Use of genomic and transcriptomic approaches in the diagnosis of rare inherited disease linked to splicing mutations

A thesis submitted to the University of Manchester for the Degree of Doctor of Bioinformatics in the Faculty of Biology, Medicine and Health

Charlie F. Rowlands

2021

School of Biological Sciences

Division of Evolution and Genomic Sciences

Table of Contents

Word count: 43,051

List of figures

List of tables

Abbreviations

Abbreviation	Definition
5'/3'ss	5'/3' splice site
AA	amino acid
ACGS	Association for Clinical Genomic Science
ACMG	American College of Medical Genetics and Genomics
AS	alternative splicing
BPP	Branchpoint Predictor
BPS	branchpoint sequence
BWA	Burrows-Wheeler algorithm
DNA	deoxyribonucleic acid
eQTL	expression quantitative trait loci
ESE/ESS	exonic splicing enhancer/silencer
hPSC	human pluripotent stem cell
ISE/ISS	intronic splicing enhancer/silencer
MaPSy	massively parallel splicing assay
miRNA	microRNA
mRNA	messenger RNA
MRSD	minimum required sequencing depth
mtDNA	mitochondrial DNA
NF1	neurofibromatosis, type I
NMD	nonsense-mediated decay
OMIM	Online Mendelian Inheritance in Man
PCR	polymerase chain reaction
PPT	poly-pyrimidine tract
PTC	premature termination codon
PWM	position weight matrix
RBP	RNA-binding protein
RNA	ribonucleic acid
RP	retinitis pigmentosa
SBL	sequencing by ligation
SBS	sequencing by synthesis
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein
SNV	single nucleotide variant
sQTL	splicing quantitative trait loci
TAD	topologically associated domain
tRNA	transfer RNA
TSS	transcription start site
uORF	upstream open reading frame
UTR	untranslated region
VEP	Variant Effect Predictor

Abstract

The identification of pathogenic variants in Mendelian disease patients underpins disease management, genetic counselling and potentially treatment. Despite recent advances in next-generation sequencing (NGS), over half of Mendelian disease patients are unable to receive a molecular diagnosis for their disorders.

A growing body of evidence suggests that disruption of pre-mRNA splicing is an under-analysed cause of pathogenesis in Mendelian disease. Variants affecting conserved splicing motifs in mRNA transcripts can lead to pathogenic mis-splicing, whereby stretches of sequence are erroneously inserted or omitted from the canonical mRNA transcript. Recent years have seen a surge in the number of bioinformatics tools available to begin to predict the effect of these variants, and to analyse their effect in empirical functional assays.

However, much remains to be learned about the efficacy of these predictive tools, and the identification of mis-splicing events in empirical datasets, such as those derived from RNA sequencing (RNA-seq), remains in its infancy. Through deepening our understanding of these areas, there is promise to improve diagnostic yield for numerous cohorts of patients.

Here, I apply novel bioinformatics analyses at multiple stages along the process from variant identification to functional corroboration, with the aim of improving diagnostic yield and the quality of variant reporting. I identify an optimal strategy for predictive analysis of splicing impact in variants identified through upstream diagnostic testing, which reveals that the predictive tool SpliceAI provides the best accuracy in analysis of clinical variants impacting splicing. I further develop a bespoke approach for the investigation of a subset of splice-impacting variants impacting the intronic branchpoint sequence, resulting in the identification of a causative pathogenic variant in the *BBS1* gene. Finally, I develop a novel metric to guide the clinical integration of RNA-seq as a tool for investigating splice impact, which reveals disease- and tissue-specific use cases for RNA-seq in the investigation of mis-splicing.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given the University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables

iv. (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

v. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=2442 0), in any

Acknowledgements

This thesis could not have come to fruition without the support and guidance of my wonderful supervisory team. To Simon, Ray, Tracy and Jamie, I would like to offer my eternal gratitude: it has been a privilege to get to know you all as people and researchers, and you have been the most amazing sounding boards and sources of support, inspiration and advice. Myself and this thesis are infinitely richer for your expertise and contributions and for that I will be forever grateful.

To all my friends and colleagues at the Manchester Centre for Genomic Medicine, I would also like to extend my thanks for your knowledge and input. In particular, the support and friendship of Katie, Leslie and Cristina has been invaluable in keeping me level-headed and giving me perspective when riding the wild rollercoaster of PhD studies. To my wonderful PhD friends and peers - Alicia, Dale, Jason, Katherine, Katie, Chris, Luke and Matt - I have been endlessly inspired by your hard work and determination, and I am excited to watch you all flourish in the scary post-PhD world.

To all my friends in Manchester, both those since day one and earlier - Brad, Tom, Josh U, Victor, Viki and Katie S - or a little later - Lydia, Natalie, Deane, Amanda, Josh T, Will, Katie A, Hannah, Carla, Monika, Paul, Trixie and Katya - my life in Manchester is all the richer with you in it, and I could never have made it this far without you all. And to my beloved ResLife family - Grace, Zosia and Iwona - I couldn't have chosen better pals to be saddled with for my time in Whitworth Park. Thank you all for the laughs, and I'm so pleased to have made friends for life in each of you.

Last but not least, I owe everything to my amazing family, who have always been in my corner, supported me and kept me going through thick and thin. To Mum, Dad, Ellie and Grandma, this thesis is for all of you.

Rationale for journal format

The work presented here constitutes a set of three related analyses designing and investigating novel bioinformatics approaches to tackle diverse issues relating to the identification and interpretation of splice variants.

Chapter 2 of this thesis comprises a paper currently in-press at Scientific Reports, evaluating the efficacy of current splice prediction algorithms in identifying variants disruptive to transcript structure.

Chapter 3 of this thesis has not been submitted for publication, but, following collaboration with another lab group, has resulted in a publication in the BMJ Journal of Medical Genetics.

Chapter 4 of this thesis, which focuses on a novel predictive metric, is currently under review at the American Journal of Human Genetics.

The above results sections are each presented in a journal format.

Author contributions

Chapter 2 – Functional assays were performed by Dr. Jamie Ellingford. All other bioinformatics analyses were conducted by the author.

Chapter 3 – All bioinformatics analyses were conducted by the author; functional demonstration of branchpoint variant splice activity was carried out by a collaborative team, as documented in Fadaie et al. (2021)

Chapter 4 – All analyses conducted by the author.

Chapter 1

Introduction

1. Introduction

1.1. Principles of human inherited disease

Human disease can result from a wide range of environmental and genetic factors. While the onset of some diseases may theoretically be entirely environmentally driven, such as a heatstroke from prolonged exposure to the sun, or cirrhosis of the liver from chronic excessive alcohol consumption, many aspects of the response to and recovery from such diseases is governed by our genetics. One of the major agents of the inherited component of disease is the body of genomic variants harboured in the DNA of an individual. Large numbers of alterations, each with varying effect size and frequency in the human population, interact in a complex way with our environment to determine disease onset and severity. In reality, therefore, most forms of human disease can be seen as existing on an aetiological spectrum, where complex interplay between our genetics and environment govern the onset and progression of the condition (**Figure 1**).

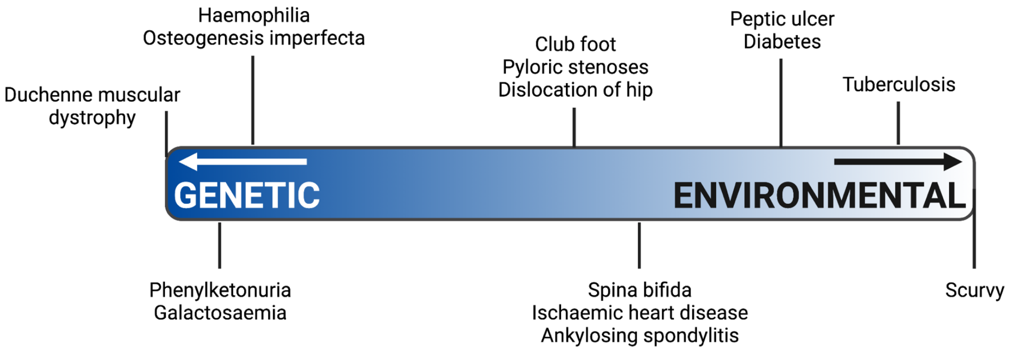


Figure 1. *The spectrum of human disease aetiology.* Many human diseases result from the complex interaction of environmental and genetic factors. On the genetic end of the spectrum, small, impactful changes lead to disease presentation independently of environment, although environmental factors may play a role in disease severity and progression. Conditions may also be predominantly caused by environmental factors, such as nutritional deficiencies; in some cases, such as in microbial infection, genetics may confer susceptibility or resistance to, and influence the recovery from, the condition, despite environmental acquisition. Adapted from (Turnpenny et al., 2017). Created using BioRender.com

Towards one extreme of this spectrum, however, exists a class of disease where widespread genetic variation of small effect size is superseded by small numbers of highly impactful variants, generally affecting one or two genes. These disorders – named Mendelian disorders – are not governed by the environment to the same extent as other diseases. Such conditions result from the inheritance of one or two pathogenic variants from parents, or the creation of *de novo* variants during fertilisation or embryonic development. While some Mendelian disorders may be effectively treated with dietary changes or pharmacological intervention, as in the case of the metabolic disorder phenylketonuria, which can be effectively managed with a low-protein diet (Al Hafid and Christodoulou, 2015), the variants underlying a Mendelian disease remain throughout the lifetime of the individual, and may still be passed down to their children.

So-called rare diseases may be loosely defined as any disease affecting fewer than one in 2,000 people in the general population. Rare Mendelian diseases, therefore, individually affect a fairly small number of people; however, the number of genetic diseases is large, with 7,897 phenotypes currently listed in the Online Mendelian Inheritance in Man (OMIM) database (https://omim.org/). It is therefore estimated that around 9.1% of individuals in the US will suffer from rare disease at some point in their lives (Haendel et al., 2020). These conditions are often associated with considerable physical distress and expensive cost of treatment, and many have a substantial impact on patient lifespan. There is therefore a moral and economic imperative to provide the best possible support for Mendelian disease patients and their families.

One of the primary goals in genetic medicine is the identification of the aforementioned genomic variants that underpin patient phenotype, which facilitates genetic counselling for family members and tailoring of treatment plans, and provides peace of mind for patients and their relatives. The journey to this *molecular diagnosis* is of variable length and ease; one factor contributing to this variability is disease subtype. In one study, for example, the screening of 105 genes associated with inherited retinal diseases provided a molecular diagnosis for approximately 51% (271/537) of Mendelian retinal disorder patients (Ellingford et al., 2016a), while another study demonstrated a lower diagnostic rate of 30% (665/2249) when screening a panel of 464 known disease-associated genes in patients with inherited neuromuscular disorders (Beecroft et al., 2020). The number of Mendelian patients lacking molecular diagnoses across all disease subtypes, however, remains significant.

There is thus an urgent unmet need for improvement in our ability to correctly identify the pathogenic variants underpinning patient phenotypes. To achieve this, it is useful to understand the diversity of the mutational landscape in the human genome.

1.1.1. Inheritance patterns

A key factor in evaluating the pathogenicity of a variant is the inheritance pattern of the disorder associated with the gene in which it is located. Mendelian disorders are broadly described as *dominant* if a single variant in a gene is sufficient for pathogenesis, or *recessive* if the presence of a pathogenic variant on both alleles of a gene is necessary for disease phenotype to occur.

Another consideration in the analysis of inheritance patterns is whether a condition is autosomal or X-linked: autosomally inherited disorders are caused by pathogenic variants in genes located on one of the 22 non-sex chromosomal pairs. Conversely, X-linked disorders are, as the name suggests, the result of pathogenic variation on genes located on the X chromosome. Autosomal disorders exhibit characteristic inheritance patterns, whereby dominant disorders are inherited by 50% of the offspring of an affected individual and do not skip generations, while autosomal recessive disorders affect, on average, 25% of offspring, with 50% being carriers of a pathogenic allele. In autosomal disorders, offspring of both sexes equally likely to inherit the pathogenic variant.

X-linked disorders are also inherited in a characteristic fashion, with male individuals being many times more likely to display an X-linked phenotype than females. A variant in a gene on the sole X chromosome of a male individual are described as being in a *hemizygous* state. The vast majority of X-linked diseases are recessive (Mehta et al., 2006), and so females require the inheritance of a pathogenic allele from both the mother and an affected father. The lower reproductive fitness of affected individuals largely accounts for the low rate of transmission of pathogenic X-linked alleles from fathers to daughters.

Although the majority of the human genome is located on chromosomes in the nucleus, the mitochondria of human cells also carry small circular chromosomes of around 16,569 bp in length

(Anderson et al., 1981). This so-called mitochondrial DNA (mtDNA) encodes a set of genes, including mitochondrial-specific ribosomal RNA (rRNA) and transfer RNA (tRNA) genes, as well as a set of 14 protein-coding genes, primarily encoding components of the mitochondrial electron transport chain (Taanman, 1999). Each mitochondrion carries between approximately 1 and 15 copies of the mitochondrial chromosome (Satoh and Kuroiwa, 1991), and, as such, mutations can accumulate independently on different copies of the mtDNA. This gives rise to the phenomenon of heteroplasmy, in which a single individual can harbour large numbers of different mtDNA variants at varying frequencies. The degree of heteroplasmy, i.e. the prevalence of a particular variant within the mtDNA population, can sometimes determine the severity of patient phenotype (Stewart and Chinnery, 2015). Due to the function of mitochondria in energy, disorders associated with variants in mtDNA, such as mitochondrial myopathy, encephalopathy, lactic acidosis and stroke-like episodes (MELAS, (Goto et al., 1992)) and myoclonic epilepsy with ragged red fibres (MERRF, (Shoffner et al., 1990)), often affect tissues with high energy consumption, such as muscle and neural cells. Crucially, an embryo's mitochondria are derived solely from those in the original maternal oocyte. As a result, disorders resulting from variants in mtDNA display a characteristic pattern of *matrilineal inheritance*, in which the disorder is passed from affected female sufferers to all of their offspring, while male sufferers are unable to pass the trait down (Stewart and Chinnery, 2015).

1.1.2. Origins and consequences of genetic variation

Genomic variation typically results from errors that occur during the replication of DNA (Pray, 2008, Tippin et al., 2004). Variants can be broadly divided at the structural level into three categories: *single nucleotide variants*, or *SNVs*, result when the identity of a single base is changed. SNVs can be further sub-divided into *transitions*, in which the substituted base is of the same biochemical class (i.e. a purine is substituted for a purine, or a pyrimidine is substituted for a pyrimidine), and *transversions*, in which a purine is replaced with a pyrimidine, or vice versa. The second category, *insertions-deletions*, or *indels*, result from the addition or removal of small stretches of nucleotides at a genomic site. The final category, *structural variants*, constitutes a wide variety of large-scale structural changes. These include, but are not limited to, translocation of genomic sequences within or between chromosomes (**Figure 2a**) and inversions of genomic sequence (**Figure 2b**). One

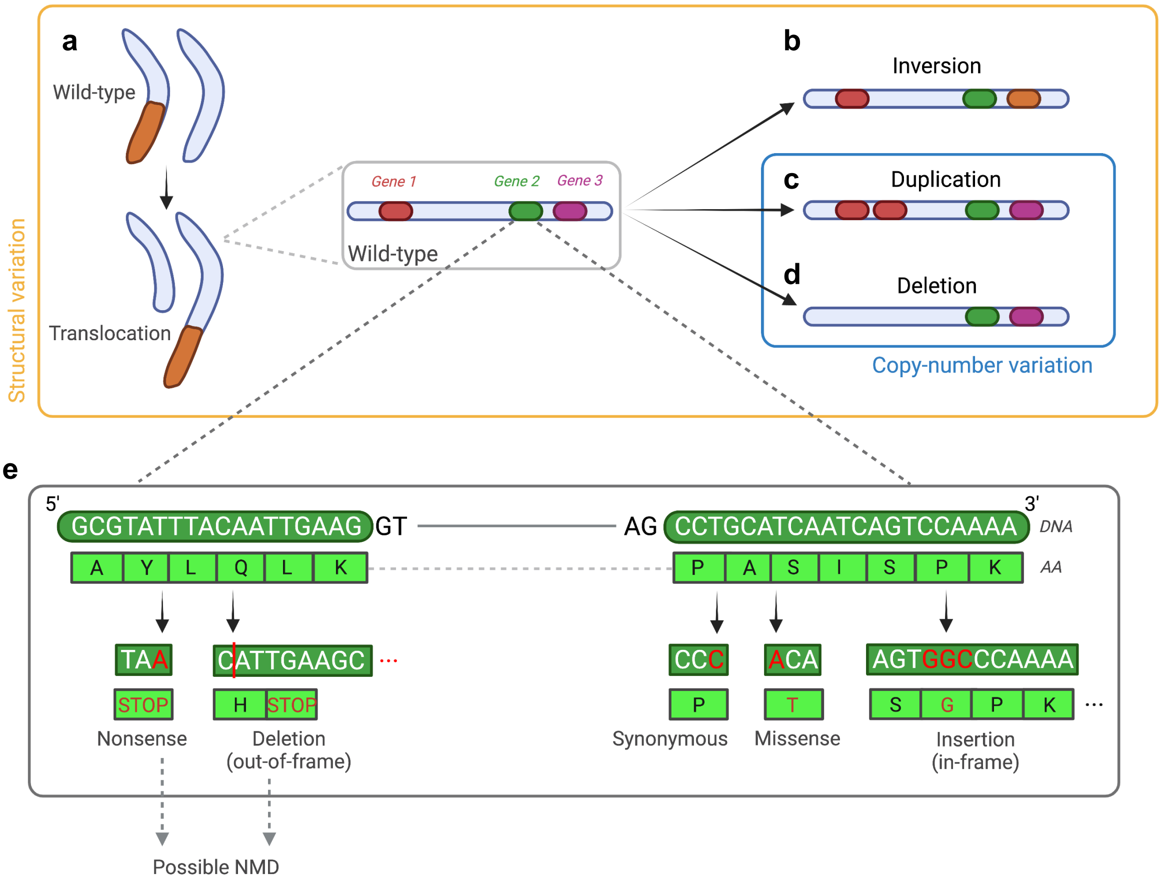


Figure 2. *Illustration of exemplar structural and coding variant types and their consequences in Mendelian disease*. At the macromolecular scale (**top**), structural variation can lead to gross inter-chromosomal events such as (**a**) translocations, while other structural variants may result in (**b**) inversions, or copy-number variants such as (**c**) duplications or (**d**) deletions. At the nucleotide level (**e**), the introduction of many different non-structural variant types (red) can lead to the perturbation of transcripts. In-frame indels lead to the omission or inclusion of whole numbers of amino acids (AAs), while out-of-frame events cause wide disruption affecting all AAs downstream of the variant, often resulting in the introduction of premature termination codons. Nonsense variants and out-of-frame indels are also liable to trigger the nonsense-mediated decay (NMD) cellular surveillance pathway and be degraded. Created with BioRender.com.

subset of structural variant is the *copy number variant*, which results from the duplication (**Figure 2c**) or deletion (**Figure 2d**) of genomic regions, typically whole exons or genes, thus leading to the namesake change in the number of copies of that region.

One less clinically studied source of genetic variation is that of transposable element insertion. Transposable elements are short sequences of DNA that can be excised from their site in the genome via the action of transposase enzymes, or copied at the RNA stage by reverse transcriptase, and inserted elsewhere in the genome (Pray, 2008). Disruptive insertion of transposons within exonic or regulatory regions has been observed as a rare cause of Mendelian disease (Holmes et al., 1994, Yoshida et al., 1998).

In extremely rare cases, catastrophic chromosomal shattering and its subsequent repair may result in numerous gross changes to chromosome structure. This phenomenon, *chromothripsis*, generates a chromosome that appears to harbour large numbers (sometimes in the order of hundreds to thousands) of concurrent structural rearrangements. Chromothripsis is emerging as a common feature in up to 50% of cancer types (Cortés-Ciriano et al., 2020, Forment et al., 2012).

1.1.3. From genotype to phenotype: mechanisms of pathogenic genetic variation

The impact of a variant on human cellular and organismal function is dependent on a number of factors, including the type of variant and its location within the genome. Variants may be defined in relation to their effect on the function of the respective gene product: variants that reduce the ability of the gene product to carry out its role are considered *loss-of-function,* and can be subdivided into *null variants*, where the function of the gene product is entirely or almost entirely compromised, and *hypomorphic variants*, where there is still residual gene product function despite some reduction in gene product function. *Gain of function* variants*,* conversely, result in the gene product acquiring some novel function, e.g. a novel binding partner or enzymatic activity, such that disease presentation results.

Variants can also broadly be divided into two groups dependent on their genomic region and consequence: *coding* and *non-coding*, both of which have the potential to be pathogenic. Some of the key characteristics of both are summarised below and illustrated in **Figure 2e**.

1.1.3.1. Coding variation as a cause of human disease

Variants are considered *coding variants* if they directly alter the protein sequence encoded by a gene; such variants are necessarily exonic. Variants that lead to a direct change in the identity of the amino acid encoded by their respective codon are termed *missense variants*. Missense variants may act to destabilise protein structure or impair key residues required for the function of the protein, such as enzymatic active sites. Changes to amino acid sequence are often the result of SNVs, but may also be caused by indels, which can lead to the merging of distinct codon boundaries, as well as the insertion or deletion of a number of amino acids. Missense variants may be loss-of-function or gain-of-function, dependent on whether the amino acid change impairs protein function or leads to novel functionalities, respectively.

Coding variants may also impair protein function through the introduction of *premature termination codons* (*PTC*s); such variants are termed *nonsense variants.* Nonsense variants are frequently null, as they result in the truncation of the resulting protein; truncated proteins are often non-functional due to the exclusion of critical domains from the final protein product. PTCs are often detected by a cellular surveillance mechanism called *nonsense-mediated decay*, or *NMD*, which targets transcripts with stop codons outside the final exon for degradation. Thus, nonsense variants can doubly abrogate protein function and reduce the number of transcripts available for translation.

Nonsense variants may be generated through SNVs via the direct mutation of an amino acid-encoding codon into a stop codon. However, indels are also a cause of PTCs: when the number of inserted or deleted bases is not a multiple of three, the boundaries between codons are shifted by one or two positions (termed a *frameshift*), resulting in gross changes to amino acid sequence at and downstream of the indel. This generally results in the introduction of a PTC (that was formerly out-of-frame) at some distance from the indel. Frameshifts and resultant PTCs are also a common consequence of splicing variants (see *1.2.5.*).

Two other rare coding variant types are *start-loss* and *stop-loss* variants. In start-loss, variants impact the first codon of the transcript, universally encoding a methionine residue in humans. In the absence of a functional initial codon, translation initiation is perturbed; resulting proteins may make use of alternative nearby start sites, or remain untranslated (Binder et al., 2003, Sargiannidou et al., 2015). Accurate interpretation of start-loss variants requires the careful analysis of which of these two consequences predominates for a given variant. Conversely, stop-loss variants instead disrupt the function of the final codon of the transcript, leading to extension of the protein at the 3' terminus. This may lead to the use of a downstream compensatory stop codon (Riedhammer et al., 2021), but may also result in activation of a surveillance pathway known as *non-stop decay*, which targets transcripts lacking stop codons for degradation. A common stop-loss variant in the human *DEF126B* gene, for example, leads to nonstop decay of the respective transcript and is associated with impaired fertility (Tollner et al., 2011).

1.1.3.2. Non-coding variation in protein-coding genes as a cause of Mendelian disease

Protein-coding regions account for just 1.5-2% of the human genome (Litwack, 2018, Mattick, 2001). The remainder of genomic space is described as *non-coding*, and pathogenic variation in non-coding regions of the genome is becoming increasingly appreciated as a cause of Mendelian disease. Here, we provide a non-exhaustive list of some of the most common pathogenic mechanisms of non-coding variation.

