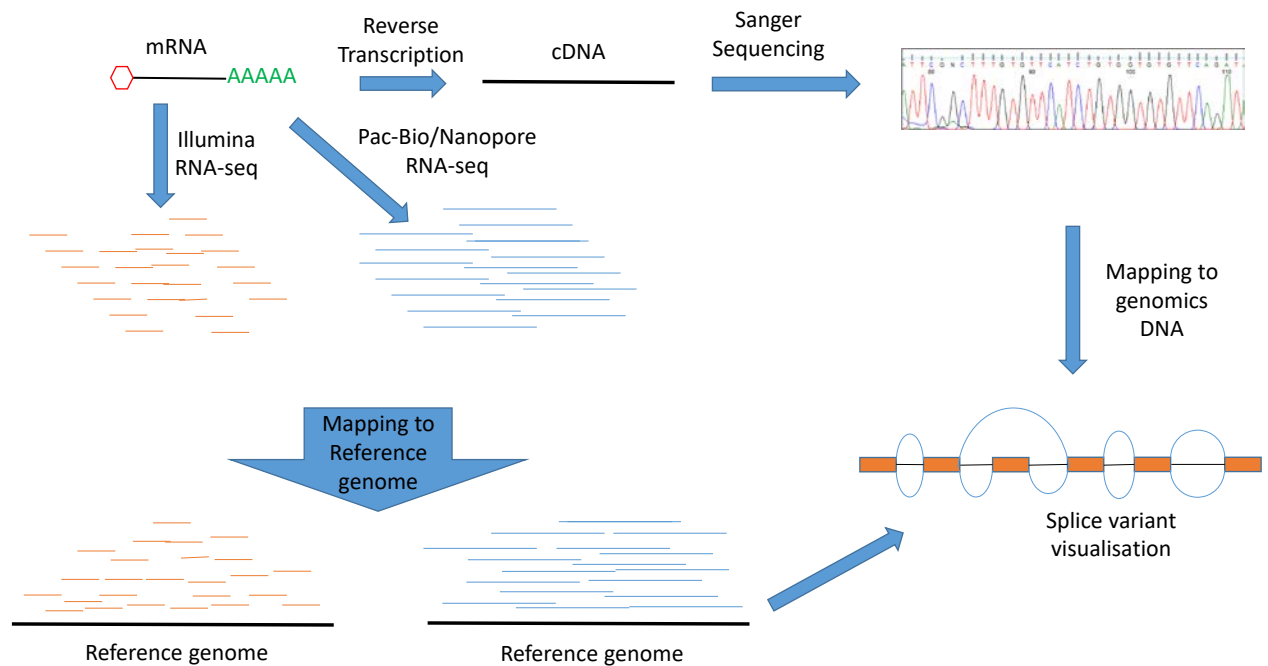


**Figure 2. Diagram of different alternative splicing patterns reported in disease phenotypes.**

A single nucleotide change in a splice acceptor site causes skipping of either (A) a single adjacent exon or (B) two or more adjacent exons. Similar exon skipping patterns have been reported in cases with sequence variants in splice donor sites (C and D). Variants in a splice donor site can also activate a new splice donor site in an adjacent exon (E). Intronic sequence variants outside of the conserved splice donor, splice acceptor, branch point or polypyrimidine tract sequences can also produce a new cryptic exon as well as retained intron splice variants (F and G). Exonic sequence variants can cause the same exon skipping in some diseases (H). Both branch point and polypyrimidine tract sequence variants are also associated with exon skipping (I and J).