Upstream of the *transcription start site* (*TSS*) of all protein-coding genes is a *promoter* sequence, a region of genomic sequence which serves as a binding site for transcription initiation complexes. One of the most common core promoter elements is the TATA box, a short sequence located around 24-30 bp upstream of many metazoan TSSs (Andersson and Sandelin, 2020). As promoters are indispensable for transcription initiation, significant perturbation of promoter function results in large decreases in, or absence of, transcript expression (**Figure 3a**). This disruption may be caused by SNVs, as in the case of the promoter of the *TERT* gene in inherited telomerase deficiency and some cancers (Gutierrez-Rodrigues et al., 2019). However, promoter disruption is also often the result of larger deletions that encompass a significant amount of their length, and has been observed as a pathogenic mechanism in diverse disease subtypes, such as in the developmental disorder Liebenberg syndrome (Kragesteen et al., 2019), the tumorigenic disorder Cowden syndrome (Zhou et al., 2003), and sickle cell anaemia (Chaouch et al., 2020).

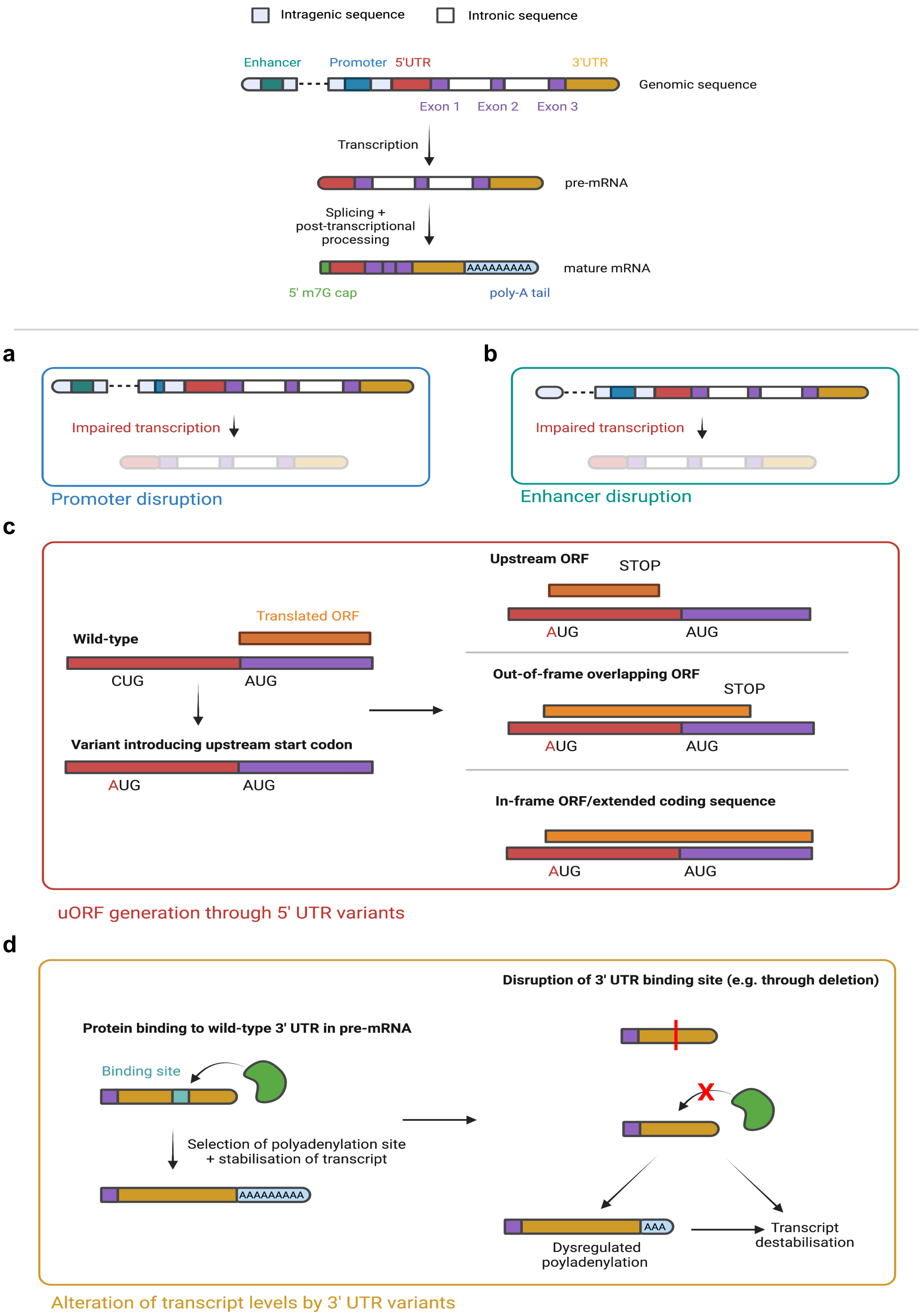

Intragenic sequence
Intronic sequence
Enhancer
Promoter
5'UTR
3'UTR
Genomic sequence
Exon 1
Exon 2
Exon 3
Transcription
pre-mRNA
Splicing +
post-transcriptional
processing
AAAAAAAAA
mature mRNA
5' m7G cap
poly-A tail
a
Impaired transcription
Promoter disruption
b
Impaired transcription
Enhancer disruption
c
Translated ORF
Wild-type
CUG
AUG
Variant introducing upstream start codon
AUG
AUG
Upstream ORF
STOP
AUG
AUG
Out-of-frame overlapping ORF
STOP
AUG
AUG
In-frame ORF/extended coding sequence
AUG
AUG
uORF generation through 5' UTR variants
d
Protein binding to wild-type 3' UTR in pre-mRNA
Binding site
Selection of polyadenylation site
+ stabilisation of transcript
AAAAAAAAA
Disruption of 3' UTR binding site (e.g. through deletion)
X
AAA
Dysregulated
poyladenylation
Transcript
destabilisation
Alteration of transcript levels by 3' UTR variants

Figure 3. *Illustration of the pathogenic mechanisms of exemplar non-coding variant types.* Non-coding variants do not directly disrupt protein structure, but rather influence other aspects of transcript dynamics. Loss-of-function variants at (**a**) promoter and (**b**) enhancer sites result in attenuation of the transcription of their target genes. (**c**) Variants creating novel transcription start sites in the 5' UTR of human transcripts are liable to generate upstream open reading frames (uORFs) that outcompete canonical transcription start sites and lead to a reduction in the levels of functional gene product. (**d**) Loss-of-function variants in the 3' UTR, such as deletions, may also impair the ability of proteins involved in RNA processing and stability to bind, thus leading to dysregulation of transcript dynamics, often resulting in an overall reduction in transcript levels. Created using BioRender.com

A related non-coding element with roles in transcription initiation is the *enhancer*, a regulatory sequence that typically interacts with the promoter through changes in chromosomal conformation. Unlike promoters, enhancers can act at great distances from their target genes, with some lying hundreds of kilobases upstream or downstream of the TSS (Karnuta and Scacheri, 2018). While some genes are associated with no enhancer elements, others are served by multiple, and a single enhancer may be associated with numerous genes, thus creating an intricate regulatory network of gene expression. Pathogenic variation affecting enhancer elements may be difficult to identify, owing to the large distance between many enhancers and their target genes. Additionally, our incomplete knowledge of enhancer distribution and targets in the human genome limits our ability to accurately annotate variants as overlapping enhancer loci. Similarly to promoters, SNVs and structural variants are both liable to disrupt enhancer function (**Figure 3b**), resulting in decreased or abrogated transcription of the target transcript, and have been identified across a broad spectrum of phenotypes: pathogenic SNVs in enhancer elements have been identified in such disease subtypes as pancreatic agenesis (Gabbay et al., 2017) and aniridia (Bhatia et al., 2013), while structural variants such as deletions and duplications at enhancer loci have been identified in, for example, inherited deafness (Naranjo et al., 2010) and disorders of sex development (Erickson et al., 2011, Hyon et al., 2015), respectively.

One non-coding region increasingly recognised as a source of pathogenic variation is the 5' untranslated region, or 5' UTR. The 5' UTR lies immediately upstream of the TSS in human genes, occasionally overlapping with promoter regions (Alexandrova et al., 2012), and is the site of ribosomal entry during translation. An emerging source of pathogenicity in 5' UTR regions is the creation of upstream open reading frames, or *uORFs* (Whiffin et al., 2020; **Figure 3c**): these are created by the aberrant generation of initiation codons upstream of the canonical TSS, and may be recognised by the ribosomal machinery as genuine reading frames, leading to the production of proteins encoding by canonically non-coding residues. This, in turns, leads to reduction in the production of the encoded wild-type protein by up to 80% (Calvo et al., 2009). Pathogenic uORFs may lie entirely upstream of the TSS (if an in-frame stop codon is present in the 5’ UTR sequence downstream of the novel initiation codon), or may overlap it. In the latter case, novel initiation codons that are out-of-frame in relation to the canonical TSS will likely result in PTCs due to frameshift, and will generally be degraded through NMD. In-frame uORFs will lead to an effective extension of the coding sequence of the transcript, and may impair protein structure and function (Whiffin et al., 2020). Interpreting the pathogenicity of suspect uORF-creating requires functional investigation to identify the nature and extent of the disruption. Pathogenic uORF-creating variants have been identified in cases of disorders such as van der Woude syndrome (de Lima et al., 2009) and neurofibromatosis type 1 (Evans et al., 2016, Whiffin et al., 2020).

At the opposite end of the transcript, immediately downstream of the termination codon, lies the 3’ UTR. The 3’ UTR is a crucial regulatory sequence that serves primarily as a binding site for mRNA processing proteins and microRNAs (miRNAs), and plays a role in the alternative polyadenylation of transcripts. Following transcription, a host of RNA-binding proteins (RBPs) may bind to stabilise the transcript or regulate its translation; thus, variants that binding sites within the 3’ UTR are liable to lead to dysregulation of transcripts (**Figure 3d**), often resulting in their decay (Pamuła-Piłat et al., 2020). Few pathogenic 3’ UTR variants have thus far been identified in Mendelian disorders; however, recent research by Griesemer et al. (2021) implemented a massively parallel assay to highlight a common SNP in the *PILRB* gene as a contributor to age-related macular degeneration.

Non-coding variants are also liable to disrupt splicing, the process in which non-coding intronic regions are removed, generally co-transcriptionally, from nascent pre-mRNA transcripts. Unlike the above variant types, splice-impacting variants can arise almost anywhere within a gene, including intronic regions. A more detailed description of splicing mechanisms, and the role of their dysregulation in pathogenesis, is given below.

1.1.3.3. Splice variants: an under-analysed mechanism of pathogenicity?

Recent years have seen a gradual shift in the focus of genetic diagnostics from the analysis primarily of coding variants to the non-coding variants. While our understanding of many non-coding variant types (and how to interpret them) is limited, research into unsolved Mendelian disease cases has begun to shed light on some of the sources of pathogenic variation that are currently under-represented in clinical diagnostics. One of the most promising such sources is those variants that impact pre-mRNA splicing: with several studies demonstrating that between 7.5-35% of unsolved Mendelian disease cases may be attributable to pathogenic splice-impacting variants (Cummings et al., 2017, Frésard et al., 2019, Kremer et al., 2017). There remain, however, many barriers to effective identification and interpretation of such variants. To overcome these obstacles, it is necessary to appreciate the mechanistic origins of splicing

1.2. pre-mRNA splicing

The excision of introns from pre-mRNA transcripts, termed *splicing*, is a tightly regulated process involving the co-ordinated activity of numerous *cis*- and *trans*-acting factors. The dysregulation of any number of these factors can result in the disruption of transcript isoform structure, and so lead to defects at the protein level that can underpin Mendelian disease. Below are described some of the key factors involved in splicing, their role in splicing biochemistry, and how disruption of these factors may play a role in pathogenesis.

1.2.1. Cis-acting sequence elements in the regulation of splicing

The boundary between introns and exons is demarcated by a host of nucleotide sequence elements that serve to guide *trans*-acting protein complexes to the correct regions of the transcript for the initiation and progression of splicing. As will be discussed below, these elements are subject to pathogenic disruption that results in gross changes to transcript structure. Major *cis*-acting sequences involved in splicing are illustrated in **Figure 4**.

Some of the best-understood sequence features in intron-exon definition are the so-called *splice donors* and *splice acceptors*, a pair of dinucleotides that constitute the 5'- and 3'-most dinucleotides of the intron, respectively. The majority – approximately 99.0% – of human introns are flanked by a GT splice donor and an AG acceptor (Sheth et al., 2006). Although these *GT-AG* introns are the most abundant in the human genome, alternative donor-acceptor dinucleotide pairs are also present, with GC-AG being the next-most common, accounting for 0.86% of introns (Sheth et al., 2006). The remaining introns are flanked by various rarer dinucleotide pairs, including AT-AC, GT-AT and GT-TG.

These dinucleotides are essential for exon definition, and so are often described as *core* splicing dinucleotides; however, other positions close to these dinucleotides can also be highly constrained. The final base of an exon, for example, often described as the donor -1 position, is often constrained to be a guanine, as is the fifth base of the intron (the *donor* +5 position; Lord et al., 2019). Such measurements of constraint can be calculated using mutability-adjusted probability of singletons, or MAPS, scores, a measure of the selection acting at a particular nucleotide position (Lek et al., 2016). Positions in the vicinity of the splice acceptor site tend to be less constrained than those adjacent to the donor. This constraint has emerged under significant selective pressure to allow splice sites to be recognised by *trans*-acting splicing complexes, and so variation at these constrained positions may be more likely to impact splicing than those at other positions (Lee et al., 1991).

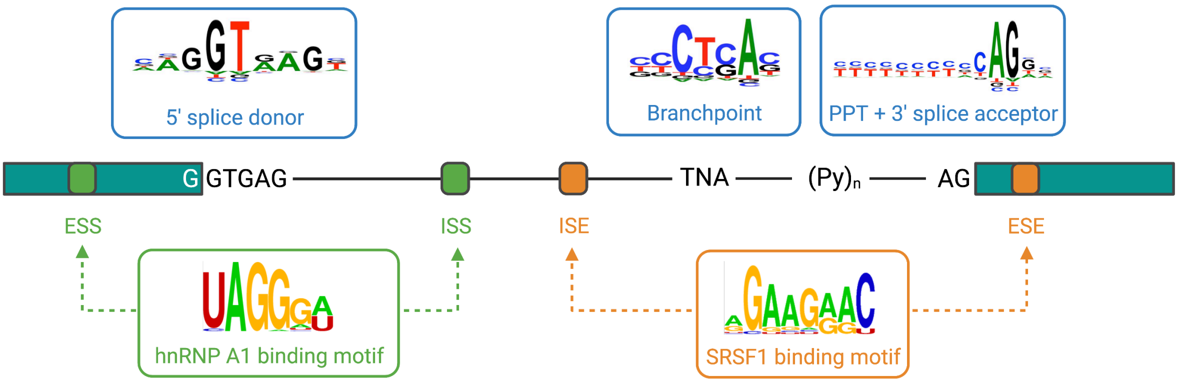


Figure 4. Cis-*acting splice elements are diverse and bind numerous consensus motifs.* Human introns are flanked by two highly conserved nucleotide pairs consisting usually of a 5’ GT and 3’ AG. Upstream of the splice acceptor is a polypyrimidine-rich region known as the polypyrimidine tract (PPT). Exonic and intronic splicing silencers (ESSs and ISSs) are often bound by members of the hnRNP family, such as hnRNP A1, depicted here, while exonic and intronic splicing enhancers (ESEs and ISEs) are often bound by SR protein family members, as exemplified here by SRSF1. The position weight matrices (PWMs) of enhancer and silencer elements show marked consistency across their length, while donor, branchpoint and acceptor positions exhibit greater redundancy outside of critical residues. 5’ splice donor, branchpoint and PPT/3’ splice acceptor PWMs were adapted from Desmet et al., (2009); PWMs for hnRNP A1 and SRSF1 were identified through the CISBP-RNA motif repository (Ray et al., 2013). Py, pyrimidine. Created using BioRender.com

Despite the lower constraint in the immediate vicinity of most splice acceptors, there is an extended region of shared structure common to the 3’ end of human introns. Most notably, the region upstream of the splice acceptor is termed the *poly-pyrimidine tract*, or *PPT*, and is characterised by a stretch of cytidine- and uridine-rich sequence that serves as the binding site for the splicing protein U2AF65 (Will and Lührmann, 2011; see *1.2.2.*) Most PPTs are around 10-20 bp in length, and the precise composition of the PPT has been demonstrated to influence splicing efficiency: *in vitro* assays have shown that a poly-uridine tract of around 13 bases produces the most efficient progression through the early stages of splicing (Coolidge et al., 1997).

Upstream of the PPT lies a short consensus sequence named the *branchpoint sequence*, or *BPS*. The BPS plays a key role in the splicing reaction by providing a nucleophilic residue that attacks the donor site and facilitates the adjoining of the two ends of the flanking exons (see *1.2.2.*). As with splice donors and acceptors, there exists a canonical central BPS, which consists of a TNA motif (where N is any nucleotide). Recent computational work has suggested a putative extended BPS motif of TRYTRAY (Taggart et al., 2017). Growing evidence has also suggested that branchpoint selection is an important characteristic of tissue-specific splicing diversity (Pineda and Bradley, 2018). Thus, pathogenic variation at the BPS may require wider consideration of the tissue(s) affected according to the patient phenotype, and functional assays of the impact of these variants may require use of the relevant cell type, where possible.

Another diverse set of sequence elements also serve to modulate the efficiency of the splicing reaction. These are termed splicing *enhancers* and *silencers*, depending on whether they facilitate or attenuate the splicing reaction, respectively, and may be either exonic or intronic. These elements constitute a diverse set of binding sites for RNA-binding proteins (RBPs), which serve as regulators of splicing at nearby junctions (Fredericks et al., 2015; see *1.2.2.*). Key splice enhancer-binding proteins include many members of the SR protein family (Cavaloc et al., 1999, Liu et al., 2000), while silencers are often bound by members of the hnRNP family (Rothrock et al., 2005, Zhu et al., 2001). However, more recent research has shown that the effect of these splice element-binding RBPs is at least partially dependent on where the element is located (Erkelenz et al., 2013). Unlike the above ubiquitous sequence features, not all introns are necessarily associated with splicing enhancers and/or silencers.

1.2.2. The major spliceosome: the primary *trans*-acting engine of intron excision

The reaction central to pre-mRNA splicing is a transesterification reaction that adjoins the two exon ends, resulting in the excision of the central intron. This two-step reaction is primarily carried out in human cells by a massive macromolecular complex named the *major spliceosome*, sometimes called the *U2-dependent spliceosome*. The spliceosome consists of a set of *core machinery* that is directly responsible for the biochemical progression of the splicing reaction; this comprises a set of five small nuclear ribonucleoproteins (snRNPs), named U1, U2, U4, U5 and U6, that are formed from the binding of a protein factor to a corresponding small nuclear RNA (snRNA). These snRNPs dynamically bind and unbind across the course of the splicing process. However, over 200 accessory proteins interact with the core spliceosomal machinery over the course of the splicing reaction, making the spliceosome one of the largest macromolecular complexes in human cells (Cvitkovic and Jurica, 2013, Schneider et al., 2002). The spliceosome shows remarkable conservation across distant eukaryotic lineages: of ~90 proteins identified as major components of the yeast spliceosome, approximately 85% had homologues in the human genome, and the process of spliceosomal assembly is almost identical in *S. cerevisiae* and humans (Fabrizio et al., 2009).

In the earliest stages of spliceosomal assembly, the 5’ splice site is recognised by the U1 complex to form the so-called *early complex*, or *complex E*. The 3’ splice site is also bound by subunits of the U2 snRNP (**Figure 5a**). An ATP-dependent remodelling of the complex then facilitates both recognition of the branchpoint by the U2 snRNP, as well as an interaction between U2 and the E complex in an ATP-dependent manner to form the so-called *pre-spliceosome*, or *complex A* (**Figure 5b**). Crucially, the base pairing between the U2 snRNA and the branchpoint is imperfect, and leads to a bulging of the branchpoint adenosine that primes it to carry out the downstream transesterification reaction.

Following this, complex A is bound by a so-called tri-snRNP, consisting of a complex of the U4, U5 and U6 snRNPs to form complex B (**Figure 5c**), which is converted to the catalytically active complex B* following conformational changes that lead to disassociation of the U1 and U4 snRNPs (**Figure 5d**). In the first step of the central transesterification reaction, the bulged branchpoint adenosine, associated with complex B*, acts as a nucleophile and attacks the end of the first exon. This forms the catalytic *complex C* (**Figure 5e**), comprising the remaining snRNPs bound to both the exposed 5’ exon end, plus the 3’ exon end with the intron attached as a *lariat* structure. In the second step of the transesterification reaction, the exposed 5’ exon end attacks the 3’ end of the intron to form a covalent bond between the ends of both exons, releasing the intron as a lariat that is quickly degraded (**Figure 5f**; Moore et al., 2002).

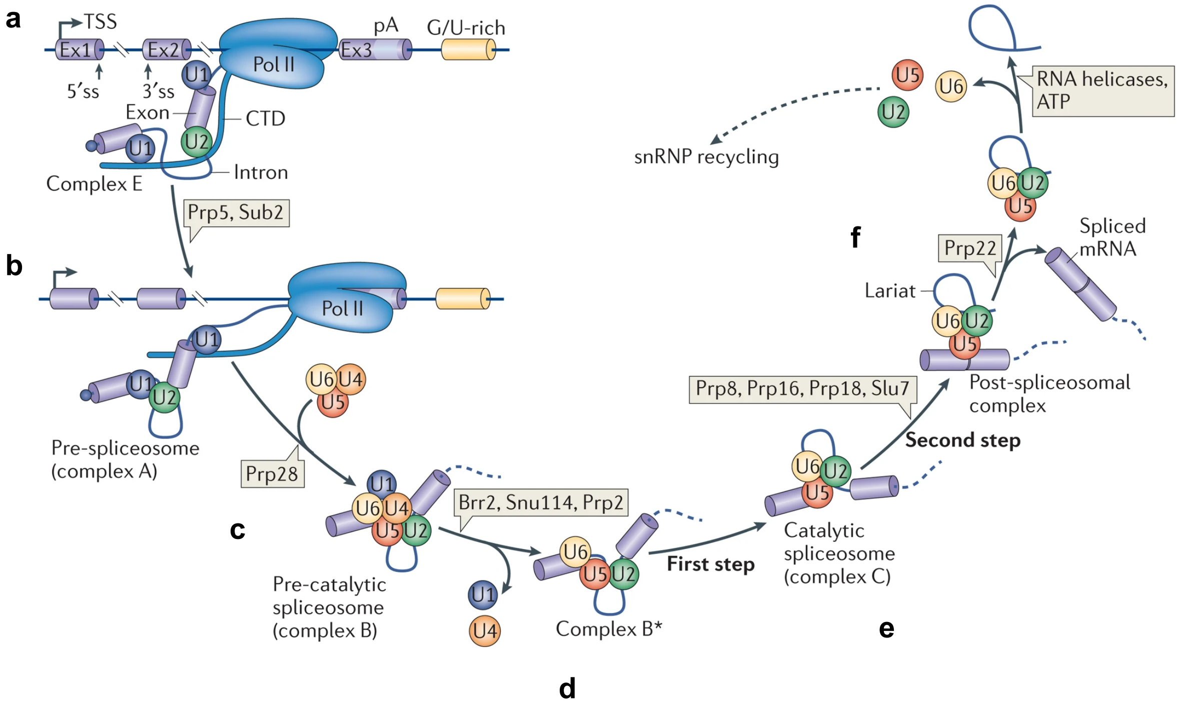


Figure 5. *The stepwise assembly of the spliceosome excises introns from pre-mRNA transcripts.* (a). In the earliest stages of spliceosomal assembly, the 5’ and 3’ splice sites of the intron are bound by the U1 and U2 small nuclear ribonucleoproteins (snRNPs), respectively to form the so-called early complex, or complex E. (b) ATP-dependent remodelling of complex E brings the U1 and U2 snRNPs into contact, forming the pre-spliceosome, or complex A. (**c**) Binding of the U4/U6.U5 tri-snRNP to complex A generates the pre-catalytic spliceosome, complex B. (**d**) Conformational changes then form the catalytically active complex B*, accompanied by the release of U1 and U4. (**e**) The transesterification reaction then takes place, resulting in complex C, consisting of the U2, U5 and U6 complexes bound to the new-free first exon and an intron-exon 2 lariat intermediate. (**f**) The splicing reaction concludes with a second transesterification reaction, which results in the joining of the exon ends and subsequent release of the U2, U5 and U6 snRNPs and the intronic lariat, which is rapidly degraded. Adapted from Matera and Wang (2014).