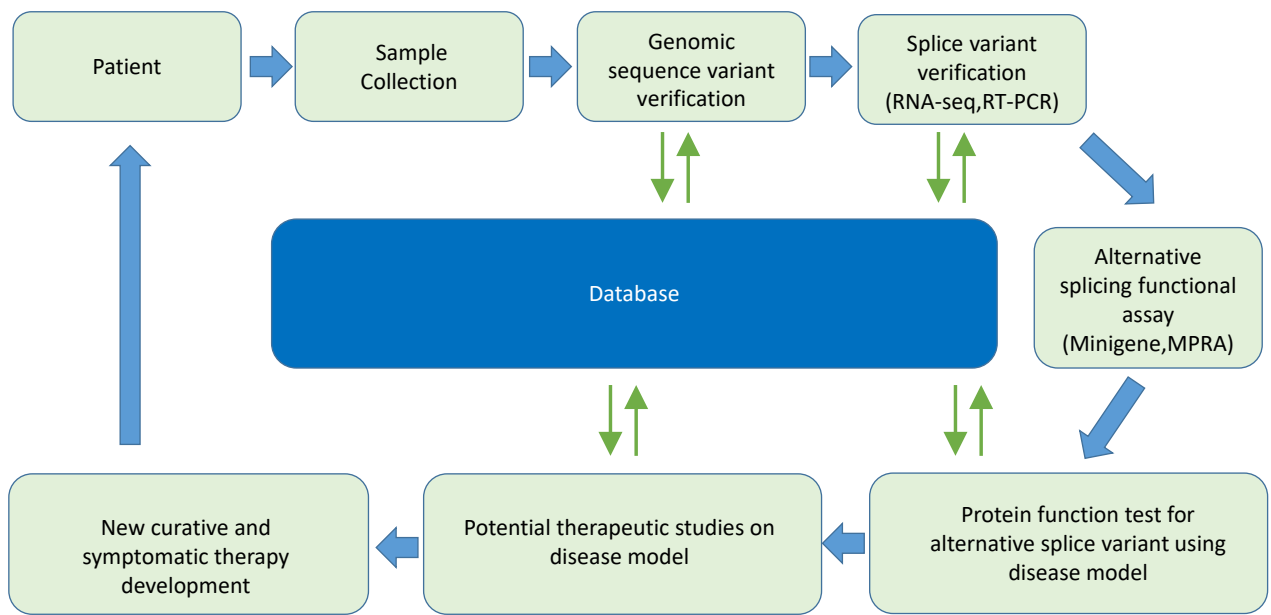


**Figure 3. Comparison between classic minigene and high-throughput MPRA minigene assays.** (A) Schematic diagram of classic minigene assay. A sequence of interest, which usually includes an exon with its flanking intron sequences, is inserted into a plasmid vector minigene which has an exon with promoter and an exon with poly-A sequence. After transfection into cells, mRNA abundance is measured by qRT-PCR. (B) Schematic diagram of high-throughput massively parallel reporter assay (MPRA) minigene. This assay uses the same principle of a classic minigene. However, it uses multiple sequence variants from a library to insert into the vectors. After transfection, the splice variant patterns are measured by RNA-seq.



**Figure 4. Applications of three different technologies to detect alternatively spliced mRNA.**

First generation sequencing or Sanger sequencing needs reverse transcription of mRNA into cDNA before sequencing. Then, the cDNA is aligned to the genomic sequence to detect alternative splicing patterns. Next generation sequencing platforms such as Illumina, PacBio and nanopore technologies can sequence cDNA to measure RNA abundance as well as alternative splicing. PacBio and nanopore sequencing generate much longer reads than Illumina.



**Figure 5. Workflow of integration of alternative splicing studies into genomic databases and diagnostics.** RNA-seq from patients with genetic disorders can be integrated into machine learning and splicing prediction tools. The performance of tools should be validated using *in vitro* assays. Pathogenicity of splice variants can be tested in animal models before the models are used for drug testing and genome editing studies.

## References

- Alamancos, G. P. *et al.* (2015) 'Leveraging transcript quantification for fast computation of alternative splicing profiles', *RNA*, 21(9), pp. 1521–1531. doi: 10.1261/rna.051557.115.
- Amoasii, L. *et al.* (2018) 'Gene editing restores dystrophin expression in a canine model of Duchenne muscular dystrophy', *Science*. American Association for the Advancement of Science, 362(6410), pp. 86–91. doi: 10.1126/SCIENCE.AAU1549.
- Anders, S., Reyes, a and Huber, W. (2012) 'Detecting diferential usage of exons from RNA-seq data', *Genome Res*, 22(10), pp. 2008–2017. doi: 10.1101/gr.133744.111.
- Angelini, C., Canditiis, D. and Feis, I. (2014) 'Computational approaches for isoform detection and estimation: good and bad news', *BMC Bioinformatics*. BioMed Central, 15(1), p. 135. doi: 10.1186/1471-2105-15-135.
- Aoyama, Y. *et al.* (2017) 'A novel mutation (c.121-13T>A) in the polypyrimidine tract of the splice acceptor site of intron 2 causes exon 3 skipping in mitochondrial acetoacetyl-CoA thiolase gene', *Molecular Medicine Reports*. Spandidos Publications, 15(6), pp. 3879–3884. doi: 10.3892/mmr.2017.6434.
- Ars, E. *et al.* (2000) 'Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1.', *Human molecular genetics*, 9(2), pp. 237–47.
- Baralle, M. *et al.* (2003) 'Identification of a mutation that perturbs NF1 agene splicing using genomic DNA samples and a minigene assay.', *Journal of medical genetics*. BMJ Publishing Group, 40(3), pp. 220–2. doi: 10.1136/JMG.40.3.220.
- Baralle, M. and Baralle, F. E. (2018) 'The splicing code', *Biosystems*. Elsevier, 164, pp. 39–48. doi: 10.1016/J.BIOSYSTEMS.2017.11.002.
- Barash, Y. *et al.* (2010) 'Deciphering the splicing code', *Nature*. Nature Publishing Group, 465(7294), pp. 53–59. doi: 10.1038/nature09000.
- Barash, Y. *et al.* (2013) 'AVISPA: a web tool for the prediction and analysis of alternative splicing', *Genome Biology*, 14(10), p. R114. doi: 10.1186/gb-2013-14-10-r114.
- Barboric, M. *et al.* (2009) '7SK snRNP/P-TEFb couples transcription elongation with alternative splicing and is essential for vertebrate development.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 106(19), pp. 7798–803. doi: 10.1073/pnas.0903188106.
- Bolisetty, M. T., Rajadinakaran, G. and Graveley, B. R. (2015) 'Determining exon connectivity in complex mRNAs by nanopore sequencing', *Genome Biology*, 16(1), p. 204. doi: 10.1186/s13059-015-0777-z.
- Bray, N. L. *et al.* (2016) 'Near-optimal probabilistic RNA-seq quantification', *Nature Biotechnology*, 34(5), pp. 525–527. doi: 10.1038/nbt.3519.
- Breathnach, R. *et al.* (1978) 'Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries.', *Proceedings of the National Academy of Sciences of the United States of America*, 75(10), pp. 4853–7.

- Brendel, V., Xing, L. and Zhu, W. (2004) 'Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus', *Bioinformatics*, 20(7), pp. 1157–1169. doi: 10.1093/bioinformatics/bth058.
- Bretschneider, H. *et al.* (2018) 'COSSMO: Predicting Competitive Alternative Splice Site Selection using Deep Learning', *bioRxiv*. Cold Spring Harbor Laboratory, p. 255257. doi: 10.1101/255257.
- Buratti, E. and Baralle, D. (2012) 'Exon Skipping Mutations in Neurofibromatosis', in *Methods in molecular biology (Clifton, N.J.)*, pp. 65–76. doi: 10.1007/978-1-61779-767-5\_5.
- Burge, C. and Karlin, S. (1997) 'Prediction of complete gene structures in human genomic DNA', *Journal of Molecular Biology*, 268(1), pp. 78–94. doi: 10.1006/jmbi.1997.0951.
- Burset, M. (2000) 'Analysis of canonical and non-canonical splice sites in mammalian genomes', *Nucleic Acids Research*. doi: 10.1093/nar/28.21.4364.
- Byrne, A. *et al.* (2017) 'Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells', *Nature Communications*. Nature Publishing Group, 8, p. 16027. doi: 10.1038/ncomms16027.
- Charlet-B, N. *et al.* (2002) 'Loss of the muscle-specific chloride channel in type 1 myotonic dystrophy due to misregulated alternative splicing', *Mol Cell*, 10(1), pp. 45–53. doi: 10.1016/S1097-2765(02)00572-5.
- Chhangawala, S. *et al.* (2015) 'The impact of read length on quantification of differentially expressed genes and splice junction detection', *Genome Biology*. Genome Biology, 16(1), p. 131. doi: 10.1186/s13059-015-0697-y.
- Conesa, A. *et al.* (2016) 'A survey of best practices for RNA-seq data analysis', *Genome Biology*. BioMed Central, 17(1), p. 13. doi: 10.1186/s13059-016-0881-8.
- Conlon, E. G. *et al.* (2016) 'The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains', *eLife*, 5, p. e17820. doi: 10.7554/eLife.17820.
- Cooper-Knock, J. *et al.* (2014) 'Sequestration of multiple RNA recognition motif-containing proteins by C9orf72 repeat expansions', *Brain*, 137, pp. 2040–2051. doi: 10.1093/brain/awu120.
- Cooper-Knock, J. *et al.* (2015) 'C9ORF72 GGGGCC expanded repeats produce splicing dysregulation which correlates with disease severity in amyotrophic lateral sclerosis', *Plos One*, 10, p. e0127376. doi: 10.1371/journal.pone.0127376.
- Cooper, D. *et al.* (2014) 'Long Non-Coding RNA NEAT1 Associates with SRp40 to Temporally Regulate PPAR $\gamma$ 2 Splicing during Adipogenesis in 3T3-L1 Cells', *Genes*, 5(4), pp. 1050–1063. doi: 10.3390/genes5041050.
- Corvelo, A. *et al.* (2010) 'Genome-Wide Association between Branch Point Properties and Alternative Splicing', *PLoS Computational Biology*. Edited by I. M. Meyer, 6(11), p. e1001016. doi: 10.1371/journal.pcbi.1001016.
- Cummings, B. B. *et al.* (2017) 'Improving genetic diagnosis in Mendelian disease with