Each stage of spliceosomal assembly is facilitated by the involvement of accessory proteins with diverse roles. Key auxiliary proteins include members of the aforementioned SR family, which stabilise interactions between U1 and the 5’ splice site (5’ss), as well as between U2 and the 3’ss (Cho et al., 2011, Staknis and Reed, 1994). The activity of helicases is also abundant at various points in spliceosomal remodelling: the helicase Prp28, for example, has been shown in yeast to facilitate dissociation of the U1 snRNP and recruitment of the U4/U6.U5 tri-snRNP (Yeh et al., 2021), while the helicase Prp5 is believed to be involved in proofreading of the U2 snRNA-BPS interaction (Zhang et al., 2021). Other auxiliary proteins, such as the U2 auxiliary proteins U2AF35 and U2AF65, are also vital in spliceosome assembly and function (Matera and Wang, 2014).

1.2.3. The minor spliceosome: an alternative biochemistry for pre-mRNA splicing

Although approximately 99.6% of introns are excised by the major spliceosome (Sheth et al., 2006), a secondary machinery, termed the *minor,* or *U12-dependent spliceosome*, is responsible for the correct splicing of the remainder of human introns, numbering approximately 700 across the human genome (Alioto, 2007, Levine and Durbin, 2001, Sheth et al., 2006). The core components vary significantly between the major and minor spliceosomes, with the major U1, U2, U4 and U6 snRNPs being replaced by snRNPs named U11, U12, U4atac and U6atac, respectively, in the minor complex. The U5 snRNA, however, is shared between both spliceosomes. U12-dependent introns (i.e. those spliced by the minor spliceosome) are characterised by distinct splice site and branchpoint motifs. Similarly, the majority of auxiliary splicing regulatory proteins are shared between both spliceosomes, suggesting the cognate minor spliceosome snRNPs function similarly to their major counterparts. While the major spliceosome primarily excises GT-AG introns, the minor spliceosome also processes a significant proportion of AT-AC introns, which account for 25.2% of U12-dependent introns (Sheth et al., 2006).

1.2.4. Sources of transcript diversity in pre-mRNA splicing

One of the primary functions of pre-mRNA splicing is to provide a source of transcript diversity such that a single gene may produce multiple isoforms of the final gene product: by providing a dynamically processed RNA template, as opposed to a single, fixed transcript for each gene, organisms can adapt the structure of the final transcript, and thus the gene product, in response to external factors. Such factors may include developmental stage: recent research in mice, for instance, has demonstrated the existence of thousands of previously unannotated transcript isoforms unique to embryonic stages (Qiao et al., 2020), while age-related changes to transcript isoforms are becoming increasingly well-characterised in humans (Ham and Lee, 2020, Wang et al., 2018). Cell signalling cascades may also result in preferential generation of specific transcript isoforms, as in the case of the preferential selection of an alternative fifth exon in transcripts of the glycoprotein CD44 in response to Ras signalling (Cheng et al., 2006).

This *alternative splicing* (*AS*) can take many forms: in one form, the deliberate skipping of exons can result in functionally distinct gene products, as seen in the *FMR1* gene, where selective exclusion of exon 14 diverts the localisation of the *FMR1* protein from the cytoplasm to the nucleus (Sittler et al., 1996), while regulated skipping of exon 6 in one isoform of the *FAS* cell surface death receptor excludes a transmembrane domain, producing a soluble isoform that can inhibit *FAS* signalling (Cascino et al., 1995).

Relatedly, pre-mRNA transcripts may contain mutually exclusive exon pairs, in which only one of two or more exons may be present in a given transcript. This is the case in fibroblast growth factor receptor 2 (*FGFR2*), for instance, in which the guided inclusion of an alternative eighth exon by the splicing regulatory protein *ESRP1* directs the epithelial-mesenchymal transition in embryonic development (Ranieri et al., 2016).

Transcripts may make use of multiple donor or acceptor sites for the same exon, resulting in the exclusion or inclusion of small numbers of bases. In a study that mapped proteomics data back to the transcriptome, Rodriguez et al. (2020) demonstrated that a small handful of alternative 5’ and 3’ splice site events constituted a total of ~10% of the tissue-specific splicing events across ten tissues, and were demonstrated to have evolved up to 400 million years ago, demonstrating their crucial role in their respective tissues.

In one mode of AS, introns can escape excision from the mature mRNA transcript, a phenomenon named *intron retention*. Although well-studied in organisms like the model plant *Arabidopsis thaliana*, in which it is a pervasive AS mechanism (Filichkin et al., 2010), the study of intron retention in humans has somewhat trailed that of many other AS types. In recent years, however, research has begun to identify regulated regimes of intron retention as key players across a wide variety of human biological processes, including haematopoiesis (Edwards et al., 2016, Wong et al.,

2013) and the differentiation of germ cells (Naro et al., 2017). These discoveries have been aided by a recent surge in the bioinformatics tools available to identify them, such as IRFinder (Middleton et al., 2017) and rMATS (Shen et al., 2014).

In an ostensibly rare AS phenomenon, named *exitron* splicing, internal stretches of long exonic sequences may be alternatively spliced out of transcripts. While still quite poorly understood, some research has suggested that exitron splicing has a role in proteome plasticity in response to factors such as stress and carcinogenesis, even in genes typically considered to be single-exon (Marquez et al., 2015).

1.2.5. The role of pathogenic mis-splicing in human disease

Given the intricacy of the splicing process, it is perhaps unsurprising that disruption to a number of the *cis*- and *trans*-acting splicing components is liable to have a substantial impact on the final transcript.

The blanket term of *splicing quantitative trait loci* (*sQTLs*) refers to the cohort of genomic variants that are known to have an impact on splicing in their respective transcripts. sQTLs comprise a diverse set of loci in terms of both effect size and location: while some sQTLs possess a high effect size, such as those at the canonical dinucleotides, many lines of recent research have demonstrated that the majority of splice-impacting variants only subtly modulate the splicing process (Garrido-Martín et al., 2021; GTEx Consortium, 2020). This small individual effect size means that many sQTLs are only identified through large-scale analyses of transcriptomic datasets (Garrido-Martín et al., 2021, GTEx Consortium, 2020, Takata et al., 2017). Due to the impact of altered transcript structure on gene expression levels (for example, through NMD), many sQTLs are also identified as expression quantitative trait loci (eQTLs), sites at which variants lead to changes in expression of the respective transcript; recent work has shown that 52% of sQTLs identified in the study overlapped with previously annotated eQTLs (Garrido-Martín et al., 2021).

As in wild-type splicing, pathogenic mis-splicing events can take a number of forms, including exon skipping, alternative splice sites, and intron retention. Both abrogation of normal splice site activity and the activation of so-called *cryptic splice sites* can lead to disruption of transcript structure. Deeply intronic variants may also lead to the formation of pseudoexons, extended regions of intronic sequence that become recognised as exons due to the presence of a variant and so are retained in the final transcript. Conversely, a small number of cases of pathogenic *exitrons* have been identified, in which pathogenic variants in long exons lead to exclusion of an internal region of exonic sequence (Wai et al., 2020).

To evaluate the potential pathogenicity of an sQTL, a number of features need to be considered: splicing defects result in the inclusion or exclusion of nucleotide sequence, and so, similarly to indels, frameshifts and resultant PTCs are a common consequence. In these cases, significant disruption to protein structure and/or transcript levels (e.g. through NMD) may be expected. Where the inserted or deleted region is in-frame, the size of the excluded region is of greater importance; however, although the addition of a small number of bases is theoretically less likely to be disruptive to protein function, recent bioinformatics analysis has shown that a significant number of in-frame single amino acid insertions and deletions are predicted to be pathogenic (Pagel et al., 2019). Another key consideration is the effect size of the variant: in theory, the greater the proportion of transcripts exhibiting a mis-splicing event, the more likely it is to cause disease. Due to the tissue-specific nature of gene expression, it is also important to consider where the variant lies within tissue-relevant transcripts.

1.2.5.1. Pathogenic variation in spliceosomal components

Pathogenic variants in genes encoding major spliceosomal components may themselves cause Mendelian disease through global disruption of splicing. Such disorders are associated with variants in many spliceosomal components, and may affect many tissues. Loss-of-function variants in one U5 snRNP component, PRPF8, for instance, lead to the onset of an autosomal dominant form of retinitis pigmentosa (RP; McKie et al., 2001, Vithana et al., 2001), while loss-of-function variants in another, EFTUD2, lead to the disorder mandibulofacial dysostosis, Guion-Almeida type, a multi-system disorder with characteristic facial dysmorphology (Lines et al., 2012).

Despite the low frequency of introns excised by the minor spliceosome, variants in minor spliceosomal components still have the capacity to cause Mendelian disorders. Homozygous or compound heterozygous loss-of-of-function variants in the minor spliceosomal U4 snRNA RNU4ATAC, for instance, are associated with a distinct trio of Mendelian disorders: microcephalic osteodysplastic primordial dwarfism type 1, or MOPD1 (He et al., 2011), Roifman syndrome (Merico et al., 2015), and Lowry-Wood syndrome (Farach et al., 2018).

1.2.5.2. The role of mis-splicing in variable expressivity and incomplete penetrance

Two confounding phenomena in variant interpretation are variable expressivity and incomplete penetrance. In the former, individuals harbouring pathogenic variant(s) in the same gene exhibit a wide spectrum of symptoms. In the case of pathogenic variants in the neurofibromin 1 (*NF1*) gene, for example, some patients may develop numerous tumours named neurofibromas, while others may have only lightly pigmented *café-au-lait* spots on their skin (Easton et al., 1993). In incomplete penetrance, the inheritance of a pathogenic allele(s) does not necessarily lead to the onset of disease, and is a feature of many disorders, including retinoblastoma (Harbour, 2001) and some forms of spastic paraplegia (Varga et al., 2013). As such, dominant traits can appear to skip generations, and incomplete penetrance can also greatly hinder segregation analysis (see *1.3.5.*).

Splice-impacting variants may provide a model to account for these phenomena to some degree. sQTLs vary widely in their predicted effect size (Garrido-Martín et al., 2021, GTEx Consortium, 2020, Takata et al., 2017). Variants with the most penetrant effects on local splicing efficiency, such as those at canonical splice sites, may be speculated to impact the majority of transcripts for the affected gene. At the opposite end of the spectrum of effect size, low-impact sQTLs are perhaps unlikely to individually cause Mendelian disorders; rather, it is more likely that sQTLs may contribute en masse to more complex disorders. In a study of transcripts in the human brain, for example, it was shown that schizophrenia-associated genes were significantly enriched for sQTLs, which may account for individual susceptibilities to onset of the condition (Takata et al., 2017).

However, an intriguing possibility is that weaker sQTLs and eQTLs may act in tandem to either attenuate or potentiate the splicing impacts of pathogenic variants that would otherwise cause, or not cause, Mendelian disease, either *in cis* or *in trans* with the causative gene. If an asymptomatic individual harbouring a pathogenic splice-impacting variant also harbours a compensatory sQTL(s), or eQTL(s), elsewhere in their genome, it may account for an absence of disease phenotype. Conversely, a variant that is non-splice-impacting against a particular genetic background may significantly disrupt splicing against another. This provides a potential framework to account for variable expressivity and incomplete penetrance of splice variants. In such cases, most current interpretation methodologies, which evaluate individual variants (see *1.3.5*), would prove insufficient, and holistic consideration of all variants would be necessary.

The impact of variants on splicing can vary widely in penetrance and magnitude. As such, splice variants can prove challenging to interpret. To better understand the complexities of splice variant interpretation, it is helpful to gain a broader understanding of how variants are identified and interpreted in a clinical context.

1.3. Identification and interpretation of clinical variants

In the case of SNVs and indels, four main steps are required for the identification of pathogenic variants from raw genomic material: in an optional first step, patient DNA must be enriched for the genomic region(s) of interest. Following this, a sequencing workflow must be selected. After sequencing, the resulting genomic data must be processed bioinformatically to generate a list of high-confidence variants. Finally, it is then necessary to use all information about the patient, variant and corresponding gene to *interpret* the variant; that is, to evaluate the likelihood that it is pathogenic or benign.

1.3.1. Target enrichment strategies for next-generation sequencing

After biosample preparation and DNA extraction, it is often (but not always) desirable to enrich the samples for particular regions of interest (see *1.3.4.*). Two enrichment workflows are particularly commonplace.

In the first, *amplicon-based enrichment*, target genomic regions are amplified through a multiplex PCR reaction. In essence, this entails the design of sets of primers that bind to regions of interest and promote their amplification through a standard multiplex PCR protocol. There is an upper limit on the number of amplicons that can be concurrently generated during a single reaction, somewhere in the realm of 5,000-10,000, which may be sufficient when there are relatively small cohorts of genes of interest. The bespoke primers used to conduct the PCR reaction mean that the enrichment protocol has a low rate of off-target enrichments, and the workflow is fairly quick.

The second method is *capture hybridisation*, in which nucleotide probes complementary to the region(s) of interest are held in solution and the genomic material added. Following binding of the targeted regions of the patient DNA to the probes, unbound DNA is washed off and the enriched regions retained for downstream sequencing. Capture hybridisation does not suffer from the same limit on scale as amplicon-based enrichment, and so can be used for larger sets of genomic regions, but has a more involved and expensive experimental protocol.

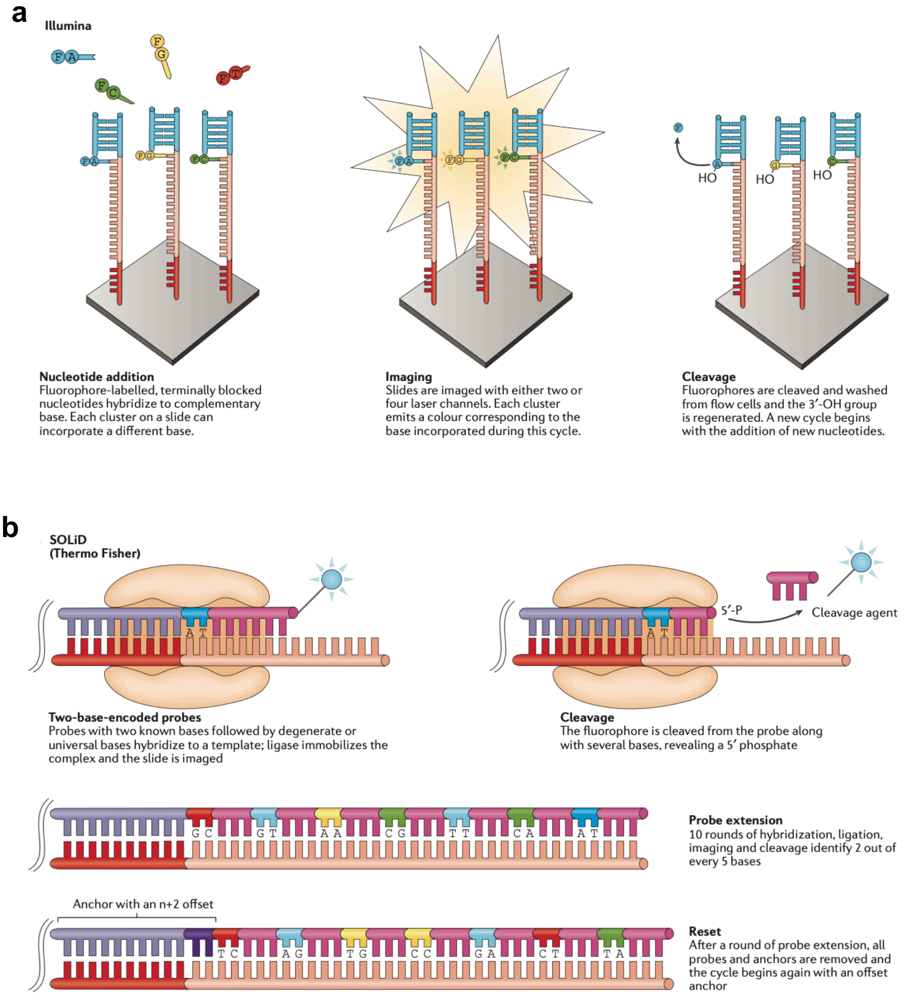
In the Oxford Nanopore long-read workflow (see *1.5.1.*), an alternative enrichment strategy is sometimes used, in which the bacterial endonuclease *Cas9* is used to excise a region of interest from the surrounding genomic sequence and adapters ligated solely to the excised fragments before the downstream sequencing reaction itself (McDonald et al., 2021).

1.3.2. Sequencing technologies in the identification of SNPs and indels

Many diverse platforms exist for the sequencing of DNA; many are based on the use of modified nucleotides bound to fluorophores, which each produce a characteristic emission spectrum to identify the precise base or bases that are present at each site within the DNA fragment. The majority of fluorophore-based sequencing platforms are described as either *sequencing-by-ligation* (*SBL*) or *sequencing-by-synthesis* (*SBS*).

Prior to sequencing, most workflows incorporate an amplification step, which allows spatial concentration of large numbers of copies of the same fragment to boost the fluorescent signal. These amplification steps are themselves diverse: in Illumina sequencing, a bridge amplification is used, whereby adapters ligated to the end of target fragments bind to oligonucleotide sequences bound to a flow cell. These adapters can then act as primers to facilitate polymerase binding and subsequent synthesis of identical daughter strands. In other methodologies, such as 454, SOLiD and GeneReader, target fragments bind to adapter sequences present on micelles or beads, and are amplified through emulsion PCR.

In SBS (**Figure 6a**), the sequencing reaction is conducted in a solution containing modified versions of the four nucleotides, with each nucleotide bound to a fluorophore that fluoresces with a characteristic emission spectrum upon incorporation by polymerase into the fragment of interest. The fluorophore is then cleaved from the newly incorporated nucleotide to allow the next round of sequencing. This process is repeated *n* times, to produce a read of length *n*. The collated fluorescence signals are then analysed to infer the order of nucleotide incorporation, and thus the sequence of the fragment.



Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.

Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.

Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

Two-base-encoded probes
Probes with two known bases followed by degenerate or universal bases hybridize to a template; ligase immobilizes the complex and the slide is imaged

Cleavage
The fluorophore is cleaved from the probe along with several bases, revealing a 5′ phosphate

Probe extension
10 rounds of hybridization, ligation, imaging and cleavage identify 2 out of every 5 bases

Reset
After a round of probe extension, all probes and anchors are removed and the cycle begins again with an offset anchor

Figure 6. *Comparison of two major exemplar next-generation sequencing chemistries: sequencing by synthesis (SBS) and sequencing by ligation (SBL).* (**a**). In sequencing by synthesis, the successive addition of fluorophore-bound nucleotides is associated with characteristic emission spectra, followed by cleavage of the fluorophore dye to allow the next sequencing cycle. (**b**) In sequencing by ligation, probes consisting of dinucleotides bound to a fluorophore and degenerate nucleotide sequence bind to a target, which is itself bonded to an anchor. Ligation of the backbone results in fluorescence of the dye, which is subsequently cleaved. This process is repeated until the desired number of bases have been added, then a new cycle begins with an offset anchor to ensure sequencing of all bases. Adapted from Goodwin et al. (2016).

Most SBL sequencing workflows begin with the binding of an anchor sequence to adapters ligated to the end of fragments of interest (**Figure 6b**). The reaction proceeds in the presence of probes consisting of a mononucleotide (e.g. in Complete Genomics sequencing) or dinucleotide (e.g. in SOLiD sequencing) downstream of a degenerate sequence of predetermined length, and bound to a fluorophore dye. At each round of the sequencing protocol, a probe carrying the complementary (di)nucleotide binds the target sequence, and its ligation produces a characteristic fluorescence. A fixed number of bases are then cleaved from the end of the probe and the next probe is ligated and fluoresces, and so on. Subsequent steps of the reaction involve removal of the bound anchors and probes and replacement of an anchor with an n + 2 offset to ensure coverage of all bases. The resulting read is constructed by deconvoluting the fluorescence signals from the multiple sequencing cycles.

Broadly speaking, SBL methodologies display a greater degree of accuracy than do SBS methodologies, with accuracy reported as high as 99.9% (Drmanac et al., 2010, Liu et al., 2012). This is largely due to the repeated sequencing of the same bases at different anchor offsets. Despite this, Ilumina SBS workflows remain the dominant sequencing technology employed by diagnostic centres, at least partly owing to their long-standing presence in the market and wide variety of platforms.

A promising emerging sequencing technology is that of long-read sequencing. As suggested by the name, long-read technologies differ from short-read in the length of nucleotide sequence that can be analysed in a single read, with one study showing reconstruction of a 4 Mb locus using reads of up to 882 kb in length (Jain et al., 2018). Through long-read sequencing, complex structural variants can be ascertained through observation of all breakpoints at once; this has been conducted across the scale of thousands of individuals to unpick the contribution of structural variation to disease and other traits (Beyter et al., 2021). Two main platforms currently predominate in the long-read market: Oxford Nanopore and PacBio. The Oxford Nanopore platform calls bases by detecting changes in electrical signals as a nucleotide strand is passed through a membrane-bound pore. Sequential combinations of nucleotides generate a characteristic signal, or “squiggle”, which is deconvoluted to identify the nucleotide sequence. The PacBio Single Molecule, Real-Time (SMRT) long-read workflow involves the circularisation of long nucleotide fragments through the ligation of single-strand nucleotides to its ends. Primers and polymerase bind to these adapters and the nucleotide is then immobilised at the bottom of wells named zero-mode waveguides. As fluorescently labelled nucleotides are added, the emission spectra are captured and converted to nucleotide sequence.

Both PacBio and Nanopore are error-prone methodologies, with erroneous base-calling and falsely called indels pervading the generated reads (Dohm et al., 2020; Fu et al., 2019). However, unlike most short-read approaches, both of these Nanopore and PacBio workflows are able to sequence the same fragment multiple times: in Nanopore, the fragment can be re-threaded back through the pore and analysed again, while the circularisation of fragments in PacBio means they can be repeatedly cycled around the polymerase. Thus, through the generation of multiple reads per fragment, the consensus over all reads can be returned, to some extent averaging out the error rate (Dohm et al., 2020; Fu et al., 2019).

1.3.3. Bioinformatics processing and analysis of sequencing data

The end product of a sequencing reaction is a file consisting of the sequencing reads generated during the run. Processing from this point onwards is purely bioinformatic, and may vary from centre to centre, but three primary steps are necessary and consistently applied across centres.

The first necessary bioinformatics step is to align the reads to the human genome. A host of alignment software is currently available, with popular examples including the Burrows-Wheeler algorithm, or BWA (Li and Durbin, 2009), TopHat (Kim et al., 2013) and its successor HISAT2 (Kim et al., 2019), and STAR (Dobin et al., 2013). The performance of alignment tools is often gauged by alignment rate and gene coverage, in which regard a recent study has demonstrated BWA to be the most effective algorithm for short-read alignment of genomic reads, while HISAT2 was found to be significantly quicker than all other analysed tools (Musich et al., 2021).

The next processing step is the calling of variants from among the aligned reads. Effective variant callers will successfully identify SNVs and indels from artefactual noise generated during the sequencing process. Variant calling tools in widespread use include GATK (Poplin et al., 2018), the VarScan method of the samtools software package (Li et al., 2009) and Strelka2 (Kim et al., 2018). A recent comparison of the performance of these three packages identified Strelka2 as being the optimal choice of variant caller in terms of both accuracy (being slightly more accurate than GATK) and processing speed, regardless of the upstream sequencing platform (Chen et al., 2019).