transcriptome sequencing.’, *Science translational medicine*. NIH Public Access, 9(386). doi: 10.1126/scitranslmed.aal5209.

Dabney, J. and Meyer, M. (2012) ‘Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries’, *BioTechniques*, 52(2), pp. 87–94. doi: 10.2144/000113809.

Desmet, F.-O. *et al.* (2009) ‘Human Splicing Finder: an online bioinformatics tool to predict splicing signals’, *Nucleic Acids Research*, 37(9), pp. e67–e67. doi: 10.1093/nar/gkp215.

Devonshire, A. S., Elaswarapu, R. and Foy, C. A. (2010) ‘Evaluation of external RNA controls for the standardisation of gene expression biomarker measurements’, *BMC Genomics*, 11(662). doi: 10.1186/1471-2164-11-662.

Dey, B. K., Mueller, A. C. and Dutta, A. (2014) ‘Long non-coding RNAs as emerging regulators of differentiation, development, and disease.’, *Transcription*. Taylor & Francis, 5(4), p. e944014. doi: 10.4161/21541272.2014.944014.

Dobin, A. *et al.* (2013) ‘STAR: Ultrafast universal RNA-seq aligner’, *Bioinformatics*, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Dogan, R. I. *et al.* (2007) ‘SplicePort--An interactive splice-site analysis tool’, *Nucleic Acids Research*, 35(Web Server), pp. W285–W291. doi: 10.1093/nar/gkm407.

Douglas, A. G. L. and Wood, M. J. A. (2011) ‘RNA splicing: disease and therapy.’, *Briefings in functional genomics*, 10(3), pp. 151–64. doi: 10.1093/bfgp/elr020.

Douglas, A. G. L. and Wood, M. J. A. (2013) ‘Splicing therapy for neuromuscular disease.’, *Molecular and cellular neurosciences*. The Authors, 56, pp. 169–85. doi: 10.1016/j.mcn.2013.04.005.

Dror, G., Sorek, R. and Shamir, R. (2005) ‘Accurate identification of alternatively spliced exons using support vector machine’, *Bioinformatics*, 21(7), pp. 897–901. doi: 10.1093/bioinformatics/bti132.

Dvinge, H. (2018) ‘Regulation of alternative mRNA splicing: old players and new perspectives’, *FEBS Letters*. doi: 10.1002/1873-3468.13119.

Fang, L. J. *et al.* (2001) ‘A novel mutation in the neurofibromatosis type 1 (NF1) gene promotes skipping of two exons by preventing exon definition’, *Journal of Molecular Biology*. Academic Press, 307(5), pp. 1261–1270. doi: 10.1006/JMBI.2001.4561.

Finkel, R. S. *et al.* (2016) ‘Treatment of infantile-onset spinal muscular atrophy with nusinersen: a phase 2, open-label, dose-escalation study’, *The Lancet*. Elsevier Ltd, 388(10063), pp. 3017–3026. doi: 10.1016/S0140-6736(16)31408-8.

Gapinske, M. *et al.* (2018) ‘CRISPR-SKIP: programmable gene splicing with single base editors’, *Genome Biology*. BioMed Central, 19(1), p. 107. doi: 10.1186/s13059-018-1482-5.

Goldstein, L. D. *et al.* (2016) ‘Prediction and quantification of splice events from RNA-seq data’, *PLoS ONE*, 11(5), p. e0156132. doi: 10.1371/journal.pone.0156132.

Habara, Y. *et al.* (2009) ‘In vitro splicing analysis showed that availability of a cryptic splice site is not a determinant for alternative splicing patterns caused by +1G--&gt;A mutations in



introns of the dystrophin gene.', *Journal of medical genetics*. BMJ Publishing Group Ltd, 46(8), pp. 542–7. doi: 10.1136/jmg.2008.061259.

Haerty, W. and Ponting, C. P. (2015) 'Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci.', *RNA (New York, N.Y.)*. Cold Spring Harbor Laboratory Press, 21(3), pp. 333–46. doi: 10.1261/rna.047324.114.

Hartley, S. W. and Mullikin, J. C. (2016) 'Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq', *Nucleic Acids Research*, 44(15), p. e127. doi: 10.1093/nar/gkw501.

Hori, T. *et al.* (2013) 'Molecular Basis of Two-Exon Skipping (Exons 12 and 13) by c.1248+5g>a in *OXCT1* Gene: Study on Intermediates of *OXCT1* Transcripts in Fibroblasts', *Human Mutation*, 34(3), pp. 473–480. doi: 10.1002/humu.22258.

Houdayer, C. *et al.* (2012) 'Guidelines for Splicing Analysis in Molecular Diagnosis Derived from a Set of 327 Combined In Silico / In Vitro Studies on BRCA1 and BRCA2 Variants', *Human Mutation*, 33(8), pp. 1228–1238. doi: 10.1002/humu.22101.

Ifuku, M. *et al.* (2018) 'Restoration of Dystrophin Protein Expression by Exon Skipping Utilizing CRISPR-Cas9 in Myoblasts Derived from DMD Patient iPS Cells', in *Methods in molecular biology (Clifton, N.J.)*, pp. 191–217. doi: 10.1007/978-1-4939-8651-4\_12.

Jha, A., Gazzara, M. R. and Barash, Y. (2017) 'Integrative deep models for alternative splicing', *Bioinformatics*. Oxford University Press, 33(14), pp. i274–i282. doi: 10.1093/bioinformatics/btx268.

Jian, X., Boerwinkle, E. and Liu, X. (2014) 'In silico tools for splicing defect prediction: a survey from the viewpoint of end users.', *Genetics in Medicine*, 16(7), pp. 497–503. doi: 10.1038/gim.2013.176.

Känsäkoski, J. *et al.* (2016) 'Complete androgen insensitivity syndrome caused by a deep intronic pseudoexon-activating mutation in the androgen receptor gene', *Scientific Reports*, 6(1), p. 32819. doi: 10.1038/srep32819.

Katz, Y. *et al.* (2010) 'Analysis and design of RNA sequencing experiments for identifying isoform regulation', *Nature Methods*, 7(12), pp. 1009–1015. doi: 10.1038/nmeth.1528.

Khan, S. G. *et al.* (2010) 'XPC branch-point sequence mutations disrupt U2 snRNP binding, resulting in abnormal pre-mRNA splicing in xeroderma pigmentosum patients', *Human Mutation*, 31(2), pp. 167–175. doi: 10.1002/humu.21166.

Kim, D. *et al.* (2013) 'TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions', *Genome Biology*, 14(4), p. R36. doi: 10.1186/gb-2013-14-4-r36.

Kim, D., Langmead, B. and Salzberg, S. L. (2015) 'HISAT: a fast spliced aligner with low memory requirements', *Nature Methods*, 12(4), pp. 357–360. doi: 10.1038/nmeth.3317.

Kol, G., Lev-Maor, G. and Ast, G. (2005) 'Human–mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation', *Human Molecular Genetics*, 14(11), pp. 1559–1568. doi: 10.1093/hmg/ddi164.