Finally, the list of variants must be annotated. Such annotations generally include a prediction of variant consequence, but may also include predictive metrics and population frequency estimates. As with aligners and variant callers, a number of annotation tools exist: Ensembl's Variant Effect Predictor, or VEP (McLaren et al., 2016), has entered widespread usage, while alternatives include Annovar (Wang et al., 2010, Yang and Wang, 2015) and SnpEff (Cingolani et al., 2012). There is perhaps a surprising discordance in the predicted consequences of variants between annotators, with one study demonstrating that annotations were consistent between Annovar and VEP for only 65% of loss-of-function variants (McCarthy et al., 2014). These results suggest that annotation of variants by multiple tools may provide more diagnostic insight for clinical staff.

1.3.4. Diagnostic sequencing options in the identification of genomic variants

The choice of enrichment and sequencing platform is largely informed by the nature of the downstream analysis. The various approaches available for clinical variant analysis are primarily defined by the size of the genomic region surveyed.

Diagnostic methodologies can be designed to target only those genes associated with a particular disease phenotype, an approach termed a *gene panel*. By limiting the genomic space surveyed, a greater depth of sequencing can be achieved for finer resolution of individual variants. Gene panels employ custom capture protocols: the Custom Target DNA Enrichment workflow by Agilent, for instance, generates probes for capture hybridisation of up to 100s of genes. As genes covered by panel-based approaches must by necessity be decided in advance, coverage of as-yet-unknown causes of disease is not accommodated by panel approaches. Additionally, gene panels generally amplify only the exonic regions of a gene; as such, deeply intronic variation is not identified using panel-based approaches.

Exome sequencing further broadens analysis to larger groups of genes. This may include all protein-coding genes (*whole exome sequencing*), or all disease-associated genes (*focused exome sequencing*). Due to the wider scope of analysis, exome sequencing is most often capture hybridisation-based. Similarly to gene panels, exomes target only exonic regions. Unlike gene panels, exomes offer the potential for novel gene discovery, making them a popular diagnostic approach for research purposes, particularly when prior gene panel testing has been negative. Research into exome efficacy has shown that they return a confirmed molecular diagnosis for around 25% of the assessed patients, many of whom have already undergone extensive upstreaming testing (Atwal et al., 2014, Yang et al., 2013).

At the broadest level, *whole genome sequencing* (WGS) aims to capture almost the entirety of the genome. Most commonly, WGS is conducted through *shotgun sequencing*, in which whole genomic DNA is fragmented through sonication or enzymatic activity, bound by adapters and then sequenced using a conventional approach. The wide genomic space covered by WGS does entail some caveats: files generated during processing of WGS data may be around ~100 GB in size (Narayanasamy et al., 2020), requiring substantial storage (particularly if kept long-term), and may return tens of millions of variants (McCarthy et al., 2014). The interpretation of WGS data, therefore, is much more difficult than WES or panel data. However, nationwide efforts, such as the 100,000 Genomes Project (Turnbull et al., 2018), have allowed collaborative input on patient data to accelerate interpretation of these large datasets.

Repetitive regions of the genome are difficult to sequence due to ambiguities they introduce in the alignment (Treangen and Salzberg, 2011), and so the above sequencing workflows are often unable to accurately call variants located in repetitive sequences. Where such regions are linked within disease-associated genes, such as ORF15 of the *RPGR* gene (Chiang et al., 2015), one of the most common causes of X-linked retinitis pigmentosa (Shu et al., 2007), may require bespoke analyses, such as Sanger sequencing, when upstream diagnostic approaches return do not return a molecular diagnosis.

The choice of diagnostic approach is particularly important when investigating non-coding variation, including splice-impacting variants. As mentioned above (see *1.3.4.*), while panel and exome sequencing are capable of identifying pathogenic variation at the canonical splice sites, deeply intronic variants are omitted from the scope of sequencing; some short lengths of intronic coverage may achieve sequencing coverage, but this may be filtered from clinical analysis in favour of exonic regions. Identification of novel non-coding variants in intergenic elements, such as enhancers, are also beyond the scope of WES or panel testing, and so require a whole genome approach.

1.3.5. The American College of Medical Genetics and Genomics pathogenicity framework for variant interpretation

Following the identification and annotation of a variant, it next must be interpreted. To develop a consistent approach for international use in interpretation of variants, a set of guidelines has been developed by the American College of Medical Genetics and Genomics (ACMG; Richards et al., 2015). These guidelines employ a tiered approach in which different lines of evidence can be used to assess the pathogenicity of a variant (**Figure 7**). Variants are assigned one of five classifications: pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign or benign, dependent on the combination of assessment criteria observed for the variant. A classification of VUS is given if conflicting benign and pathogenic criteria are identified for a single variant. The ACMG guidelines have become widely used, and the framework forms the basis of variant interpretation annotations in repositories such as ClinVar (Landrum et al., 2018) and the Leiden Open Variation Database (LOVD; Fokkema et al., 2011). Further, in 2016, the Association for Clinical Genetic Science issued an official statement recommending the adoption of the ACMG guidelines across all UK diagnostic centres (McMullan, 2016).

These guidelines incorporate consideration of a wide variety of variant- and gene-level features in the interpretation of variants. The highest level of support for pathogenicity, termed PVS1, is assigned only for the following variants:

“[A] null variant (nonsense, frameshift, canonical ± 1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease”

Richards et al. (2015)

Other characteristics considered strong evidence of pathogenicity include: variants causing the same amino acid change as an existing, known pathogenic variant; a variant arising *de novo* in a patient in a gene known to cause the phenotype; increased prevalence of the variant in affected individuals compared to controls; and the existence of *in vitro* or *in vivo* functional studies corroborating the deleterious impact of a variant.

Population-level information may also highlight potential pathogenicity: as Mendelian diseases are generally very rare, we would expect the variants underpinning them to exist at a very low level in the general population (with recessive disorders potentially existing at a slightly higher level due to the presence of carriers in the general population). Historically, the use of publicly available population frequency databases such as dbSNP (Sherry et al., 2001) and ExAC (Lek et al., 2016) have facilitated interpretation of this criterion. Recent years have seen the widespread adoption of the gnomAD database (Karczewski et al., 2020), which encompasses a larger cohort of individuals across a wider range of ethnic backgrounds, and includes variant counts from whole genome sequencing data, allowing evaluation of frequency for intronic variants. This is particularly beneficial in the analysis of putatively splice-impacting variants, which may lie deep in the intron regions.

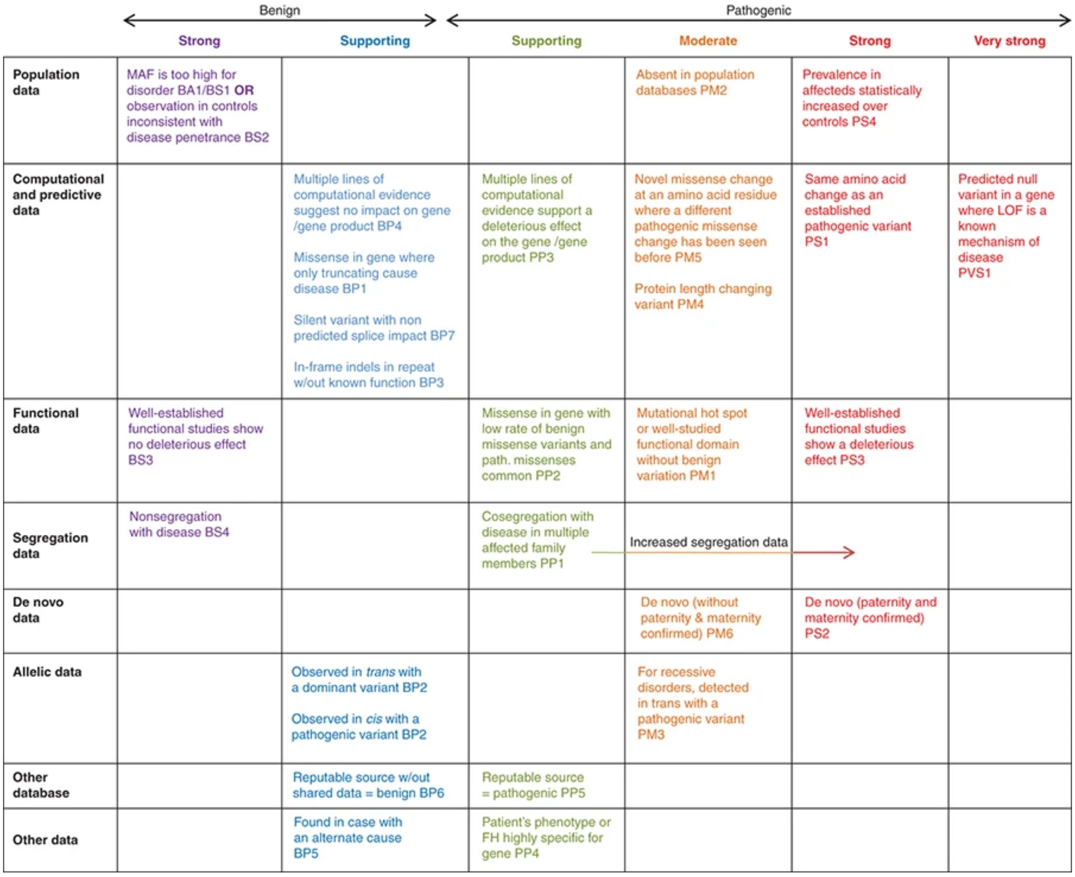
Multiple lines of evidence may also be used to support a benign classification: for example, variants present at a higher frequency in the population than expected given the inheritance pattern, prevalence and penetrance of the disease are assigned the strongest level support for a benign classification. Nonsegregation of a variant with the disease phenotype – that is, the inconsistent presentation of disease phenotype in related individuals known to harbour the variant – also results in the assignment of this level of support.

ACMG guidelines permit the use of 16 missense prediction tools as weakly supporting evidence for pathogenicity, including the tools CADD (Rentzsch et al., 2019), PolyPhen2 (Adzhubei et al., 2010) and SIFT (Ng and Henikoff, 2003). Six splice prediction tools are also listed for use in this regard, namely GeneSplicer (Pertea et al., 2001), Human Splicing Finder (Desmet et al., 2009), MaxEntScan (Yeo and Burge, 2004), NetGene2 (Hebsgaard et al., 1996), NNSplice (Reese et al.,

1997) and FSPLICE (http://www.softberry.com). As results from different tools may be discordant, evidence from predictive analysis can only be considered if supported by multiple tools.

1.3.5.1. Transcript choice in variant interpretation

An important factor in the investigation of variant pathogenicity is the transcript, or transcripts, in relation to which the variant is being interpreted. As a result of alternative splicing (see *1.2.4.*), the transcription of a single genetic locus can result in many different transcripts. Thus, a variant with a particular impact in one transcript may have a different impact in another, and variants that are exonic in one transcript may be intronic in another (and vice versa).



	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 **OR** observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Figure 7. *ACMG criteria for classification of pathogenic variants.* According to ACMG guidelines, the pathogenicity of a variant is assessed by the overall body of supporting criteria shown here. Bioinformatics predictions of splicing impact are considered only a supporting criterion for pathogenicity (criterion PP3). Criteria range across a spectrum of classes, with the strongest support being assigned for null variants in loss-of-function genes. Taken from Richards et al. (2015)

Alternative splicing is highly tissue-specific, and so transcript selection should reflect the isoform(s) present in the most disease-relevant tissue. The importance of correct transcript selection in variant interpretation has been highlighted in recent research: one study demonstrated mis-reporting of variant pathogenicity in three different genes (*CKDL5*, *KMT2C* and *OFD1*) due to either lack of coverage of tissue-specific isoforms in the sequencing protocol or interpretation of variants against a non-tissue-relevant transcript (Schoch et al., 2020). This resulted in the missed reporting of pathogenicity for two variants, and the false reporting of pathogenicity for the third. Another such misannotation was recently described in the DYNC2H1 gene (Vig et al., 2020), in which three patients with retinal degeneration harboured an ostensibly intronic variant in a homozygous state. Analysis of retinal organoids derived from patient cells revealed that this variant in fact affected a retinal microexon present in a non-canonical transcript that would not have been conventionally surveyed in a clinical context. These findings also led to a new genotype-phenotype association for DYNC2H1, which is conventionally associated with a form of short-rib thoracic dysplasia with only occasional ocular involvement (Dagoneau et al., 2009). Both examples highlight that blanket use of canonical transcripts may hinder accurate variant interpretation.

Guidance for accurate transcript selection is given in the ACMG guidelines, which state the following:

"*A reference transcript for each gene should be used and provided in the report when describing coding variants. The transcript should represent either the longest known transcript and/or the most clinically relevant transcript.*" (Richards et al., 2015)

Fulfilling this provision, however, is dependent on an accurate understanding of tissue-specific isoform structure. Transcriptomic analysis of a diverse range of cell and/or tissue types will be of great benefit to ensuring variants can be interpreted in a context relevant to the biological system(s) being studied.

The MANE project is a collaborative project aiming to produce high-quality transcript models for all human genes, through cross-referencing of features – such as 5’ and 3’ UTR co-ordinates – between transcripts listed in the Ensembl (Howe et al., 2021) and RefSeq (O'Leary et al., 2016) repositories. While such efforts will indeed provide a valuable set of gold-standard known transcripts, caution should be taken in interpreting variants blindly against these them. Rather, there is a major unmet need to construct tissue-specific maps of alternative splicing, and evaluate variants in relation to tissue- and disease-relevant transcripts.

1.3.5.2. Interpretation of splice variants against ACMG criteria

Explicit guidelines for the interpretation of splice-impacting variants are scarce in the ACMG guidelines; however, canonical splice site variants at the ± 1 or 2 position are deemed null. Selection of the correct transcript is important in this regard, as a variant that occupies a canonical splice site in one transcript may be annotated otherwise against a different transcript model. Indeed, splice-impacting variants may be annotated by standard annotation pipelines as a number of different variant types, including synonymous (Pagani et al., 2005, Zeng and Bromberg, 2019) or missense (Dionnet et al., 2020, Uddin et al., 2020) variants. It is thus important to consider mis-splicing as an alternative pathogenic mechanism when interpreting variants of these types.

Variants at residues in the extended splice site are not currently incorporated into the ACMG guidelines, due to the high inconsistency in their splicing impact. Thus, assigning pathogenicity to these variants is largely reliant on the conduction of functional studies (ACMG criterion PS3). The results of these assays may themselves be difficult to interpret when variants affect a smaller proportion of transcripts, or lead to the in-frame inclusion of small numbers of amino acids. Conflicting interpretations of splicing impact may also be given dependent on the nature of the assay(s) conducted, and variants annotated as being splice-impacting have been shown to be at a higher risk of reclassification than most other variant types (Esterling et al., 2020).

Despite the provision for *in silico* analysis of splicing impact in the ACMG guidelines, recent years have seen a surge in the number of bioinformatics tools predicting impact of variants on splicing, many of which are based on modern machine learning models (of the six tools listed above, only NetGene2 and NNSplice are based on such models). Their newness means such tools are absent from the existing guidelines. The accuracy of the wide variety of newly-available machine learning tools is only just beginning to be assessed, and these tools may offer an improved ability to identify variant impact on splicing than that of existing tools. To determine the likely benefit of integration of these tools into clinical practice, there is thus a need to empirically evaluate their accuracy on a cohort of functionally evaluated variants. The performance of these splicing prediction models may depend on the machine learning paradigm on which they are based.

1.4. Machine learning approaches for the prioritization of genomic variants impacting pre-mRNA splicing

Machine learning is already being applied to great effect in diverse biological fields, such as the modelling of social networks in animal behaviour studies (Valletta et al., 2017; Psorakis et al., 2012) and protein secondary structure prediction (Wang et al., 2016). The application of machine learning to the prediction of variant impact on splicing has been accelerated by the recent availability of large-scale transcriptomic datasets, such as the GTEx project (GTEx Consortium, 2013), which allow researchers to link genomic diversity with transcriptomic variation across large numbers of individuals and tissue types (Castel et al., 2019, GTEx Consortium, 2020, Ferraro et al., 2019).

Depending on sequencing strategies, clinical scientists will be expected to interpret and triage hundreds to millions of genomic variants per individual, although many variants can be immediately excluded due to their frequency in the general population (Richards et al., 2015). The development of effective machine learning tools for the prediction of splicing impact will allow prioritization of likely pathogenic variants among the mass of genomic variants returned by standard diagnostic pipelines. Ultimately, these tools may prove a valuable asset in improving diagnostic yield globally.

Here, I provide a summary of some of the major machine learning-based splice analysis tools released to date. While the focus here is largely on the functionality of these tools, some basics of machine learning are introduced to allow easier understanding of their underlying logic.

1.4.1. Early computational methodologies for the prediction of splicing

Despite the relatively recent advent of machine learning-based splicing models, many other computational approaches to the prediction of splice disruption have been described over the last two decades.

Many early tools for the prediction of splicing regulatory element (SRE) binding sites were based on position weight matrices (PWMs)—log-scaled representations of the frequency of particular nucleotides within sequences predicted to bind splicing factors. Experimental derivation of such PWMs formed the basis of tools such as ESEFinder (Desmet et al., 2009) and Human Splicing Finder (Cartegni et al., 2003), and decreased fitting of mutant sequences to the PWM model was seen as evidence for impairment of splice factor binding.

Many computational and experimental approaches to splice prediction have involved the investigation of nucleotide hexamers (i.e., sequences of 6 bases length). The method RESCUE-ESE, for example, computationally identified 10 splice-enhancing hexanucleotides in the vicinities of weak splice sites (Fairbrother et al., 2002). An approach named ESRseq (Ke et al., 2011) made use of a saturation technique in which all 4096 possible nucleotide hexamers were scored for splicing impact based on in vitro minigene splicing assays. The tabulated results of these experiments, published online, could then be used to speculate on the splicing potential of a mutant sequence versus its wild-type counterpart.

One early tool that remains widely used in splice site prediction is MaxEntScan (Yeo and Burge, 2004). Based on the principles of maximum entropy modelling (MEM) from the field of information theory, MaxEntScan generates two models based on a set of real and decoy splice sites. It then compares the probability that a nucleotide sequence belongs to each of the two distributions and returns how much more likely it is that the sequence is a real, rather than decoy, site.

1.4.2. Basics of machine learning methodologies

All machine learning models require both training and testing—to do this, a relevant data set is divided into both a training set and a test set. Importantly, no entry in one set is present in the other; were there to be overlap, the model would be over-trained to recognize those items in the test set that it had already seen, and measurement of model efficacy would overestimate its accuracy and efficacy. The variables or characteristics in each dataset that are input to a model are termed features. In the earliest stages of training, some model-specific algorithm is applied (usually iteratively) to this training set to develop an initial model. The model is then applied to the test set and its efficacy quantified. Measurement across different versions of the model then allows the model to be fine-tuned to maximize its efficacy.

The efficacy of a machine learning model is generally measured as some kind of loss function – in essence, a measurement of how far a model's predictions deviate from the expected outcome, and machine learning algorithms strive to minimize this value over the course of the generation of the model. In other words, these models are gradually tweaked so that their ability to accurately classify data improves over the training process.

1.4.2.1. Features

A key element of machine learning is the use of features: these are the underlying characteristics or variables that are input to the models and from which inferences are ultimately made. It is these features by which data are classified or separated. In the context of genomic and transcriptomic analysis, many of these features are often sequence-based, representing the frequency or position of particular nucleotide sequences over a given region. Biochemical features, such as GC content and thermodynamic properties, are often also employed. Moreover, some tools adopt a meta-analytical approach through the incorporation of output from other tools as features, such as the use of SPANR (Xiong et al., 2015) and CADD (Rentzsch et al., 2019) scores in S-CAP (Jagadeesh et al., 2019; see *1.4.5.6.*). Differences in choice of features may often underlie the various strengths and caveats of particular tools.

Table 1. *Glossary of machine learning terms.* SVM, support vector machine

Term	Definition
Backpropagation	The computational process by which a neural network adjusts the weights and biases of the network in such a way as to reduce the loss of the model.
Bagging	Abbreviation for bootstrap aggregation. The training of a model on random subsets of data entries and features to improve generalizability of a model (usually a decision tree-based model).
Bias	A (usually negative) value that represents a neuron's inherent tendency towards inactivity. Usually randomized for each neuron before the training of a network.
Classification	A type of machine learning system in which the output is assignment of a data point to a discrete group. Usually contrasted with regression.
Feature	One of a set of variables in a dataset that are input to a machine learning model. Machine learning models classify data according to the values of features in the dataset.
Hidden layer	One of any number of layers of neurons lying between the input and output layers of a deep neural network.
Hyperplane	A surface with one fewer dimensions than the space it occupies. SVMs separate datasets with *n* features using a hyperplane of *n–1* dimensions. For example, if there are 6 features, an SVM attempts to create a 5-dimensional hyperplane that best separates data.
Kernel trick	The use of a mathematical function allowing inference of relational qualities of data without explicitly carrying out computationally expensive mathematical calculations.
Loss function	A mathematical function measuring the degree to which a model's predictions deviate from the true classifications of data.
Machine learning	The use of computer systems to detect patterns in and make inferences from data without explicit instruction.
Multiclass SVM	A subtype of SVM used when data may be classified into more than two classes.
Neuron	The basic unit of a neural network, taking in input from previous neurons and propagating a weighted response to subsequent ones.
Regression	A type of machine learning system in which the output is the prediction of a continuous or ordered value. Usually contrasted with classification.
Support vectors	Data points that lie along the margins between classifications in an SVM model.
Training set	A dataset containing the data that is presented to a machine learning system and then used to make inferences and learn patterns present within the data.
Test set	The dataset used to evaluate performance of the model. The test set is generally taken from the same source as the training set, but may come from elsewhere.

1.4.2.2. Training and test sets

One major contributing factor to the rapid surge in the number of machine learning-based splice prediction tools is the increased availability of publicly-available datasets. Particularly valuable are experimentally-derived RNA-seq datasets, which allow effective linking of genome- and transcriptome-level features. Several tools also incorporate measurements of pathogenicity in the form of variant classifications from ClinVar (Landrum et al., 2014). Many tools use raw sequence data as input; in such cases, these sequences are taken from a reputed transcript model, most often GENCODE (Harrow et al., 2012), as in the cases of MMSplice (Cheng et al., 2019) and SpliceAI (Jaganathan et al., 2019).

1.4.2.3. Outputs of machine learning methodologies

Machine learning models broadly fall into the categories of *regression* and *classification* models. Classification models identify the class (of a set of classes) to which an unseen data entry is most likely to belong. On the other hand, regression models use input data to predict a quantitative value. Thus, the output of a model depends on its design, the types of features which are utilized as input and the objectives of the prediction tool. Most splice prediction tools utilize regression models, and generate predictive scores corresponding to, for example, the strengthening or weakening of a novel or existing splice site (SpliceAI), the magnitude of an exon skipping event (SPIDEX), or variant pathogenicity (S-CAP). How scores from these tools are utilized and interpreted is thus highly dependent on the tool being used.

1.4.2.4. Evaluating the performance of a machine learning model

As described above (see *1.4.2.*), many machine learning models refine themselves over training iterations by minimizing some kind of loss function. However, comparative analysis of the relative performance of different models usually relies on the construction of an unseen test dataset that can be applied to both/all models. Model performance metrics, such as the area under curve (AUC) of both receiver-operating characteristic (ROC) and precision-recall (PR) curves, can then be used to more directly compare model performance, although this may be confounded by many factors (see Discussion).

1.4.3. Common machine learning models in splice prediction

1.4.3.1. Support vector machines (SVMs)

SVM models aim to use a *hyperplane* (a surface with one fewer dimension than the space around it) to separate data belonging to different classes. This is done such that the distance between the hyperplane and data that lie closest to the overlap between two classes—the so-called *support vectors*—is maximized (**Figure 8a**). Data presented to an SVM are then classified according to which side of the hyperplane they lie on. *Multiclass SVM* approaches can also be used where there are more than two outcome classes to which data may be assigned. Finally, data which cannot be separated by a single continuous hyperplane (**Figure 8b**) are able to be transformed using the *kernel trick*. This approach makes use of kernel functions—mathematical operations that allow inference of relational qualities between data points in a computationally inexpensive manner. Common kernels used in machine learning are the polynomial and radial basis function (RBF) kernels, although a multitude of others exist.

Importantly, standard SVMs are only able to classify data as belonging to one group or another; to provide probabilistic measures of confidence or effect size, models need to be adapted and extended.

1.4.3.2. Decision Trees

Decision trees are a simple but powerful form of machine learning model in which a series of binary choices is designed that produces the most effective classification or prediction of a dependent variable **(Figure 8c**)—this is done through selecting whichever choice allows most accurate separation of data at each stage in the tree-building process. The single decision tree that is generated for a given training set, however, is prone to overfitting and bias for the input data. To remedy this, random forest models are often used (**Figure 8d**). Here, iterative *bagging* (bootstrap aggregating) of the training data, as well as of the variables considered at each stage of the tree-building process, allows the model to be more generalizable to unseen data. Gradient tree boosting (**Figure 8e**) adopts a different approach to bypass overfitting by using the generation of successive trees, each of small contribution to the final model, until decreases in the model loss are negligible.

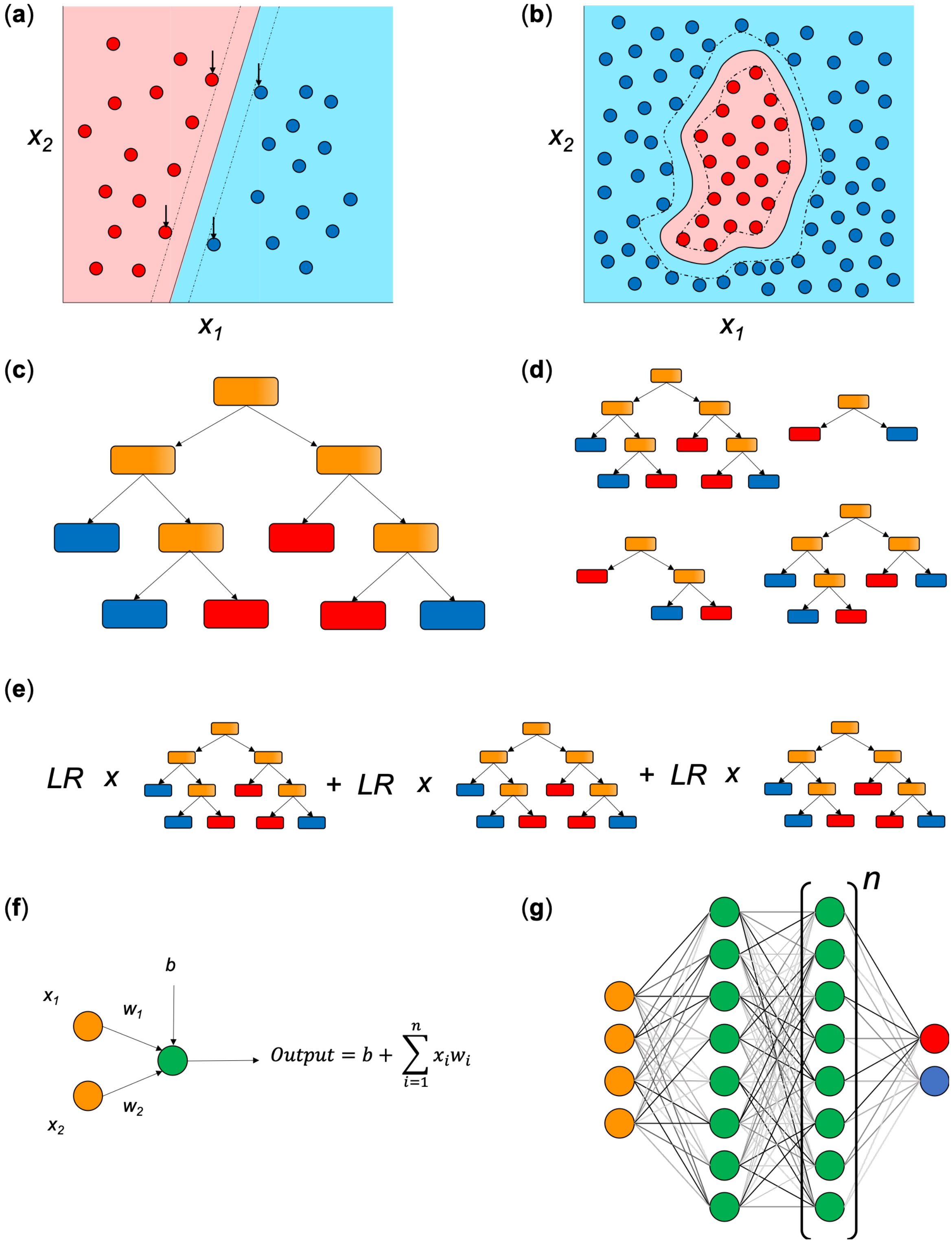

(a)
x_2
x_1
(b)
x_2
x_1
(c)
(d)
(e)
LR x + LR x + LR x
(f)
b
x_1
w_1
x_2
w_2
$Output = b + \sum_{i=1}^{n} x_i w_i$
(g)
n

Figure 8. *Basic machine learning models.* (**a**) Support vector machines (SVMs) classify linearly separable data using a single hyperplane (solid line), with points classified according to the side of the hyperplane on which they lie. Construction of the hyperplane is done using support vectors (indicated by arrows), data points that mark boundaries (dotted) within which the hyperplane must lie. (**b**) Where data are not linearly separable, they may be transformed using kernel functions (radial basis function, or RBF, kernel shown here) which infer relational qualities of data in a computationally inexpensive manner. (**c**) Decision trees use a series of binary choices (orange) to most effectively separate data into different categories (red and blue). (**d**) Random forest models consist of large numbers (often hundreds or thousands) of trees each derived from bootstrap aggregating (bagging) of both input features and data entries in the original training set. (**e**) To mitigate overfitting problems common to decision trees, gradient tree boosting generates successive trees of fixed structure that each contribute a small amount to the final classification, with each tree scaled by a *learning rate* between 0–1. (**f**) In a neural network, a single neuron receives quantitative input *(x_i)* from neurons in the preceding layer and scales them according to the weights of its connection to them *(w_i)*. Each neuron also has a "bias" *(b)*, representing a tendency for inactivity. The output (or *activation*) of a neuron is the sum of each input neuron multiplied by its respective weight, plus this bias value. (**g**) A deep neural network has an initial layer of input neurons (orange), which are coded representations of data features. These are connected to 1 or more layers of "hidden neurons" (green), which are, in turn, connected to an output layer of neurons (red and blue) corresponding to the possible classifications of the data. Predictions may be categorical or continuous and are based on the relative activation of the output neurons. Biases for each neuron and weights for each connection are randomized before the network is trained. After a set of training data is presented, the loss function of the model is calculated (i.e., how accurately or inaccurately the model has classified the known data) and an approach termed *backpropagation* is used to modulate each weight and bias so as to reduce this loss. More data is then presented and this process repeated iteratively to refine the model.

1.4.3.3. Deep Neural Networks (DNNs)

DNNs are computational networks modelled on the activity of biological *neurons* (**Figure 8f**). These neurons are arranged in layers (**Figure 8g**): the first is an input layer, where each neuron is assigned a value corresponding to a feature of the model for that data entry. The final layer contains neurons corresponding to the possible outcomes of the model. Between these are a number of *hidden layers*. Hidden layer neurons receive weighted input from all the neurons in the previous layer, and subsequently distribute the sum of these inputs to all neurons in the next layer by another series of weighted connections. These weightings are assigned at random before the training of the model.

Training data are presented sequentially to a DNN and the resulting output in the final layer recorded and averaged over many training iterations. The efficacy of the model is then compared in relation to expected results. Through a process termed *backpropagation*, the weightings of the connections between neurons are proportionally adjusted so as to minimize the loss function of the model. This is repeated over multiple presentations of training data, or *epochs*, gradually refining the model. Particularly popular in the analysis of nucleotide sequences is a variation termed the convolutional neural network, or CNN, in which input data are ordered in the form of an *n*-dimensional array—that is, nucleotides are input to the model in windows.

1.4.4. Machine learning-based tools for splicing prediction

Here, I describe 7 tools incorporating different aspects of splicing prediction. Below is a tabulated summary of key characteristics of each model for reference (**Table 2**). For ease of visualisation, there are also tabulated and schematic representations of the transcript regions amenable to analysis by each tool, using the pre-mRNA transcript of the *APO3* gene as an exemplar (**Table 3**, **Figure 9**).

Table 2. *Summary of splice prediction bioinformatics tools.* "Citation" denotes references to articles describing tools themselves. SVM, support vector machine; RBF, radial basis function; MPRA, massively parallel reporter assay; HGMD, Human Gene Mutation Database; PSSM, position-specific scoring matrix; pLI, probability of loss-of-function intolerant; RVIS, residual variation intolerance score; AUC, area under receiver-operator characteristic (ROC) curve; PR-AUC, area under precision-recall curve.

Tool Name	Function	ML Model	Training/Testing Data	Features	Efficacy	Citation
CADD	General purpose pathogenicity scoring	v1.0: linear SVM Later releases: L_2-regularized logistic regression	Benign training: evolutionarily neutral variants; pathogenic training: simulated de novo pathogenic variants Benign testing: common benign variants; pathogenic testing: pathogenic ClinVar variants, somatic cancer mutation frequencies	60, covering conversation scores, epigenetic modifications, functional analyses, and genetic context	AUC = 0.916, across all variant types	Rentzsch et al., 2019; Kircher et al., 2014
TraP	Quantification of variant impact on transcripts	Random forest of 1000 individual decision trees	Benign: De novo mutations in healthy individuals Pathogenic: Curated pathogenic synonymous variants	20, including several PSSM-based splice site scores, GERP++ conservation scores, and models of feature interactions	AUC = 0.88, all ClinVar variants AUC = 0.83, ClinVar intronic variants only	Gelfman et al., 2017
SPANR	Cassette exon skipping prediction	Group of neural networks modeled on Bayesian framework	ψ values for all human exons across 16 tissues, based on the Illumina Human Body Map project	1393, including exon/intron lengths, distances to nearest alternative splice sites, conservation and RNA secondary structure	AUC = 0.955, when distinguishing between high (≥67%) and low (≤33%) ψ values	Xiong et al., 2015
CryptSplice	Effect of variants on existing splice sites and cryptic splice site prediction	SVM with RBF kernel	True and false splice sites from GenBank-derived datasets	3 types, all sequence-based, relating to the probability of finding given nucleotide sequences at certain points in splice region	Sensitivity = 97.8% and 88.9% in correctly labeling canonical donors and acceptors, respectively	Lee et al., 2017
MMSplice	Prediction of exon skipping, competitive interactions, changes in splicing efficiency and pathogenicity	Modular neural networks, and linear and logistic regression	Donor/acceptor modules: GENCODE v24 true and false splice sites Exon/intron modules: MPRA data from Rosenberg et al. (2015) Downstream models: various	Direct encoding of the sequence	R = 0.87 and 0.81, correlation between predicted and actual Δψ values for acceptor and donor mutations, respectively PR-AUC = 0.41, exon skipping prediction	Cheng et al., 2019
S-CAP	Variant pathogenicity scoring with the compartmentalization of genomic space	Gradient boosting tree	Pathogenic variants curated from HGMD and ClinVar; benign variants curated from gnomAD	Features across chromosomal, gene, exon and variant levels, e.g., pLI, RVIS, CADD and SPIDEX scores, exon length, splice site strengths	AUC: 0.828–0.959, across 6 regions	Jagadeesh et al., 2019
SpliceAI	Prediction of variant impact on acceptor/donor loss or gain	32-layer deep neural network	GENCODE v24 pre-mRNA transcript sequence for human protein-coding genes	Direct encoding of the sequence	PR-AUC = 0.98 in correct prediction of splice site location from raw sequence	Jaganathan et al., 2019

1.4.4.1. CADD (Combined annotation-dependent depletion)

CADD (Rentzsch et al., 2019; Kircher et al., 2014) was among the earliest machine learning-based variant scoring systems; it generates a score that is approximately interpretable as a measure of pathogenicity.

To train the CADD model, both benign and pathogenic variant sets were derived. For the former, variants with high mean allele frequency (≥95%) in the 1000 Genomes dataset (Auton et al., 2015) were chosen that had arisen since the split between humans and chimpanzees, based on the assumption that such variants had been fixed under natural selection, and so are, at worst, weakly pathogenic. De novo pathogenic variants—both indels and SNVs—were simulated genome-wide using a model informed by local mutation rates and CpG dinucleotide mutation asymmetry.

A wide range of features were incorporated into the CADD model. Such features included: conservation metrics, such as phyloP (Pollard et al., 2010), GERP (Cooper et al., 2005), and phastCons (Siepel et al., 2005); regulatory information, such as transcription factor binding (Johnson et al., 2007) and DNAse I hypersensitivity regions (Boyle et al., 2008); and protein-level predictions, for example Grantham (Grantham et al., 1974), SIFT (Ng et al., 2003), and PolyPhen (Adzhubei et al., 2010) scores. Transcript-level features, such as gene expression levels, were also derived, along with some consideration of splicing in the inclusion of variant distance to the nearest canonical splice site. The initial releases of CADD adopted an SVM-based approach (**Figure 8a-b**) with a linear kernel. However, with later releases, L_2-regularized logistic regression—a form of regression model allowing the modelling and prediction of a binary dependent variable—was shown to lead to improved sensitivity and specificity, and so became the model of choice (Rentzsch et al., 2019).
CADD has been rapidly and widely adopted since its creation, with uses in pathogenicity prediction for many disease subtypes, both Mendelian and complex. In a study of autism spectrum disorder (ASD) in 85 quartet families, for example, CADD scoring was used to filter genomic variants of interest, resulting in the identification of ASD-relevant mutations in 69.4% of affected siblings (Yuen et al., 2015).

The use of CADD scoring has become a gold standard for the prediction of protein-coding variant impact. This ubiquity has led to CADD becoming a benchmark against which many predictive tools are measured. However, its efficacy in terms of splicing prediction is undermined by certain features: the use of conservation scores, for example, may not be informative at the poorly-conserved bases of introns, where cryptic splice sites and pseudoexonisation events are liable to occur. Thus, while a highly effective tool for protein-coding impact prediction, CADD lacks the splice-specific considerations to accurately predict variant effect at the transcript level.

1.4.4.2. TraP (Transcript-inferred pathogenicity) scores

TraP (Gelfman et al., 2017) is a random forest-based tool (**Figure 8d**) for the analysis of non-coding variant impact at the transcript level, providing a score between 0–1 to reflect the scale of this impact. This score corresponds to the proportion of decision trees in the model that predict a variant as pathogenic, and may thus be used as a proxy for the degree of impact a variant is likely to have on a transcript.

TraP was trained on 75 pathogenic and 402 benign variants. To source the former, the authors curated a list of solely synonymous variants associated with rare disease to avoid any incorporation of protein-coding consideration in the model. Synonymous de novo variants in healthy individuals were selected as the benign dataset. These rare variants were selected over common variants in the population to avoid training the model to distinguish solely between rare and common variants.
The TraP model consists of 20 features, primarily splicing-related, including whether or not the variant lies within the splice site region (as pre-defined by the authors); the score of new splice sites where cryptic GT-AG dinucleotides are introduced, according to a position-specific scoring matrix (PSSM); and a bespoke “variant regulatory score”, which incorporates several other features that do not directly affect existing splice sites. The model further incorporates the GERP++ conservation metric (Davydov et al., 2010). The random forest model underlying TraP consists of 1000 decision trees harbouring various combinations of these 20 features.

Table 3. *List of loci pre-mRNA transcript loci amenable to predictive analysis by each of 7 splice prediction tools.*

Tool	Loci Covered
SPANR	Any internal exon, plus 300 bp flanking intronic sequence
CryptSplice	Within 60 bp of a canonical splice junction; >100 bp into intron if novel donor/acceptor is created
MMSplice	Any exon, plus 50 bp upstream or 13 bp downstream
S-CAP	Any exon, plus 50 bp flanking intronic sequence
CADD	All loci
TraP	All loci
SpliceAI	All loci

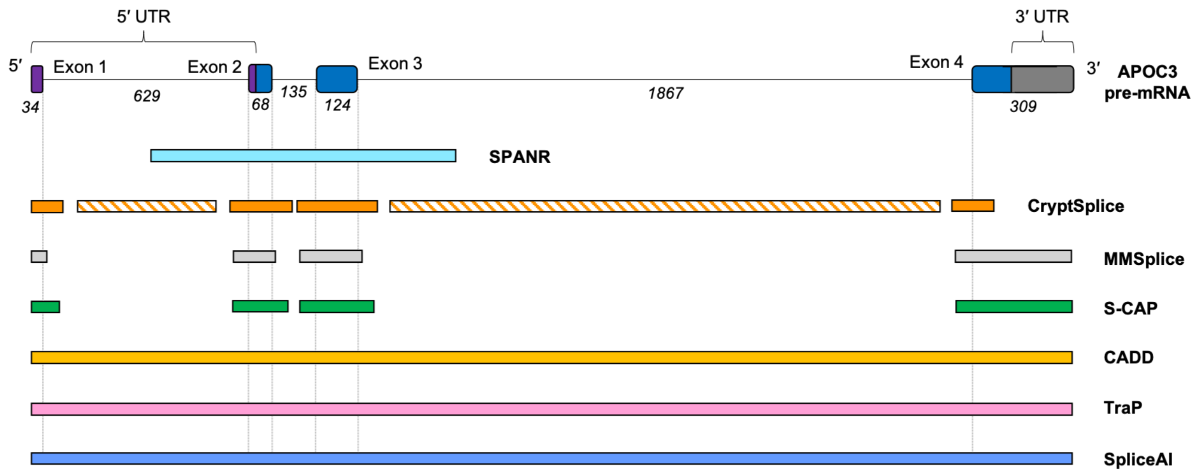


Figure 9. *Location of variants amenable to analysis by splice prediction software.* With diverse underlying training sets and purposes, different splice prediction tools are only able to analyze variants at particular sites in a pre-mRNA transcript. To-scale representation of the loci amenable to analysis by each of 7 tools for the pre-mRNA transcript of the human *APOC3* gene (RefSeq accession NM_000040.3). Dotted lines signify canonical exon-intron boundaries. Hashed bars represent loci where the variant effect can be modeled only if a novel splice donor or acceptor is created. Italicized numbers show exon/intron length in nucleotides. UTR, untranslated region.

The authors suggest a 3-tier threshold system for TraP scoring: variants with a TraP score below 0.495 are considered likely benign. Variants scoring ≥0.495 but below 0.93 are in an intermediate range, representing variants that may possibly have an impact at the transcript level. Variants scoring ≥0.93 are likely pathogenic. When considering intronic variants, the authors suggest a threshold of 0.75 to avoid inclusion of large numbers of false positives.

The authors compared the performance of TraP compared to CADD in distinguishing pathogenic and benign variants, both intronic and synonymous. They demonstrated that matching the specificity of TraP at a 0.495 threshold would give CADD a sensitivity of just 6% or 18.8%, for synonymous and intronic variants, respectively. Thus, it is evident that TraP scoring offers a marked improvement on the CADD model for the prioritization of variants impacting splicing.

In addition, TraP considers the potential impact of variants across multiple transcripts, a feature not considered by many splicing prediction tools. The efficacy of the model is also impressive, particularly given the relatively small size of the training and test sets. While the model works well in identifying pathogenic intronic variants, retraining a second model using such pathogenic intronic variants, rather than synonymous ones, may improve the performance of TraP yet further.

1.4.4.3. SPANR (Splicing-based analysis of variants)

SPANR (Xiong et al., 2015) seeks to model variants impacting cassette exon splicing—the inclusion or skipping of a given internal exon—across a number of human tissues. It achieves this using a Bayesian deep learning model based on the percentage spliced in (PSI, or Ψ) metric, a measure of the percentage of mature mRNA transcripts containing, rather than excluding, a particular exon. This model seeks to maximize a "code quality" metric that is a measure of the improvement of the model to predict Ψ values over a random guesser. SPANR works on a variation of a two-layer neural network, where the hidden layers of the model are common to all tissues, but each tissue has a distinct output layer.

Transcripts from RefSeq (Pruitt et al., 2007) were mined, and Human UniGene data from NCBI analysed, to identify instances of cassette and constitutive exon splicing in the normal human transcriptome, leading to the identification of 10,689 cassette and 33,159 constitutive exons (all flanked by an exon on either side). The Ψ metrics for each of these central exons was then computed genome-wide using RNA-seq data from the Illumina BodyMap 2.0 project (NCBI GSE30611) and used as input for training an ensemble of DNN models. $\Delta\psi$ values representing the predicted change in exon inclusion were then able to be generated, with the paper using $|\Delta\psi \geq 5\%|$ as a general threshold over which a variant is considered to impact cassette exon splicing.

In the original paper, the authors demonstrated the utility of SPANR in the analysis of specific variant cohorts in patients with spinal muscular atrophy (SMA) and Lynch syndrome, implicating common causative variants in these disorders as splice-impacting. They also showed that predicted effects of simulated variants in intron 7 of the SMN2 gene are recapitulated with RT-PCR. They conducted a wider analysis of SNVs in genome data from 5 patients with autistic spectrum disorder and observed an enrichment of splice-impacting variants in genes associated with neurodevelopmental roles, thus demonstrating a wide range of potential uses for the tool in the study of both Mendelian and complex disease.

The model is somewhat limited by the scope of the cassette exon model—a variant must lie within 300 bp of an exon that itself lies between two other exons, meaning variants in first or terminal exons are not analysable. This also renders the model obsolete for analysis of pathogenic variant types such as cryptic splice sites and deep intronic mutations. However, a webserver is provided, allowing easy analysis of small batches of variants, while a tabulated version of the SPANR dataset called SPIDEX, comprising pre-computed scores for all eligible variants in the genome, can be downloaded by the user and used during variant annotation with the ANNOVAR package for larger variant sets (Wang, Li and Hakonarson, 2010). SPANR may thus be a powerful component of a predictive pipeline, but is likely too limited in scope to be considered proof of pathogenicity in isolation.

1.4.4.4. CryptSplice

CryptSplice (Lee et al., 2017) aims to predict the effects of the generation of cryptic splice sites. Namely, it considers three scenarios: the weakening of a canonical site by the introduction of a new splice site nearby, the outcompeting of a canonical site by a novel site, and the introduction of a functional deep intronic splice site.

An SVM forms the basis of CryptSplice, with input data being transformed with an RBF kernel, which was shown to yield the greatest accuracy. To provide probabilistic estimates to accompany classifications, the model was trained using 10-fold cross validation; that is, the training set was randomly divided into 10 equal parts and each part successively used to generate a new model. The distribution of accuracies across different models then formed the basis of probability metrics.

For training, CryptSplice was trained on a series of “true” splice sites derived from the NN269 (Reese et al., 1997) and HS^{3}D (Pollastro et al., 2002) datasets, repositories of splice junctions curated from GenBank annotations following various quality control and cleaning processes. An equal number of “false” sites were derived, consisting of sequences with GT or AG dinucleotides at least 60 bp from a canonical splice site. All features for the model were sequence-based and fell into one of three categories (**Table 2**).

If a cryptic donor or acceptor is created, CryptSplice is able to cover regions >100 bp into the intron (**Figure 9**), lending it some strength over some tools that lack applicability far from splice junctions. However, some weaknesses in the model are apparent—the training junctions, for example, are derived from transcript annotations over 20 years old. Thus, the model may be underpowered to detect weaker splice sites that may not have become part of standard transcript models until more recently, and other tools are likely more effective for analysis of variants lying outside deeply intronic regions.