- Krawczak, M., Reiss, J. and Cooper, D. N. (1992) *The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences*, *Hum Genet*.
- LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*. Nature Publishing Group, 521(7553), pp. 436–444. doi: 10.1038/nature14539.
- Lee, H. *et al.* (2016) 'External RNA Controls Consortium Beta Version Update', *Journal of Genomics*, 4, pp. 19–22. doi: 10.7150/jgen.16082.
- Leung, M. K. K. *et al.* (2014) 'Deep learning of the tissue-regulated splicing code.', *Bioinformatics (Oxford, England)*. Oxford University Press, 30(12), pp. i121–9. doi: 10.1093/bioinformatics/btu277.
- Libbrecht, M. W. and Noble, W. S. (2015) 'Machine learning applications in genetics and genomics.', *Nature reviews. Genetics*. NIH Public Access, 16(6), pp. 321–32. doi: 10.1038/nrg3920.
- Lim, K. H. and Fairbrother, W. G. (2012) 'Spliceman--a computational web server that predicts sequence variations in pre-mRNA splicing', *Bioinformatics*, 28(7), pp. 1031–1032. doi: 10.1093/bioinformatics/bts074.
- Lin, X. *et al.* (2006) 'Failure of MBNL1-dependent post-natal splicing transitions in myotonic dystrophy', *Human Molecular Genetics*, 15(13), pp. 2087–2097. doi: 10.1093/hmg/ddl132.
- Lloyd, J. P. B. (2018) 'The evolution and diversity of the nonsense-mediated mRNA decay pathway', *F1000Research*, 7, p. 1299. doi: 10.12688/f1000research.15872.1.
- López-Bigas, N. *et al.* (2005) 'Are splicing mutations the most frequent cause of hereditary disease?', *FEBS Letters*, 579(9), pp. 1900–1903. doi: 10.1016/j.febslet.2005.02.047.
- Lord, J. *et al.* (2018) 'The contribution of non-canonical splicing mutations to severe dominant developmental disorders', *bioRxiv*. Cold Spring Harbor Laboratory, p. 256636. doi: 10.1101/256636.
- Markmiller, S. *et al.* (2014) 'Minor class splicing shapes the zebrafish transcriptome during development.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 111(8), pp. 3062–7. doi: 10.1073/pnas.1305536111.
- McAlinden, A. *et al.* (2008) 'Missense and nonsense mutations in the alternatively-spliced exon 2 of COL2A1 cause the ocular variant of Stickler syndrome', *Human Mutation*, 29(1), pp. 83–90. doi: 10.1002/humu.20603.
- Mendell, J. R. *et al.* (2016) 'Longitudinal effect of eteplirsen versus historical control on ambulation in Duchenne muscular dystrophy', *Annals of Neurology*, 79, pp. 257–271. doi: 10.1002/ana.24555.
- Moles-Fernández, A. *et al.* (2018) 'Computational tools for splicing defect prediction in breast/ovarian cancer genes: how efficient are they at predicting RNA alterations?', *Frontiers in Genetics*, 9(September), p. 366. doi: 10.3389/fgene.2018.00366.
- Mortazavi, A. *et al.* (2008) 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nature Methods*, 5(7), pp. 621–628. doi: 10.1038/nmeth.1226.

- Nudelman, G. *et al.* (2018) 'High resolution annotation of zebrafish transcriptome using long-read sequencing', *Genome Research*, 28(9), pp. 1415–1425. doi: 10.1101/gr.223586.117.
- Pan, Q. *et al.* (2008) 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nature Genetics*, 40(12), pp. 1413–1415. doi: 10.1038/ng.259.
- Patro, R. *et al.* (2017) 'Salmon provides fast and bias-aware quantification of transcript expression', *Nature Methods*, 14(4), pp. 417–419. doi: 10.1038/nmeth.4197.
- Pertea, M., Lin, X. and Salzberg, S. L. (2001) 'GeneSplicer: a new computational method for splice site prediction.', *Nucleic acids research*, 29(5), pp. 1185–90.
- Polymenidou, M. *et al.* (2011) 'Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43', *Nature Neuroscience*. Nature Publishing Group, 14(4), pp. 459–468. doi: 10.1038/nn.2779.
- Raponi, M. *et al.* (2008) 'Polypyrimidine tract binding protein regulates alternative splicing of an aberrant pseudoexon in NF1', *FEBS Journal*. Wiley/Blackwell (10.1111), 275(24), pp. 6101–6108. doi: 10.1111/j.1742-4658.2008.06734.x.
- Raponi, M. *et al.* (2011) 'Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6', *Human Mutation*, 32(4), pp. 436–444. doi: 10.1002/humu.21458.
- Raponi, M., Upadhyaya, M. and Baralle, D. (2006) 'Functional splicing assay shows a pathogenic intronic mutation in neurofibromatosis type 1 (NF1) due to intronic sequence exonization', *Human Mutation*, 27(3), pp. 294–295. doi: 10.1002/humu.9412.
- Reddy, K. *et al.* (2013) 'The disease-associated r(GGGGCC)<sub>n</sub> repeat from the C9orf72 gene forms tract length-dependent uni- and multimolecular RNA G-quadruplex structures.', *The Journal of biological chemistry*, 288(14), pp. 9860–6. doi: 10.1074/jbc.C113.452532.
- Richards, S. *et al.* (2015) 'Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology', *Genetics in Medicine*, 17(5), pp. 405–423. doi: 10.1038/gim.2015.30.
- Rinaldi, C. and Wood, M. J. A. (2018) 'Antisense oligonucleotides: the next frontier for treatment of neurological disorders', *Nature Reviews Neurology*. Nature Publishing Group, 14(1), pp. 9–21. doi: 10.1038/nrneurol.2017.148.
- Romero-Barrios, N. *et al.* (2018) 'Splicing regulation by long noncoding RNAs.', *Nucleic acids research*. Oxford University Press, 46(5), pp. 2169–2184. doi: 10.1093/nar/gky095.
- Rosenberg, A. B. *et al.* (2015) 'Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences', *Cell*, 163(3), pp. 698–711. doi: 10.1016/j.cell.2015.09.054.
- Ruan, G.-X. *et al.* (2017) 'CRISPR/Cas9-Mediated Genome Editing as a Therapeutic Approach for Leber Congenital Amaurosis 10', *Molecular Therapy*, 25(2), pp. 331–341. doi: 10.1016/j.ymthe.2016.12.006.

Ryan, M. C. *et al.* (2012) 'SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts', *Bioinformatics*, 28(18), pp. 2385–2387. doi: 10.1093/bioinformatics/bts452.

Sanz, D. J. *et al.* (2017) 'Cas9/gRNA targeted excision of cystic fibrosis-causing deep-intronic splicing mutations restores normal splicing of CFTR mRNA', *PLOS ONE*. Edited by E. Buratti, 12(9), p. e0184009. doi: 10.1371/journal.pone.0184009.

Seki, M. *et al.* (2018) 'OUP accepted manuscript', *Dna Research*. doi: 10.1093/dnares/dsy038.

Shen, S. *et al.* (2014) 'rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data.', *Proceedings of the National Academy of Sciences of the United States of America*, 111(51), pp. E5593–5601. doi: 10.1073/pnas.1419161111.

Shibata, A. *et al.* (2016) 'IntSplice: prediction of the splicing consequences of intronic single-nucleotide variations in the human genome', *Journal of Human Genetics*, 61(7), pp. 633–640. doi: 10.1038/jhg.2016.23.

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), pp. 177–181. doi: 10.2144/000112208.

Singh, R. *et al.* (2016) 'Regulation of alternative splicing of Bcl-x by BC200 contributes to breast cancer pathogenesis', *Cell Death & Disease*, 7(6), pp. e2262–e2262. doi: 10.1038/cddis.2016.168.

Soemedi, R. *et al.* (2017) 'Pathogenic variants that alter protein code often disrupt splicing', *Nature Genetics*, 49(6), pp. 848–855. doi: 10.1038/ng.3837.

Song, Y. *et al.* (2017) 'Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation', *Molecular Cell*, 67(1), p. 148–161.e5. doi: 10.1016/j.molcel.2017.06.003.

Sorek, R. *et al.* (2004) 'A non-EST-based method for exon-skipping prediction.', *Genome research*. Cold Spring Harbor Laboratory Press, 14(8), pp. 1617–1623. doi: 10.1101/gr.2572604.

Sterne-Weiler, T. *et al.* (2018) 'Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop', *Molecular Cell*. Elsevier, 72(1), p. 187–200.e6. doi: 10.1016/j.molcel.2018.08.018.

Symoens, S. *et al.* (2011) 'A Novel Splice Variant in the N-propeptide of COL5A1 Causes an EDS Phenotype with Severe Kyphoscoliosis and Eye Involvement', *PLoS ONE*. Edited by F. Palau, 6(5), p. e20121. doi: 10.1371/journal.pone.0020121.

Sznajder, Ł. J. *et al.* (2018) 'Intron retention induced by microsatellite expansions as a disease biomarker', *Proceedings of the National Academy of Sciences*, 115(16), pp. 4234–4239. doi: 10.1073/pnas.1716617115.

Teraoka, S. N. *et al.* (1999) 'Splicing Defects in the Ataxia-Telangiectasia Gene, ATM: Underlying Mutations and Consequences', *The American Journal of Human Genetics*, 64(6), pp. 1617–1631. doi: 10.1086/302418.