1.4.4.5. MMSplice (Modular modelling of splicing)

The tool MMSplice (Cheng et al., 2019) aims to model the competitive interaction between splice sites in close proximity, supplementing this with predictions of exon skipping, splicing efficiency (i.e., the proportion of transcripts undergoing, rather than bypassing, splicing at a particular junction) and pathogenicity.

MMSplice has a complex underlying modular architecture containing 6 basic models of the transcript space (**Figure 10a**), covering donor and acceptor sites, plus 3′ and 5′ intronic and exonic sequences. Each was generated by a neural network with 2–4 layers, and all but the donor model had at least one convolutional layer. To generate the donor and acceptor models, all splice donor and acceptor sites present in the GENCODE v24 annotation (Harrow et al., 2012) were derived as examples of positive sites. Random sequences from within the same genes were then used as negative sequences, provided they did not overlap the position of the positive splice sites. The output of these models is a positive or negative score, corresponding approximately to the strength of the presented variant sequence as a donor/acceptor.

To generate the 5′ and 3′ exonic and intronic models, the authors leveraged a massively parallel reporter assay (MPRA) generated by Rosenberg et al. (2015), in which the relative splicing efficiencies of pairs of random 25-mer oligonucleotides were evaluated on both the exonic and intronic sides of an intronic splice junction. These models derive either $\Delta\psi_5$ or $\Delta\psi_3$ metrics, corresponding to the relative usage of a variant sequence as a splice acceptor or donor, respectively, compared to the canonical site.

A series of regression models were then designed based on the output of these models in order to predict variant impact on splicing. Four linear regression models were constructed: one analysed variant impact on exon skipping through analysis of data from the splice analysis pipeline Vex-seq (Adamson et al., 2018); two were designed to predict $\Delta\psi_5$ or $\Delta\psi_3$ values based on cross-referencing of genome and RNA-seq data from the GTEx study (GTEx Consortium, 2013); the fourth leveraged a massively parallel splicing assay, or MaPSy (Soemedi et al., 2017; see *1.5.3.*), to predict splicing efficiency. In addition to this, a logistic regression model to predict pathogenicity was derived based on known pathogenic and benign variants in the splice region, as listed on ClinVar (Landrum et al., 2014). Thus, MMSplice provides a powerful combination of both biological and clinical predictions.

MMSplice is highly intricate and versatile, and is also easily clinically applicable, being able to take variant call format (VCF) files as input, and incorporating both SNV and indel predictions (unlike many tools) to predict a wide range of variant impacts on splicing. However, the training set of all splice junctions in the GENCODE v24 annotation may also contain substantial numbers of false positives where particular transcripts have been computationally predicted and remain experimentally unverified. Furthermore, modelling of competitive splice site interactions using GTEx data was based solely on samples from brain and skin tissue, which may underpower the model for predicting competitive interactions that predominate in other tissue types.

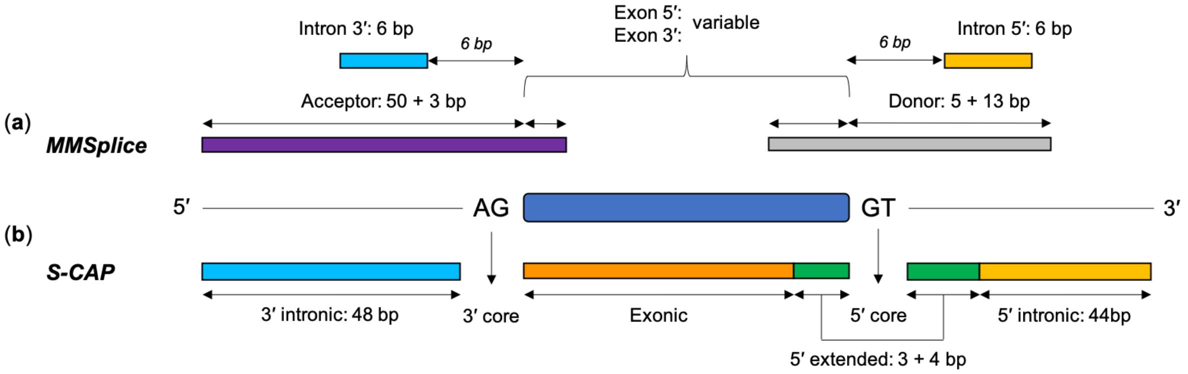


Figure 10. *Compartmentalization of the splice region by S-CAP and MMSplice.* Both MMSplice and S-CAP divide the splice region into six sub-regions, although the length and location of these divisions are different between the two tools. MMSplice (**a**) consists of 6 initial deep neural network modules corresponding to each region, with exonic and intronic modules both trained on the results of a massively parallel reporter assay (MPRA) experiment (Rosenberg et al., 2015) and the acceptor and donor modules trained to predict functional acceptors and donors based on the real and decoy sites in the GENCODE v24 annotation. The scores from all modules are then passed to linear and logistic regression models to predict downstream effects, such as exon skipping, alteration of splicing efficiency, and competitive splice site interactions. S-CAP (**b**) consists of six separate models trained on pathogenic and benign variants curated for each region. The most significant consequence is returned for a given variant. Length of bars not to scale.

1.4.4.6. S-CAP (Splicing clinically applicable pathogenicity prediction)

S-CAP (Jagadeesh et al., 2019) is a splice prediction tool designed to directly predict the pathogenicity of splice-impacting variants. Much like MMSplice, S-CAP compartmentalizes the splicing landscape. In S-CAP, this compartmentalization comprises 6 distinct regions: 3′ intronic, 3′ core, exonic, 5′ core, 5′ extended, and 5′ intronic (**Figure 10b**), all lying within 50 bases of the canonical exon-intron junction. This approach aims to counter the tendency for prioritization of core splice site mutations in most machine learning models, which may understate the pathogenicity of more intronic variants.

The creators of S-CAP took both the Human Gene Mutation Database (HGMD; Stenson et al., 2014) and ClinVar (Landrum et al., 2014) as sources for pathogenic variation, while benign variants were sourced from gnomAD (minor allele frequency ≥1%). The model is trained on 29 different features, classified as chromosome, gene, exon or variant level features. These features include highly tailored analyses, such as the number of rare variants found in the given exonic locus, or the SPANR and CADD scores for the variant. Intolerance of the gene as a whole to mutation is incorporated into the model through the use of pLI (probability of being loss-of-function intolerant; Lek et al., 2016) and RVIS (residual variation intolerance; Petrovski et al., 2013) scores.

In cases of 5′ and 3′ core mutations, the downstream consequence is almost universally impairment of splicing, removing the requirement of evaluating splice impact. This leaves only the question of whether this impairment of splicing is likely pathogenic. This is highly dependent on whether the variant is present in a heterozygous or hemi/homozygous state. To this end, core splice variants are run through two models, one based on a recessive and the other on a dominant inheritance model, and a score returned for each possibility.

Pre-computed scores are available for all variants lying within 1 of the 6 regions considered by the model, and individual thresholds are predefined for analysis of each of these regions. These thresholds, however, are designed for 95% sensitivity, coming somewhat at the expense of specificity and leading to the generation of large numbers of false positives also being identified. There is also substantial variety in the efficacy of the 6 models: exonic and 5′ intronic mutations are particularly difficult to characterize. This is most likely accounted for by the method of generation of these two models, for which variants had to be co-opted from other compartments prior to training, in order to boost an otherwise small pool of pathogenic variants. While S-CAP is underpowered to detect these variant types compared other types, it regardless outperformed SPIDEX, CADD and TraP in both sensitivity and specificity.

Although it doubtless plays a huge part in the efficacy of the tool, the division of the genomic landscape also comes at the expense of universal applicability: variants lying more than 50 bp into the intron are not covered by the model. Despite this, for the cohort of variants lying within these predefined regions, S-CAP has the potential to be a highly effective predictive tool.

1.4.4.7. SpliceAI

The deep learning tool SpliceAI (Jaganathan et al., 2019) analyses each position in a pre-mRNA transcript and evaluates whether it is likely to be a splice donor, acceptor, or neither. The model considers all bases within 50 bp of a presented variant and returns the one with the most substantial gain or loss of acceptor or donor potential as a result of the mutation. The model analyses the impact of a variant on the splicing potential of residues in the surrounding genomic space.

SpliceAI consists of a 32-layer deep residual neural network, a subtype of neural network in which the network is arranged into so-called “residual blocks”—sub-networks containing “skip connections” that output directly to deeper layers in the model. This helps bypass common pitfalls for particularly deep neural networks, such as vanishing/exploding gradients, and also improves the speed with which the network learns (He et al., 2016).

To train the model, the authors selected over 20,287 principal protein-coding transcripts from the GENCODE v24 annotation, and used those from a selection of particular chromosomes (all except chr1, chr3, chr5, chr7, and chr9) as a training set, with the remainder acting as the test set, following removal of paralogs within the set. Each base within these transcripts was designated either a splice donor, acceptor or non-splice site. Four architectures were specifically designed: SpliceAI-80nt, SpliceAI-400nt, SpliceAI-2k, and SpliceAI-10k, where the suffix denotes the total number of bases flanking the variant that are input to the model.

SpliceAI is designed to infer features from the transcript sequence itself; as such, the only input to the model is a coded representation of the variant of interest and the flanking sequence of variable length, dependent on the above choice of model. Scores of gain or loss of acceptor or donor potential are generated for all residues lying within 50 bp of the variant on the pre-mRNA transcript. The residue within this flanking region that experiences the most significant change is then returned for each of these 4 consequences.

The authors demonstrated the ability of SpliceAI to faithfully identity true splice sites from nucleotide sequence alone, allowing recreation of entire gene transcripts; SpliceAI-10k exhibits 95% top-*k* accuracy and a PR-AUC (area under precision recall curve) of 0.98, both markedly high figures. While the authors demonstrate very favourable model performance in comparison to earlier tools, e.g., MaxEntScan (Yeo and Burge, 2004), GeneSplicer (Pertea et al., 2001) and NNSplice (Reese et al., 1997), they did not analyse performance against any more recent tools. Such comparisons will prove very valuable in ascertaining the utility of SpliceAI in clinical practice.

In using a near-agnostic approach to model training, SpliceAI is able to identify features that may not be apparent to most humans. Because of this, it is quite possible that many features of the above tools, such as the modelling of competitive interactions between neighbouring and novel splice sites, are already encompassed within the model. As acknowledged by the authors, however, this agnosticism may mean that certain features incorporated into the model do not truly reflect phenomena with biological meaning. Despite this, the power of the model, as well as the public availability of precomputed scores for all possible single nucleotide substitutions in the genome, suggest that SpliceAI may prove the gold standard for clinical interpretation of splice-impacting variants.

1.4.5. Future prospects for machine learning-based splice prediction tools

The ever-growing range of splice prediction tools complicates variant interpretation by providing a surplus of choices for bioinformatics analysis. Identifying the optimal choice through direct, head-to-head comparisons of these tools is not a simple task. The genomic loci analysable by different tools vary considerably, thus making construction of a universal test variant set difficult without the introduction of missing data points for at least one of the tools. The diverse functions of these tools also complicate comparative analysis. Comparing the performance of a tool predicting competitive splice site interactions with one predicting exon skipping, for example, may not ultimately prove informative.

Despite this, many of the papers describing the above tools do attempt such comparisons. SpliceAI, for instance, significantly outperforms the splice prediction tools GeneSplicer (Pertea et al., 2001), MaxEntScan (Yeo and Burge et al., 2004), and NNSplice (Reese et al., 1997) in both top-*k* accuracy and precision-recall. However, these latter tools were created over a decade ago, when sizeable training datasets were not so readily available, and so may be underpowered in splice prediction. MMSplice (Cheng et al., 2019), meanwhile, shows favourable performance over the similar tool COSSMO (Bretschneider et al., 2018), and S-CAP (Jagadeesh et al., 2019) outperforms SPANR

(Xiong et al., 2015), CADD (Kircher et al., 2014), TraP (Gelfman et al., 2017), and others across all six of its considered regions.

The different approaches adopted by these models offers clinical geneticists the opportunity to consider variant impact from many perspectives, both in terms of the specific splicing consequences predicted by the given model and the value it outputs. Broadly, tools may predict either pathogenicity or splicing impact. Care may need to be taken with the former, as training of a pathogenicity score is reliant on human annotations of pathogenicity, such as through ClinVar (Landrum et al., 2014). These annotations may be inaccurate, and may also suffer from ascertainment bias, whereby the main body of pathogenic variants in the database reflect the current state of our understanding of splice-impacting variants, thus underpowering models in the analysis of less apparent splice variant types. The ACMG have produced detailed guidelines for the scoring of variant pathogenicity (Richards et al., 2015); consideration of splicing impact first and then following these guidelines on a variant-by-variant basis may prove a more robust and sensitive way to characterize pathogenic variants.

Machine learning models are often seen as “black boxes”, in that the inner workings of the model are not discernible to the user, and it is thus difficult for meaningful biological inferences to be made. However, variants flagged by these tools may prove a valuable jumping-off point for research into the mechanisms underlying the inability of earlier tools to correctly predict certain variants.

One such mechanism is the existence of long-distance splicing interactions: SpliceAI has demonstrated that consideration of wider genomic context significantly improves model performance. Such an improvement likely reflects the interactions between *trans*-acting splicing complexes bound across the often substantial lengths of introns, as well as their respective *cis*-acting binding sites (De Conti et al., 2013; Ke et al., 2010). Thus, SpliceAI may provide a useful resource in the investigation of long-range determinants of splicing, and ultimately improve our understanding of splicing in both a healthy and pathogenic context.

Many of these tools share common caveats. Few tools, for example, are able to predict the splice impact of indels, with the exceptions of CADD, MMSplice, and SpliceAI. Future tools will certainly benefit from more thorough consideration of such variants, which may have a significant impact on ultimate transcript structure. Indels affecting the poly-pyrimidine tract (PPT), for example, are known to have significant effects on splicing that may be more marked than the effect of many PPT SNVs, as spacing between the branchpoint and 3′ splice site is crucial for correct assembly of the spliceosome (Coolidge, Seely and Patton, 1997; Bryen et al., 2019).

It should also be noted that atypical splice sites (i.e., those not consisting of GT-AG dinucleotide pairs) comprise just 1% of the body of human introns (Burset et al., 2000), and so do not feature prevalently in training sets. Some tools, such as CryptSplice, actively exclude such introns from model training. Thus, many models may be underpowered to predict changes affecting these low-frequency sites. The effect of variants in AT-AC introns (also known as U12 introns), which are instead processed by the biochemically distinct “minor spliceosome” (Turunen et al., 2013), may be particularly difficult to predict. While the relative occurrence of such introns is low, they nonetheless represent a possible source of pathogenic variants (Verna et al., 2018), with mutations affecting the U12 5′-splice sites of introns in the *STK11* (Hastings et al., 2005) and *TRAPPC2* (Shaw et al., 2003) genes being shown to cause Peutz–Jeghers syndrome and spondyloepiphyseal dysplasia tarda, respectively. Special care may need to be taken, therefore, when considering variants in the vicinity of splice sites for such introns.

A final valuable consideration for models is the inclusion of more personalized and patient-specific prediction of splicing. The single-variant functionality of most of the above tools, for example, neglects to consider the interactions between multiple variants in close (or even distant) genomic space. Studies in mice suggest such interactions between common SNPs (i.e., an individual’s genetic background) and rare variants may underlie phenomena such as incomplete penetrance and variable expressivity (Bourgeois et al., 1998; Doetschman, 2009) in Mendelian disorders. Consideration of these common genomic variants in tandem with variants of interest may allow further clarification of variants of uncertain significance.

1.5. Functional analysis of variant impact on mis-splicing

As described above (see *1.3.5.*), computational predictions cannot be taken as concrete evidence of the splicing impact of a variant. In accordance with ACMG guidelines, variants outside the canonical ± 1 or 2 splice site cannot be presumed to be a null variant. This stipulation may be a reasonable one, given the weaker levels of constraint found at sites distal to canonical splice acceptors/donors (see *1.2.1*). Thus, functional investigation is necessary to provide evidence that a variant causes, or does not cause, loss of function through generation of aberrant splicing. A handful of tools are available to functionally assess splicing impact, and, as with sequencing methodologies, these functional splicing assays vary in scope, ranging from analysis of individual transcripts to the surveying of entire transcriptomes.

1.5.1. RT-PCR: targeted amplification and sequencing of transcripts of interest

One of the most direct ways to scrutinise splicing impact is to look directly at the isoform structure of RNA extracted from patient biosamples. A commonly used technique in this regard is *reverse transcription-PCR*, or *RT-PCR* (**Figure 11a**). In RT-PCR, reverse transcriptase is used to convert the RNA in a patient sample into cDNA; typically, this targets poly-adenylated RNA, which is bound by an oligo-dT primer prior to addition of reverse transcriptase. The generation of double-stranded DNA then acts as the starting material for a conventional PCR reaction: using primers specific to the transcript of interest, the transcript sequence is amplified and can then be sequenced.

1.5.2. Minigenes and midigenes: cell-based assays of mis-splicing

As described above, the splicing process shows remarkable conservation across most eukaryotic lineages (*see 1.2.2.*). This conservation is leveraged in two targeted, vector-based methodologies: *minigenes* and *midigenes* (**Figure 11b**). In both techniques, DNA sequences harbouring exons of interest are cloned into a plasmid vector that is subsequently transformed into a model eukaryote of choice, most commonly the brewer's yeast *Saccharomyces cerevisiae* (Gaildrat et al., 2010, Smith and Lynch, 2014)*.* RNA can then be harvested from cells and RT-PCR carried out to assay potential changes in transcript structure. Mini- and midigenes offer an alternative targeted approach to RT-PCR alone when expression of the transcript of interest is expected to be poor in the patient biosample.

Minigenes and midigenes differ in the size of the investigated genomic region, with minigenes generally encompassing only one or two exons (Gaildrat et al., 2010), and midigenes being associated with longer, multi-exon loci, or exons flanked by particularly long introns (Sangermano et al., 2019, Verbakel et al., 2019). Notably, minigenes and midigenes may produce conflicting results: in one study of VUSs in patients with the inherited retinopathy Stargardt disease, variants previously observed to have no effect on splicing when investigated with minigenes did exhibit significant and pathogenic mis-splicing of the *ABCA4* gene when the experiment was repeated with the insertion of a wider genomic region in the plasmid construct (Sangermano et al., 2019).

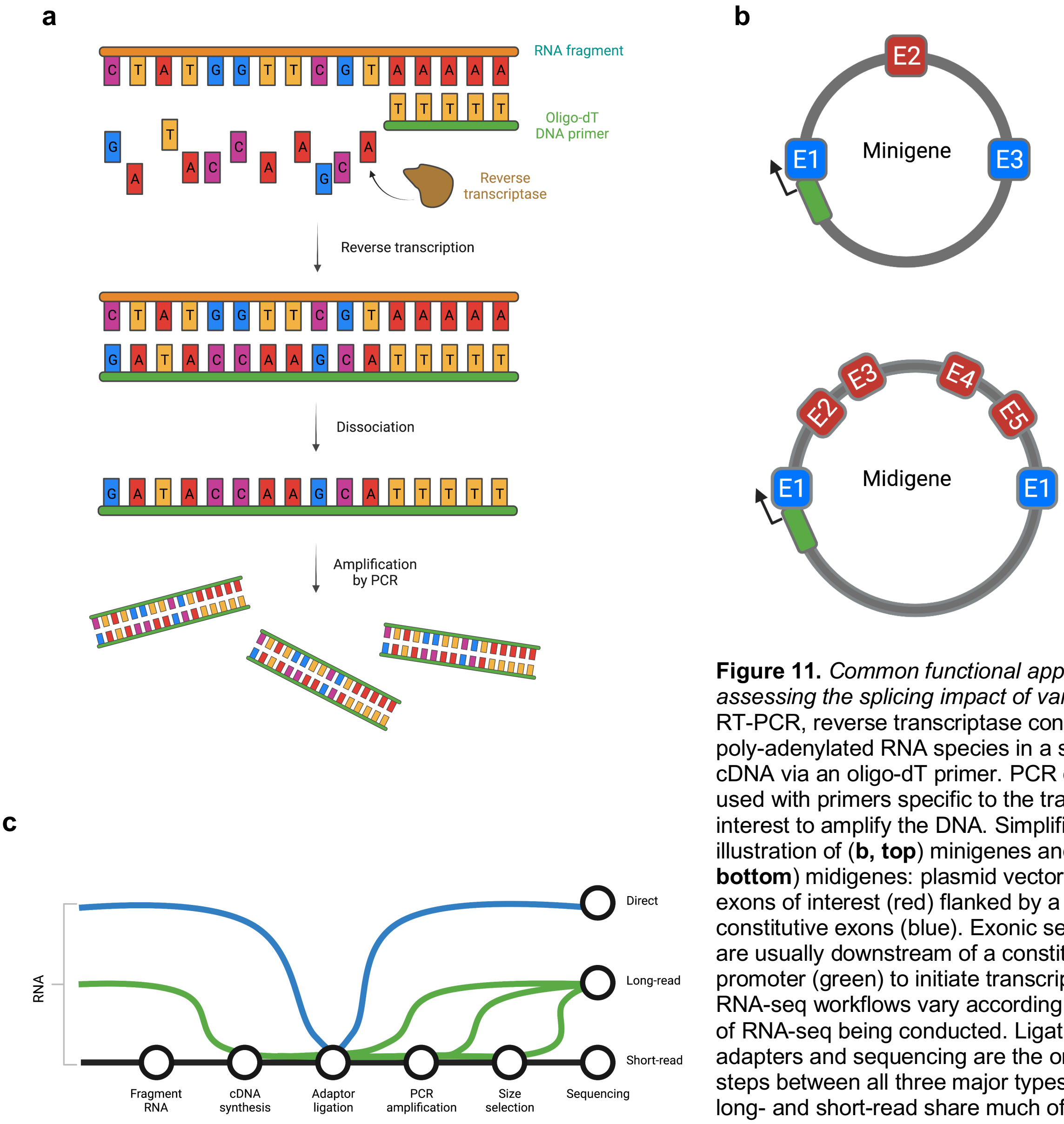


Figure 11. *Common functional approaches in assessing the splicing impact of variants.* (**a**) In RT-PCR, reverse transcriptase converts all poly-adenylated RNA species in a sample into cDNA via an oligo-dT primer. PCR can then be used with primers specific to the transcript of interest to amplify the DNA. Simplified illustration of (**b, top**) minigenes and (**b, bottom**) midigenes: plasmid vectors containing exons of interest (red) flanked by a pair of constitutive exons (blue). Exonic sequences are usually downstream of a constitutive promoter (green) to initiate transcription. (**c**) RNA-seq workflows vary according to the type of RNA-seq being conducted. Ligation of adapters and sequencing are the only shared steps between all three major types, although long- and short-read share much of the sample preparation stages. (**c**) adapted from Stark et al. (2019). Created using BioRender.com

1.5.3. High-throughput workflows for functional investigation of splice-impacting variants

While RT-PCR and mini-/midigenes are able to validate the splicing impact of variants, they are generally only able to do so for one variant at a time. Soemedi et al. (2017) devised a novel workflow named a Massively Parallel Splicing Assay (or MaPSy), using which they analysed the splicing impact of a total of 4,964 published disease-causing exonic variants.

The MaPSy study analysed the effect of the exonic variants both *in vivo* and *in vitro*. Exons of interest and their flanking intronic sequence (at least 55 bp upstream and 15 bp downstream) were recreated using oligonucleotide synthesis and inserted into a reporter; in the *in vitro* assay, this consisted of an upstream adenoviral exon, while the *in vivo* assay comprised the same upstream adenoviral exon downstream of an enhancer and promoter, and a downstream intron and exon derived from the *ACTN1* gene. *In vivo* constructs were transfected into human tissue cells, and *in vitro* constructs added to a solution consisting of HeLa cell nuclear extract, which contained the splicing machinery required for assessment of mis-splicing events.