- Tombácz, D. *et al.* (2018) 'Long-Read Sequencing Revealed an Extensive Transcript Complexity in Herpesviruses.', *Frontiers in genetics*. Frontiers Media SA, 9, p. 259. doi: 10.3389/fgene.2018.00259.
- Trapnell, C. *et al.* (2010) 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nature Biotechnology*, 28(5), pp. 511–515. doi: 10.1038/nbt.1621.
- Tripathi, V. *et al.* (2010) 'The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation', *Molecular Cell*, 39(6), pp. 925–938. doi: 10.1016/j.molcel.2010.08.011.
- Vaz-Drago, R., Custódio, N. and Carmo-Fonseca, M. (2017) 'Deep intronic mutations and human disease', *Human Genetics*. doi: 10.1007/s00439-017-1809-4.
- Volden, R. *et al.* (2018) 'Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 115(39), pp. 9726–9731. doi: 10.1073/pnas.1806447115.
- Wang, E. T. *et al.* (2008) 'Alternative isoform regulation in human tissue transcriptomes.', *Nature*, 456(7221), pp. 470–476. doi: 10.1038/nature07509.
- Wang, Z. *et al.* (2004) 'Systematic Identification and Analysis of Exonic Splicing Silencers', *Cell*. Cell Press, 119(6), pp. 831–845. doi: 10.1016/J.CELL.2004.11.010.
- Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: a revolutionary tool for transcriptomics', *Nature Reviews Genetics*, 10(1), pp. 57–63. doi: 10.1038/nrg2484.
- Wani, S. and Kuroyanagi, H. (2017) 'An emerging model organism *Caenorhabditis elegans* for alternative pre-mRNA processing *in vivo*', *Wiley Interdisciplinary Reviews: RNA*. Wiley-Blackwell, 8(6), p. e1428. doi: 10.1002/wrna.1428.
- Van De Water, N. S. *et al.* (2004) 'Factor IX polypyrimidine tract mutation analysis using mRNA from peripheral blood leukocytes', *Journal of Thrombosis and Haemostasis*, 2(11), pp. 2073–2075. doi: 10.1111/j.1538-7836.2004.00989.x.
- Weirather, J. L. *et al.* (2017) 'Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis', *F1000Research*, 6(1), p. 100. doi: 10.12688/f1000research.10571.2.
- Wu, J. *et al.* (2011) 'SpliceTrap: a method to quantify alternative splicing under single cellular conditions', *Bioinformatics*, 27(21), pp. 3010–3016. doi: 10.1093/bioinformatics/btr508.
- Wu, J. *et al.* (2013) 'OLego: Fast and sensitive mapping of spliced mRNA-Seq reads using small seeds', *Nucleic Acids Research*, 41(10), pp. 5149–5163. doi: 10.1093/nar/gkt216.
- Wu, T. D. and Nacu, S. (2010) 'Fast and SNP-tolerant detection of complex variants and splicing in short reads', *Bioinformatics*, 26(7), pp. 873–881. doi: 10.1093/bioinformatics/btq057.
- Xie, Y. *et al.* (2014) 'SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-

Seq reads', *Bioinformatics*, 30(12), pp. 1660–1666. doi: 10.1093/bioinformatics/btu077.

Xiong, H. Y. *et al.* (2015) 'RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease.', *Science (New York, N.Y.)*. NIH Public Access, 347(6218), p. 1254806. doi: 10.1126/science.1254806.

Xiong, H. Y., Barash, Y. and Frey, B. J. (2011) 'Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context', *Bioinformatics*, 27(18), pp. 2554–2562. doi: 10.1093/bioinformatics/btr444.

Yeo, G. W. *et al.* (2005) 'Identification and analysis of alternative splicing events conserved in human and mouse', *Proceedings of the National Academy of Sciences*, 102(8), pp. 2850–2855. doi: 10.1073/pnas.0409742102.

Yuan, J. *et al.* (2017) 'The MBNL3 splicing factor promotes hepatocellular carcinoma by increasing PXN expression through the alternative splicing of lncRNA-PXN-AS1', *Nature Cell Biology*. Nature Publishing Group, 19(7), pp. 820–832. doi: 10.1038/ncb3538.

Zeng, L. *et al.* (2013) 'A novel donor splice-site mutation of major intrinsic protein gene associated with congenital cataract in a Chinese family.', *Molecular vision*, 19, pp. 2244–9.