The group did also note that the most common splicing outcome *in vivo* was exon skipping, while the *in vitro* assay more commonly showed an absence of splicing (resembling retention of the upstream intron) due to the absence of a downstream exon. Further, there was imperfect concordance between both approaches, with approximately 80% of *in vivo* findings being recapitulated *in vitro*.

A similar approach named Vex-seq was designed by Adamson et al. (2018). Unlike MaPSy, in which linear RNA was directly transfected into cells or added to nuclear extract, Vex-seq involved insertion of sequences of interest into barcoded plasmids, which were subsequently transfected into cells to assess the impact on splicing of 2,059 human variants.

High-throughput techniques hold promise as a tool to investigate clinical variants of interest, but as yet remain unused by most genomic centres due to their complexity and the infrastructure required to conduct them. The maintenance of a cell line for the MaPSy *in vivo* test, for example, requires personnel and has cost implications. Further, limits to the length of the test exon and flanking intronic sequences mean that both MaPSy and Vex-seq are likely unsuitable in their current forms for investigation of deeply intronic variation, or mis-splicing of long exons.

1.5.4. RNA-seq: transcriptome-wide snapshots of isoform structure

RT-PCR, mini/midigenes, MaPSy and Vex-seq are effective approaches for targeted analysis of splice variant impact, but all by their nature require prior knowledge of the transcript or variant of interest. They are unsuitable, therefore, for discovery of unidentified mis-splicing events, as may be the case with the large proportion of unsolved Mendelian disease cases. RNA sequencing (RNA-seq) overcomes this limitation to some extent by aiming to capture and sequence all or the majority of RNA transcripts present in a given sample.

As with DNA sequencing, RNA-seq approaches can be broadly categorised as long-read or short-read. 95% of published RNA-seq datasets on the Short Read Archive (SRA) were generated using the Illumina short-read sequencing workflow (Leinonen et al., 2011), which involves the generation of short reads through fragmentation of the initial RNA sample. Long-read RNA-seq approaches, such as those employed by Oxford Nanopore and PacBio have already been used to resolve structural variants (Dutta et al., 2019; Merker et al., 2018) and to sequence over tandem repeats (De Roeck et al., 2017; Ishiura et al., 2018). In both short-read and long-read sequencing, conversion of RNA to cDNA is required to stabilise the target transcripts and facilitate the downstream sequencing reaction. Recent work in *Arabidopsis thaliana* and the viral pathogen HSV-1 has also shown the capacity of the Oxford Nanopore platform to conduct *direct RNA-seq* (Depledge et al., 2019, Parker et al., 2021). In this approach, RNA molecules are fed through the Nanopore without prior conversion to cDNA. This allows the identification of epigenetically modified RNA bases, such as N6-methyladenine (m^{6}A), a methylated version of adenine with wide-ranging roles, including modulation of gene expression (Roignant and Soller, 2017), alternative splicing (Liu et al., 2015, Liu et al., 2017) and transcript stability (Wang et al., 2014). Such bases are usually replaced with their unmodified counterparts during the RNA-to-cDNA conversion. The variation in experimental protocol according to RNA-seq methodology is depicted in **Figure 11c**.

Most RNA-seq workflows begin with the enrichment of desirable RNA species. In the absence of an enrichment step, up to 95% of sequenced reads may map to ribosomal RNAs, which are the most abundant RNA species in human cells (Morlan et al., 2012), thus obscuring analysis of other RNA species. Two of the most common methodologies to prevent the predominance of rRNA in samples are oligo-dT-enrichment, in which poly-T probes are used to capture and enrichment the poly-A tails found on many RNA species (not including rRNA), and ribodepletion, in which oligonucleotide probes complementary to the major rRNA subtypes enable immobilisation of rRNAs in a column and elution of remaining rRNA species.

While oligo-dT-enriched RNA-seq datasets are able to capture the majority of mature RNA transcripts in the cell (provided sequencing depth is sufficient), they do result in the exclusion of some other RNA species which lack poly-A tails, such as miRNAs and enhancer RNAs (eRNAs). Ribodepletion, conversely, is often considered a *whole-transcriptome* methodology, as it aims to solely exclude rRNAs from the resulting library, while retaining most other RNA species. By capturing transcripts independently of their poly-A tails, ribodepletion RNA-seq datasets are also able to capture immature mRNA transcripts that may be un-spliced or partially spliced. This may complicate analysis of mis-splicing, as sequencing reads may map across unspliced junctions and within introns (Zhang et al., 2018). It may be difficult, for instance, to delineate genuine intron retention events from background levels of intronic coverage.

1.5.4.1. Bioinformatics steps in the processing of RNA-seq datasets

As with DNA sequencing methodologies, RNA-seq requires the alignment of reads back to the genome. Unlike DNA sequencing, however, reads derived from RNA can be expected to contain large numbers of discontinuities relative to the DNA sequence when aligned back to the genome. Such reads are called *split reads*, and represent the apparent jump in mapping co-ordinates that is observed when looking across the genomic boundaries of a splice junction. A host of *split-read aligners* have been developed to assist in the alignment of RNA-seq data. Many of the genomic aligners described above are also capable of handling split reads, although they may not originally have been designed to do so. Popular short split-read aligners include STAR (Dobin et al., 2013), HISAT2 (Kim et al., 2019) and TopHat2 (Kim et al., 2013).

Aligned RNA-seq datasets are powerful tools for investigating many aspects of transcript dynamics, becoming more so with the presence of large control datasets. Transcript abundance can be measured using tools like edgeR (Robinson et al., 2010), RNA-SeQC (DeLuca et al., 2012) and RSEM (Li and Dewey, 2011). Application of statistical models to this quantification data can allow identification of genes exhibiting differential expression between test conditions, or, in the context of disease, between healthy and affected individuals. Calculation of allelic imbalance, too, can shed light on the dynamics of transcript degradation, and both can signpost potentially pathogenic events (see *1.5.4.2*).

1.5.4.2. Identifying pathogenic variants from RNA-seq data

A major challenge in the clinical integration of RNA-seq is the identification of individual, or small numbers of, pathogenic mis-splicing events from among the hundreds of thousands of splicing events observed in a single sample. Recent years have seen a surge in the number of bioinformatics tools aiming to identify *splicing outliers* that may constitute pathogenic mis-splicing events (Cummings et al., 2017; Jenkinson et al., 2020; Mertes et al., 2021). Many of these approaches attempt to fit a statistical distribution to the read counts observed spanning canonical and non-canonical junctions, and identify outliers along that distribution. These may constitute rare events, but this does not strictly imply pathogenicity.

As discussed above (see *1.3.5.2.*), while explicit provision for splice-impacting variants is largely not covered by ACMG guidelines, there are many characteristics of mis-splicing events that can shed light on the likelihood of their pathogenicity. The impact of a change in splicing is dependent on the penetrance of the event: a mis-splicing event affecting only a small proportion of transcripts is less likely to cause disease. Conversely, high penetrance may support, but does not necessarily imply, pathogenicity. The examination of mis-splicing event penetrance can be undermined by NMD by leading to ostensibly sub-pathogenic levels of aberrant isoform structure. However, two signatures of NMD can be used to identify transcripts in which aberrant transcripts are being degraded: firstly, NMD naturally leads to a reduction in transcript levels. Using quantification methodologies like those described above, genes with significantly decreased total transcript level can be identified and flagged as potential NMD targets and investigated further. A downstream consequence of allele-specific decay is the presence of allelic imbalance, which can be detected by the relative read counts supporting the existence of nearby SNPs, or the variants in question if they are in an exonic region.

There are thus a host of signals that may facilitate the interpretation of pathogenic mis-splicing events. However, our current understanding of what signals characterise pathogenic mis-splicing is incomplete, and more work is needed on analysing known pathogenic events to delineate these characteristics.

1.5.4.3. Tissue selection in the clinical application of RNA-seq

RNA-seq holds particular promise as a diagnostic strategy in simultaneously facilitating both corroboration of the splicing impact of previously identified variants, and identification of the splicing impact of variants that may have been missed by upstream sequencing approaches. The sequencing of patient peripheral blood samples is commonplace in the clinic; sample preparation workflows such as the PaxGene DNA and RNA extraction kits (Harrington et al., 2020) allow quick and simple isolation of transcriptomic material from the most easily available biosample, human blood. While blood is the most convenient source of RNA, it also has a major caveat. Whole blood samples have a highly specific transcriptomic profile; that is, their repertoire of expressed transcripts largely does not overlap lap with other tissues. Its value in clinical diagnostics, therefore, is limited in certain disease contexts. There is thus a need to develop a method of readily identifying the optimum tissue of choice for RNA-seq-led investigations.

The presence of tissue-specific alternative splicing events and expression patterns has significant implications for the clinical implementation of RNA-seq. For the effect of a variant on splicing to be determined, a sample must (a) express transcripts that may be expected to predominate in the affected tissue, and (b) exhibit coverage of the specific junction that will likely be disrupted. As yet, few frameworks exist to quantitatively evaluate how well covered clinically relevant transcripts are in a tissue of interest.

1.6. Research aims

Multiple factors between variant identification and interpretation impact our ability to assign pathogenicity to splice-impacting variants. Selected diagnostic methodologies that do not include intronic coverage risk omitting large sections of potentially pathogenic deeply intronic variation, while approaches that do cover introns (such as WGS) result in vast numbers of variants of uncertain significance, thus leading to difficulties in variant interpretation.

The work described here aims to develop novel bioinformatics approaches to address issues often encountered at different stages of splice variant identification and interpretation.

The first is the investigation of the efficacy of novel predictive tools. Given the older nature of the splice variant predictors encompassed by the ACMG guidelines (Richards et al., 2015), there is substantial scope for evaluation of the efficacy of more up-to-date predictive models. Considering the increased complexity of these machine learning paradigms, it is possible that modern predictive tools may significantly outperform existing tools in identifying splice variation. The integration of more accurate predictive approaches into a clinical setting may ultimately lead to an increase in the molecular diagnosis rate.

The second aim is to identify subsets of variants for which these tools perform poorly, and generate bespoke analyses that can outperform them.

The final aim is to develop a bioinformatics framework to guide diagnostic decision-making around RNA-seq. The goal is to develop a resource to inform clinicians of the likely benefits, or lack of benefits, in adopting a transcriptome-led approach in the diagnosis of a given patient. Through this, we hope to boost pathogenic variant identification, and allow the diagnosis of cohorts of previously undiagnosed patients.

Chapter 2

Comparison of *in silico* strategies to prioritize rare genomic variants impacting RNA splicing

2. Comparison of *in silico* strategies to prioritize rare genomic variants impacting RNA splicing

2.1. Abstract

The development of computational methods to assess pathogenicity of pre-messenger RNA splicing variants is critical for diagnosis of human disease. We assessed the capability of eight algorithms, and a consensus approach, to prioritize 249 variants of uncertain significance (VUSs) that underwent splicing functional analyses. The capability of algorithms to differentiate VUSs away from the immediate splice site as being 'pathogenic' or 'benign' is likely to have substantial impact on diagnostic testing. We show that SpliceAI is the best single strategy in this regard, but that combined usage of tools using a weighted approach can increase accuracy further. We incorporated prioritization strategies alongside diagnostic testing for rare disorders. We show that 15% of 2783 referred individuals carry rare variants expected to impact splicing that were not initially identified as 'pathogenic' or 'likely pathogenic'; 1 in 5 of these cases could lead to new or refined diagnoses.

2.2. Introduction

Pinpointing disease-causing genomic variation informs diagnosis, treatment and management for a wide range of rare disorders, and helps bring an end to the "diagnostic odyssey" undergone by some Mendelian disease patients. Molecular testing, in a healthcare setting, now frequently includes genome and exome sequencing (Lee et al., 2015, Yang et al., 2014, Turnbull et al., 2018). Accurate interpretation and categorization of identified variants remains a key limiting factor despite the availability of guidelines for variant analysis (Richards et al., 2015, Tavtigian et al., 2018).

The capability to interpret variation within the non-coding genome is particularly challenging. Variant interpretation is hindered by the vast number of rare/novel non-coding variants identified in each individual (Taylor et al., 2015, Ellingford et al., 2016b), the depleted levels of evolutionary conservation within non-coding regions (Short et al., 2018), and our current lack of understanding of the motifs and interactions that are required for appropriate control of gene expression and regulation (ENCODE, 2012, Kundaje et al., 2015).

Intragenic genomic variants have the potential to impact splicing (Faustino and Cooper, 2003), the ubiquitous process in eukaryotic cells of converting nascent pre-messenger RNA (pre-mRNA) molecules into mature messenger RNA (mRNA) which can be transported out of the nucleus to provide a template for protein synthesis. Genomic variation in protein-coding, splice junction and intronic regions of genes can disrupt normal splicing mechanisms and underpin the onset of rare disease (Stenson et al., 2014). Known mechanisms of splicing disruption include the introduction of cryptic splice sites, disruption of canonical splice acceptor and donor sites, and the disruption of other motifs essential for splicing, e.g. branchpoints and the polypyrimidine tract (Stenson et al., 2014). The significant impact these events have on transcript and protein structure means such disruption is likely to be pathogenic when in transcripts of genes associated with loss-of-function disease mechanisms.

This has already been observed in many disease types, for example in autism and intellectual disability (Jagadeesh et al., 2019), and rare ophthalmic disorders (Weisschuh et al., 2021).
A number of computational tools have been developed to assist in the interpretation of genomic variation impacting splicing, and these tools have been expanded recently to include an array of machine learning tools that have been trained to prioritize splice disrupting variation through diverse means (Cheng et al., 2019; Jagadeesh et al., 2019; Jaganathan et al., 2019; Lee et al., 2017; Xiong et al., 2015). Developing standards and recommendations for variants in non-coding regions is an important and emerging area for genome diagnostic services. However, in a similar manner to guidance for missense variants, *in silico* tools may be used as supporting evidence (*PP3*) to prioritize variants that impact splicing and can thereby assist in variant classification. While the initial reports of these *in silico* prioritization tools have shown promising results, there is yet to be a formal assessment of their integration, utilization and comparative performance in clinical environments.
The aim of this study was to compare the performance of nine *in silico* strategies, including eight state-of-the-art algorithms and a consensus approach, to prioritize variants impacting splicing. By applying these findings to known cohorts of variants identified through clinical testing, we aimed to identify the likely diagnostic benefit of routine integration of bioinformatics splicing predictions into diagnostic pipelines.

2.3. Materials and methods

Patient recruitment and genomic variant dataset generation

All individuals included in this study have provided informed written consent for the analysis of relevant disease-causing genes through tertiary healthcare centers within the UK. All genetic testing procedures have been approved by and are available through the UK National Health Service and were performed in a UK Accreditation Service Clinical Pathology Accredited Medical Laboratory (North West Genomic Laboratory Hub, Manchester, UK; ISO 15189:2012; UKAS Medical reference 9865). All data collected is part of routine clinical care and all investigations were conducted in accordance with the tenets of the Declaration of Helsinki. Analyses to improve genomic services, as reported in this study, have been approved by the North West Research Ethics Committee (11/NW/0421 and 15/YH/0365). Patients reported in individual case reports have provided informed written consent for publication. All individuals with genome sequencing datasets have consented through the Genomics England 100,000 Genomes Project.

Patients were identified with ‘variants of uncertain significance’ (VUSs) according to ACMG guidelines for variant interpretation (Richards et al., 2015). Variants were generated through genome sequencing or gene panel sequencing (see *Whole genome sequencing* and *Gene panel sequencing*, below). All variants investigated are reported in **Supplementary Table S1** and their HGVS cDNA nomenclature and genomic co-ordinates (GRCh37 and GRCh38) were validated using VariantValidator (Freeman et al., 2018).

Whole genome sequencing

Whole genome sequencing datasets were generated through the UK 100,000 Genomes Project (Turnbull et al., 2018), using Illumina X10 sequencing chemistry. Sequencing reads were aligned to build GRCh37 of the human reference genome utilizing Isaac (Raczy et al., 2013). Small variants were identified through Starline (SNVs and small indels ≤ 50 bp), and structural variants were identified utilizing Manta (Chen et al., 2016) and Canvas CNV caller (Roller et al., 2016). Variants were annotated and analyzed with the Ensembl Variant Effect Predictor (v75), bcftools and bespoke Perl scripts within the Genomics England secure research embassy.

Gene panel sequencing

Enrichments were performed on DNA extracted from peripheral blood using Agilent SureSelect Custom Design target-enrichment kits (Agilent, Santa Clara, CA, USA). Enrichment kits were designed to capture known pathogenic intronic variants and the protein-coding regions +/-50 nucleotides of selected NCBI RefSeq transcripts; conditions tested included inherited retinal disease (105 genes or 176 genes), ophthalmic disorders (114 genes), cardiac disorders (72 genes comprised of 10 sub-panels) and severe learning difficulties (82 genes). All genes tested and relevant testing strategies are available through the UK Genetic Testing Network (https://ukgtn.nhs.uk/). All samples included in the large cohort analysis were generated through a previously described methodology, (Ellingford et al., 2016a) and had been completed prior to August 2017. Briefly, samples were pooled and paired-end sequencing was performed using the manufacturer protocols for the Illumina HiSeq 2000/2500 platform (Illumina, Inc., San Diego, CA, USA). Sequencing reads were demultiplexed with CASAVA v.1.8.2. and aligned to the GRCh37 reference genome using Burrows-Wheeler Aligner short read (BWA-short v0.6.2; Li and Durbin, 2009, Li, 2013) software before duplicate reads were removed using samtools v0.1.18. The detection and clinical analysis of single nucleotide variants and small insertions/deletions was performed as described previously (Ellingford et al., 2016a, Gillespie et al., 2014), and in accordance with ACMG guidelines for variant interpretation (Richards et al., 2015). During variant analysis, we considered inheritance modes associated with monogenic disorders available in OMIM (https://omim.org/) or PanelApp (https://panelapp.genomicsengland.co.uk/), the zygosity of identified variants, additional variants identified to impact the same gene, phenotype-genotype correlations and scores determined by *in silico* splicing tools. We identified rare variants within our cohort for prioritization (<20 heterozygous variants and <10 homozygous variants) by each of the *in silico* splicing prediction tools, resulting in 18,013 unique variants and 43,744 total variants (42,281 het and 1,463 hom). The region of impact for each rare variant was extracted from S-CAP pre-computed files where available, (Jagadeesh et al., 2019) or determined through Ensembl Variant Predictor (v75) for specified transcripts where unavailable through S-CAP.

In silico *splicing prediction scores*

We utilized scores available from CADD (Kircher et al., 2014), SpliceAI (Jaganathan et al., 2019), SPIDEX (Xiong et al., 2015), S-CAP (Jagadeesh et al., 2019), MMSplice (Cheng et al., 2019), TraP (Gelfman et al., 2017), KipoiSplice (Avsec et al., 2019) and MaxEntScan (Yeo and Burge, 2004) to prioritize the 249 variants (we noted on revision that one duplicate variant existed in our dataset). Where multiple scores were available for a variant from the *in silico* tool, we selected the highest for consideration. To enable comparisons of tool performance and correlation between scores, we converted negative values from SPIDEX, MaxEntScan and MMSplice to positive integers. Whilst these conversions removed directional impact information, i.e. reduced or increased splice site usage, they still reflected the absolute splicing impact of variants. Where scores were unavailable, we assigned the variant a score of 0, i.e. no impact could be predicted. Pre-defined thresholds were applied to determine whether a variant was ‘disruptive’ or ‘undisruptive’ to splicing, as suggested by the authors of the original papers (Xiong et al., 2015, Jaganathan et al., 2019), by recent refinements of thresholds (Jagadeesh et al., 2019), or through nationally recommended guidelines (**Supplementary Table S4**). A consensus score was generated by considering whether the variant exceeded the threshold of each *in silico* prediction tool. ROC curves were generated and compared using the pROC package in R. A comparison of accuracy of the tools was performed through 2,000 iterations of sampling with replacement for the 249 samples. Statistical differences in accuracy were identified through the Kruskal-Wallis test in R.

A novel scaled metric was generated for each variant:

$$score = \sum_{i=1}^{n} {}^{x_i}/_{max_i}$$

Where, *n* = a given combination of the nine prediction strategies, *max* = maximum score from prediction tool *i*, and *x* = variant score from prediction tool *i*. For example, for a variant with a SpliceAI score of 0.85 (the maximum SpliceAI score being 1) and above the threshold of 5/8 tools using the consensus approach:

$$score = {}^{0.85}/_{1} + {}^{5}/_{8} = 1.475$$

RNA investigations

Appropriate functional assays were selected after consideration of gene expression profiles in GTEx (https://gtexportal.org/home/), and the availability of relevant patient samples. We performed assessments on available patient samples or through cell-based minigene assays.

RNA investigations from patient samples – LCLs and blood

Lymphoblast cell cultures were established for control samples and probands. RNA was extracted using the RNeasy® Mini Kit (Qiagen, UK, Catalogue No. 74104) following the manufacturer's protocol. RNA was extracted from whole-cell blood using the PAXgene™ Blood RNA System Kit (Qiagen, UK. Catalogue No. 762174), following the manufacturer’s protocol for control samples and probands. Extracted RNA was reverse transcribed using the High Capacity RNA to cDNA Kit (Applied Biosystems, UK. Catalogue No. 4387406) following the manufacturer's protocol. Gene specific primers (available on request) amplified relevant regions of the genes being investigated. PCR products were visualized on an agarose gel using a BioRad Universal Hood II and the Agilent 2200 Tapestation. Visualized bands were cut out and prepared for capillary sequencing on an ABI 3730xl DNA Analyzer.

RNA investigations using cell-based minigene assays

Assays were designed to amplify appropriate genomic regions from patient DNA templates. For variants nearby to wild-type exons, we amplified regions containing one or multiple exons along with flanking ~200 intronic nucleotides. For deeply intronic variants we amplified regions containing at least 500bp of flanking intronic sequence. Primer sequences are available upon request. All regions were amplified from patient DNA templates. For homozygous variants, we also generated a minigene plasmid from a control DNA template. Amplified fragments were checked for size using gel electrophoresis, purified using the QIAquick Gel Extraction kit (Qiagen, UK, Catalogue No. 28706) and then cloned into a customized minigene plasmid (a derivative of the pSpliceExpress vector; Kishore et al., 2008) containing an RSV-promoter and two control exons (rat insulin exons 2 and 3) using the NEBuilder® HiFi DNA assembly (NEB, E2621). Amplified fragments were inserted between the two control exons. Plasmids were transformed into competent bacteria (XL-1 blue) and incubated overnight at 37°C on LB plates containing Carbenicillin. Individual colonies were cultured overnight before isolation of plasmid DNA using the GenElute™ miniprep kit (Sigma-Aldrich, Catalogue No. PLN350). Purified plasmids were Sanger sequenced to confirm successful cloning and identify plasmids containing the wild-type and variant sequence. Plasmids were transiently transfected into HEK-293 cells using Lipofectamine, and incubated for up to 48h in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum at 37°C and 5% CO_2.

RNA was isolated using TRI Reagent® and further purified using the RNeasy Mini Kit (Qiagen, UK, Catalogue No. 74106) which included a DNase digestion step. cDNA was synthesized from up to 4µg of purified RNA using SuperScript™ reverse transcriptase (ThermoFisher Scientific, Catalogue No. 18091200) and subsequently amplified by Phusion high-fidelity polymerase (ThermoFisher Scientific, Catalogue No. F553) using primers designed to amplify all minigene transcripts. PCR products were visualized by electrophoresis on a 1-2% agarose gel and purified using the QIAquick Gel Extraction kit. Purified PCR products were Sanger sequenced and aligned to the reference sequence for the minigene vector using the SnapGene software suite and assessed for differences in splicing between wild-type and variant minigene constructs.

Comparison with GTEx datasets

Variants identified in GTEx v7 datasets were cross-referenced with prioritized variants from our cohort. FASTQs were downloaded from the Database of Genotypes and Phenotypes (dbGaP) under the project accession phs000424.v8.p2 for GTEx control individuals carrying prioritized variants. RNA-seq datasets for samples carrying prioritized variants were identified, and the TPM value of the tissues available were considered. RNA-seq data from tissues with a TPM value > 5 were considered and FASTQ datasets were processed as described previously (Cummings et al., 2017). Read alignments were visualized in IGV and Normalized Read Count (NRC) and intron retention levels were quantified. NRC is calculated as the proportional usage of non-canonical splice junctions compared to canonical splice junctions for any given site. NRC and intron retention levels for individuals carrying prioritized variants were compared to 10 control individuals in the GTEx dataset.

2.4. Results and Discussion

2.4.1. Functional assessment of variants of uncertain significance identified through clinical genetic testing strategies

First, we ascertained and performed functional analyses for 249 VUSs to observe their impact on splicing (**Supplementary Table S1**). To the best of our knowledge, this is the largest set of VUSs that have been functionally interrogated for impact on splicing as part of diagnostic services for individuals with rare disease. All VUSs investigated are in genes where loss-of-function is an expected mechanism of disease causation. Variants had been identified in individuals undergoing genome sequencing and targeted gene panel analysis, with diverse phenotypic presentations including familial susceptibility to breast cancer (MIM #604370), syndromic disorders such as Marfan syndrome (MIM #154700) and isolated inherited retinal disorders such as retinitis pigmentosa (MIM #300029). The approaches for VUS functional analysis are described elsewhere (Wai et al., 2020) and in **Supplementary Table S1**. We observed that 80/249 (32%) of the VUSs significantly impacted splicing, and as a result may be reclassified as 'likely pathogenic' according to ACMG guidelines for variant interpretation (Richards et al., 2015). This formal reclassification is not conducted as part of this study which focused on the capability of *in silico* tools to distinguish variants which impact splicing (true positives) and variants which did not impact splicing (true negatives). All VUSs impacted regions outside of canonical splice acceptor and donor sites, and included examples of deeply intronic cryptic splice sites, exonic cryptic splice sites and branchpoint variants. In some cases, functional investigations demonstrated a range of consequences on mRNA splicing (**Figure 12**), reinforcing that the precise effect of splicing variants is an important piece of evidence for consideration during clinical variant interpretation that, in the future, may enable refinements in appropriate targeted treatments (Shen and Corey, 2018, Bauwens et al., 2019).

2.4.2. Assessment of in silico prediction strategies to prioritize variants of uncertain significance

We obtained *in silico* prediction scores for each of the 249 functionally assessed variants using eight *in silico* prioritization algorithms (**Supplementary Table S2**) and calculated sensitivity, specificity and receiver operating characteristic area under the curve (AUC), observing significantly variable performances (**Figure 13**). Pairwise statistical comparisons of AUC for the 249 functionally assessed VUSs, after Bonferroni correction for multiple testing, demonstrated that SpliceAI outperformed other single algorithm approaches (**Figure 13**; **Supplementary Table S2**). The AUC analysis for single algorithms calculated the optimal score (based on Youden's J statistic, as calculated using the pROC software package) for each of the algorithms to distinguish between true positives (80 variants shown to impact splicing in our functional assays) and true negatives (169 variants shown functionally not to impact splicing) in this dataset. We acknowledge that splicing machinery may be influenced by cell- or

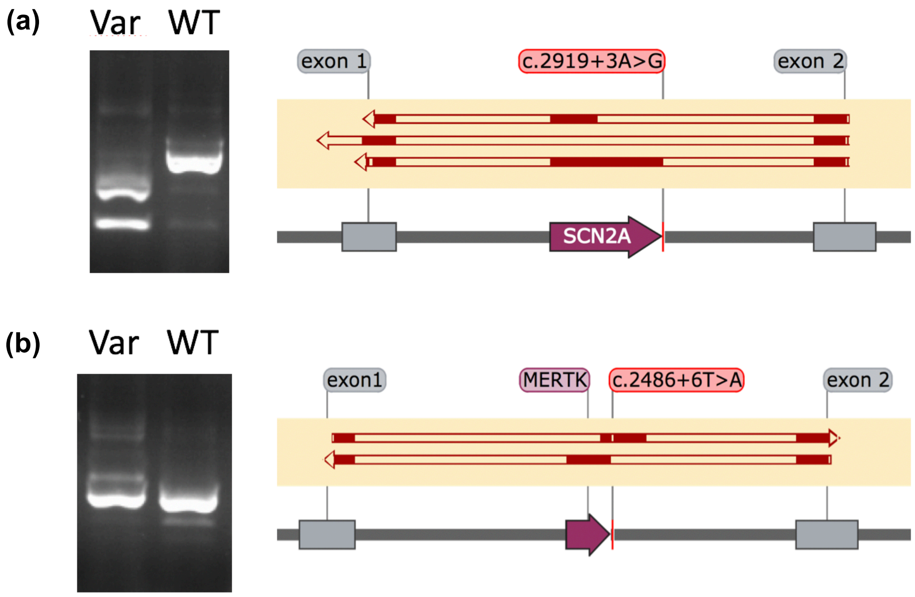
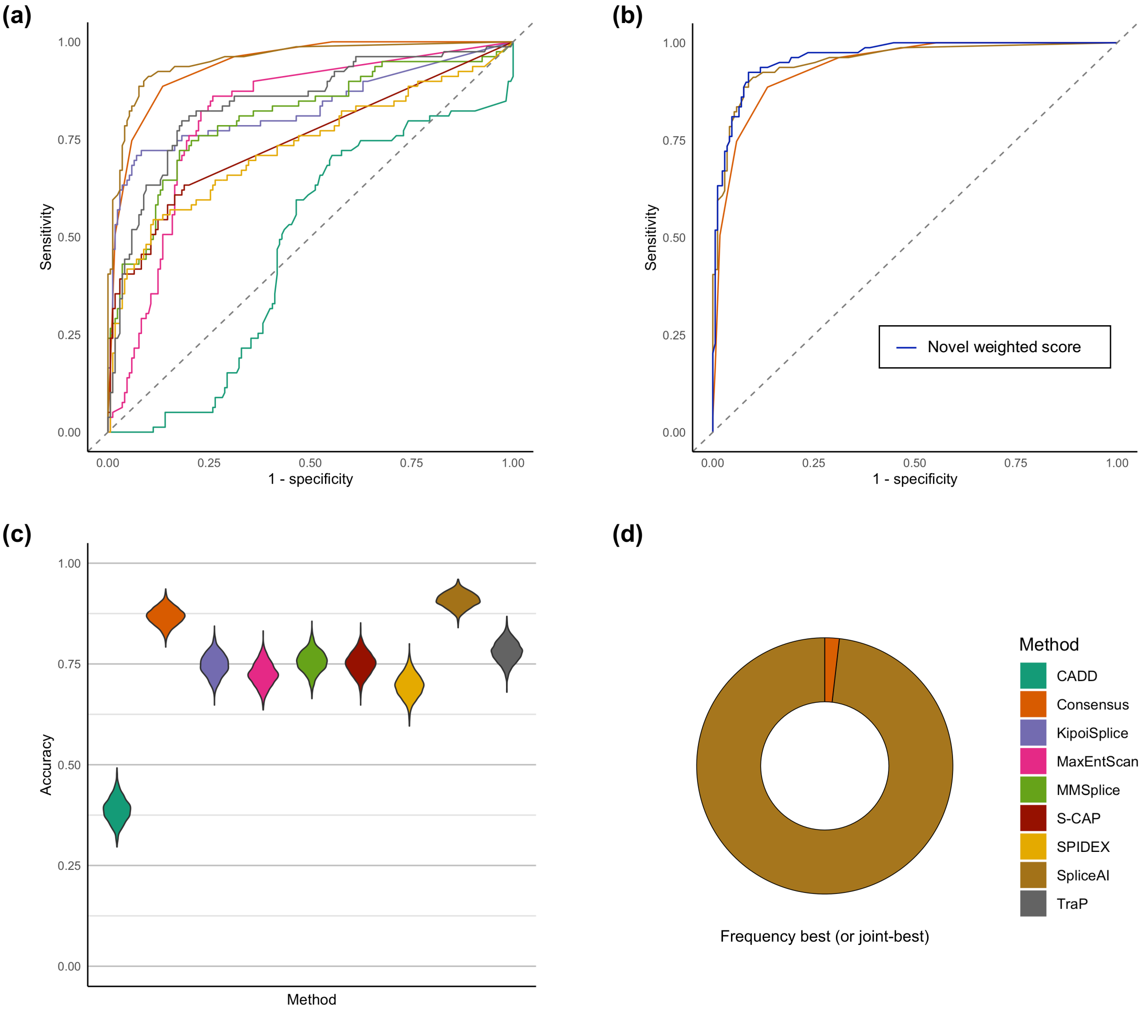


Figure 12. *Results from* in vitro *minigene assays demonstrating multiple consequences as a result of variants proximal to the canonical splice site.* (**Left**) gel electrophoresis snapshots of cDNA products amplified from primers designed for control exons within the minigene (*exon 1* & *exon 2*). All prominent bands were cut out and Sanger sequenced. *Right*, solid red blocks illustrate alignment of sequenced cDNA transcripts to features within the minigene vector: control exons (*grey boxes*) and inserted exons (*purple boxes*); **(a) SCN2A c.2919+3A>G**, showing complete exon exclusion and exon truncation in minigene vectors containing the c.2919+3A>G variant (*top* two alignments) and normal splicing in minigene vectors containing the WT sequence (*bottom* alignment). The first resulted in a transcript with a truncated exon, NM_001040142.1:r.2563_2710del, and the second resulted in a complete exon skip, NM_001040142.1:r.2563_2919del. It is noteworthy that if these events were also observed *in vivo* then they may be considered differently using ACMG criteria; the exon truncation event resulted in a frameshift and introduction of a premature stop codon (*PVS1*), whereas the complete exon skipping event resulted in the inframe removal of 119 amino acids from the transcript (*PM4*); **(b) MERTK c.2486+6T>A**, showing a shifting of the exon included in the reading frame in minigene vectors containing the c.2486+6T>A variant (*top* alignment) and normal splicing in minigene vectors containing the WT sequence (*bottom* alignment). This novel variant is present in two individuals with severe rod-cone dystrophy, and resulted in the simultaneous usage of a cryptic exonic splice acceptor site and a cryptic intronic splice donor site creating a novel exon (chr2:112,779,939-112,780,082, *GRCh37*), and a premature stop codon in the penultimate exon, p.(Trp784Valfs*10). Original images for both *SCN2A* c.2919+3A>G and *MERTK* c.2486+6T>A are presented in **Supplementary Figure S2**.

tissue-specific factors which are outside the scope of assays performed here (Aicher et al., 2020; Vig et al., 2020; Cummings et al., 2020), and variants may have pathogenic impacts on gene expression and/or regulation without any detrimental impact on splicing (Castel et al., 2018; Evans et al., 2018; Short et al., 2018; Zhang et al., 2020). Such factors will influence comparative metrics between algorithms, and future investigations may uncover pathogenic roles for variants reported here. However, the optimal thresholds calculated in light of these limitations for the 249 functionally assessed VUSs in this study are reported in **Supplementary Table S3**.

Global approaches to variant analysis, as assessed through the AUC, may fail to capture region-specific intricacies in splicing disruption (Jagadeesh et al., 2019). For example, variants could be sub-divided by their pathogenic mechanism, their effect on pre-mRNA splicing, their predicted molecular consequence or the location of the variant with respect to known splicing motifs, and each of these sub-groups may require different approaches or thresholds for accurate prioritization of pathogenic variation. We therefore predicted variants to be 'disruptive' or 'undisruptive' according to thresholds pre-defined by the developers of the tools. This included region-specific thresholds for S-CAP and CADD, across six and five different regions, respectively, dependent on variant location in relation to its nearest exon (**Supplementary Table S4**, **Figure 10b**). These regions illustrate if a variant lies in the core splicing dinucleotides, the immediate vicinity of these sites, or at a greater distance. We utilized a single score threshold for tools where region-specific thresholds have not been previously identified (**Supplementary Table S4**). We compared accuracy of each of the prioritization strategies across 2,000 iterations of sampling with replacement. This analysis highlighted differences across the tools and significantly differentiated their ability to accurately predict pathogenicity (Kruskal-Wallis, df = 8, p < 0.0001; **Figure 13c-d**). Similar to the AUC analysis, SpliceAI (using a threshold of 0.2) was

Figure 13. *Comparison of* in silico *strategies to prioritize 249 variants of uncertain significance with functional investigations performed.* (**a**) Receiver operating characteristic area under the curve (AUC) comparisons for nine *in silico* prioritization strategies demonstrating that SpliceAI (AUC=0.95, 95%CI=0.92-0.97) and a consensus approach (AUC = 0.94, 95% CI = 0.91-0.97) outperform other strategies for prioritization; (**b**) AUC comparisons between SpliceAI, a consensus approach and a novel metric, demonstrates that a weighted approach slightly increases accuracy of prioritization over single approaches alone (AUC = 0.96, 95% CI = 0.94-0.98); (**c-d**) Accuracy comparisons of each *in silico* prioritization approach across 2000 bootstraps utilizing region-specific pre-defined thresholds: (**c**) Violin plot demonstrating the calculated accuracy of each *in silico* prioritization approach; (**d**) Frequency that each strategy is the best or joint-best performing.



significantly the best performing strategy across all assessed single algorithms for our set of analyzed VUSs (Kruskal-Wallis, p < 0.0001 for all pairwise comparisons of accuracy between SpliceAI and other tools; **Figure 13c-d**).

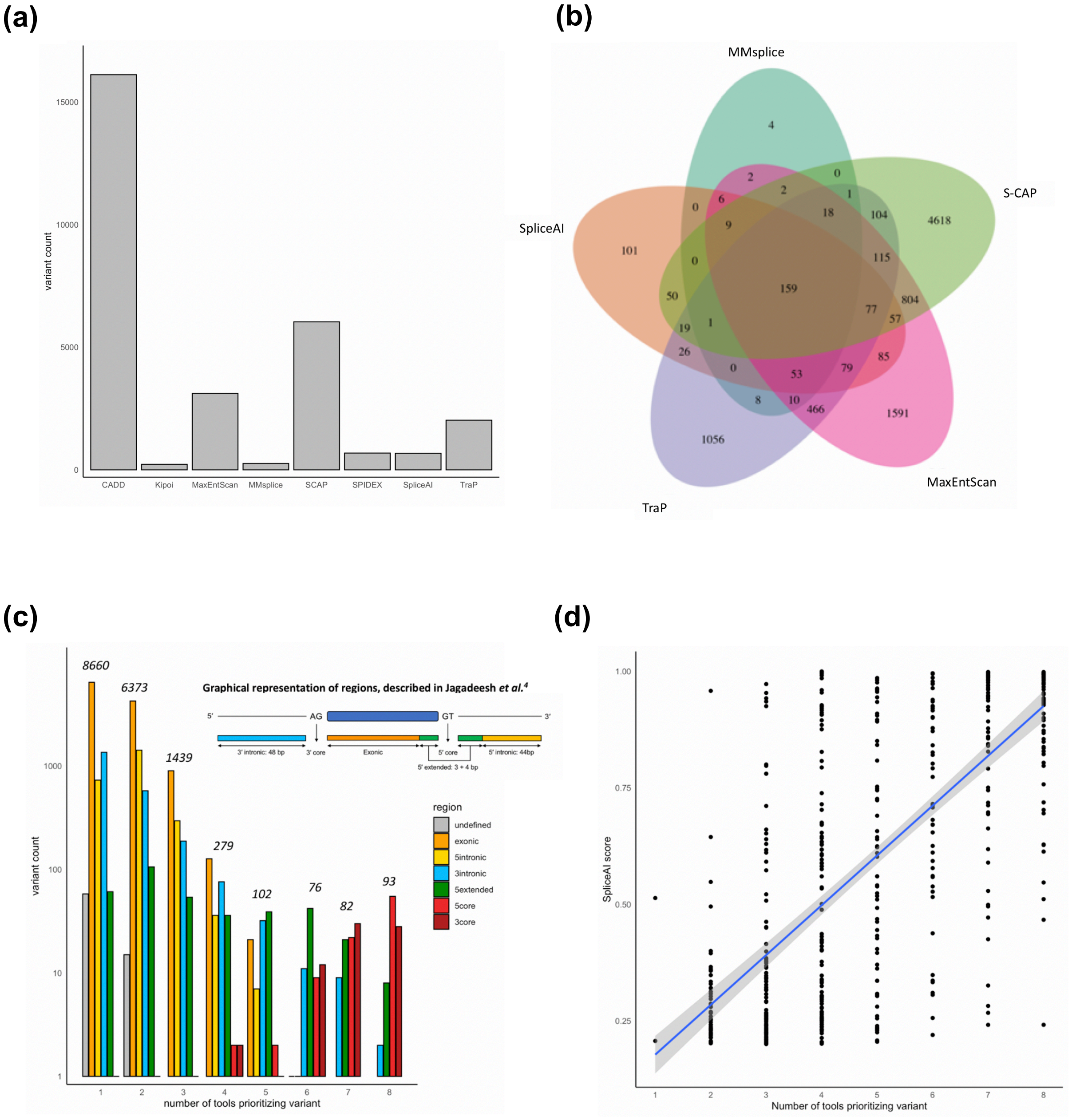
2.4.3. Combining in silico tools improves accuracy to identify variants of uncertain significance impacting splicing

To determine if combining one or more of these metrics could achieve greater accuracy than prioritization scores in isolation, we developed a consensus score for each variant which considered the region-specific thresholds for each tool (**Supplementary Table S4**). The score ranged from 0-8 and represented the number of tools for which a variant's score exceeded the respective threshold. We observed that the consensus approach performed similarly to SpliceAI when assessed through the receiver operating characteristic AUC (**Figure 13a**; **Supplementary Tables S2 & S3**). The consensus approach (using a threshold of 4/8 algorithms supporting splicing disruption) also performed more similarly to SpliceAI than other strategies when measuring accuracy across sampling iterations (**Figure 13c**), but was less frequently the best performing approach (**Figure 13d**). Variability in model accuracy was consistently low across sampling iterations for all tools (**Supplementary Table S5**). To understand if the relative scores from each algorithm could assist interpretation we developed a novel metric which incorporates weighted scores from the prioritization strategies. This analysis considered the actual score of the variant relative to the maximum score possible from each prediction algorithm (see *2.3.*). Of note, the weighted approach considering scores from SpliceAI and a consensus approach performed better than these two approaches in isolation (**Figure 9b**; **Supplementary Table S3**). Although not mutually exclusive and underpowered to detect significant statistical differences in the AUC from this combined analysis – due to marginal gains in accuracy and sample size – this demonstrates the potential utility of combined approaches utilizing combinations of scores to improve accuracy for the identification of variants impacting splicing.

2.4.4. Integration of in silico strategies to prioritize variants impacting splicing for a large cohort of individuals with rare disease

Next, we sought to examine the impact of these approaches on clinical variant analysis. Therefore, we integrated region-specific prioritization strategies (**Supplementary Table S4**) for intronic gene panel variants identified during routine diagnostic analyses for 2,783 individuals with rare diseases (Ellingford et al., 2016a). All individuals included in this analysis have received genetic testing for rare disease within the UK National Healthcare Service through a clinically accredited laboratory. We calculated *in silico* scores for 20,617 variants (of which 18,013 were rare) observed a total of 1,346,744 times in the cohort. We observed substantial variability in the number of rare variants prioritized by each *in silico* tool (**Figure 14a**; **Supplementary Table S6**) and in the specific variants prioritized by the most correlated *in silico* splicing tools (**Figure 14b**). We observed that while variants which show the highest consensus between *in silico* splicing tools impact the canonical splice site (**Figure 14c**; **Supplementary Table S7**), 99% (*n* = 17,871) of variants analyzed impact exonic or intronic regions of genes outside of the canonical splice sites.

Figure 14. *Summary of the overlap and correlations observed between the scores from in silico splicing prediction algorithms for 18,013 unique rare variants identified in a large cohort of 2783 individuals with rare disease.* (**a**) Bar chart showing overall count of unique variants prioritized using pre-defined thresholds for each *in silico* prediction algorithm. (**b**) Overlap between the unique variants prioritized by the five most correlated *in silico* prediction tools. (**c**) Grouped bar chart demonstrating the overlap of variants prioritized by each tool segregated by the region of the genome that the variant impacts, as defined by Jagadeesh *et al.* (2019), demonstrating that variants prioritized by many tools are highly likely to be close the canonical splice sites (5'core, 3'core and 5'extended). (**d**) Correlation between SpliceAI score and the number of additional tools also prioritizing the variant for the 528 unique rare variants prioritized by SpliceAI.



Splicing variants are often considered as a single class of variants and canonical splice site variants are therefore highly susceptible to over-prioritization by *in silico* tools, as such variants represent the majority (~70%) of known pathogenic splicing variants (Stenson et al., 2014, Xiong et al., 2015, Krawczak et al., 2007). Our data further underline the need to develop effective and unbiased strategies for prioritizing variants impacting splicing outside of the canonical splice sites, and this will be especially important for VUSs in known disease genes. Overall, these data demonstrate that different *in silico* strategies for splicing variant prioritization will alter the burden of variant analysis for clinical scientists. This is an important consideration for the analytical specificity and throughput of diagnostic testing.

To assess the clinical impact of such strategies, we integrated a single prioritization strategy, SpliceAI (using a threshold of 0.2, as above), in parallel to outcomes from routine diagnostic testing. This analysis involved extensive curation of genomic findings for the 2,783 referred individuals, all of which were classified in accordance with ACMG guidelines by clinically accredited scientists. We added SpliceAI predictions alongside these analyses and observed that this approach influenced analysis for 420 (15%) individuals receiving genomic testing for rare disease, and prioritized variants that could result in new or refined molecular diagnoses in 81 (3%) cases. Overall, we prioritized 758 variants (528 unique variants) in 646 individuals (23% of cohort) with a range of predicted molecular consequences. Most (99.6%, 526/528) variants were prioritized by at least one other *in silico* tool (**Supplementary Table S8**). The strength of the score from SpliceAI correlated highly with prioritization from other *in silico* tools (**Figure 14d**) and differed between regions of genome that were impacted (**Supplementary Table S9**). We classified prioritized variants as being:

- *New*: variant not previously highlighted or reported through diagnostic testing
- *Clarified*: variant previously reported through diagnostic testing but pathogenicity or pathogenic mechanism was unclear
- *Reported*: variant already described or established as ‘pathogenic’ or ‘likely pathogenic’ through diagnostic testing

In this regard, we identified 379 *new* variants in 337 individuals, 87 *clarified* variants in 83 individuals and 292 *reported* variants in 274 individuals. We found most (91%, 697/758) variants to be in genes known as a recessive cause of genetic disease. To understand if these variants impacted normal splicing, we interrogated the GTEx datasets (GTEx Consortium, 2013) for individuals carrying these variants in a heterozygous state, identifying 40 carriers of variants prioritized by this analysis. Of these, 21 had suitable RNA-seq datasets available for evaluation, and we were able to clearly observe significant alterations to splicing in four cases (**Table 4**). Whilst most variants will require bespoke functional investigations to establish precise effects on splicing and protein synthesis, leveraging publicly available datasets for individuals carrying potentially pathogenic rare variants in the GTEx dataset can quickly increase certainty of variant impact and refine clinical variant analysis.

Variant	Gene	Tissue	Metric Type	Controls Mean (95% CI)	Cases
20-3899342-G-A	*PANK2*	Fibroblasts	Intron Retention	0.12 (0.10-0.14)	0.32
12-88448136-G-A	*CEP290*	Thyroid	NRC	0.15 (0.02-0.27)	0.91
10-73567463-C-T	*CDH23*	Ovary	NRC	0.003 (0-0.01)	0.11
2-110922263-G-A	*NPHP1*	Testis	NRC*	0.51 (0.48-0.55)	0.7

Table 4. *Metrics obtained from the analysis of GTEx v7 datasets to observe the impact of variants prioritized as splice-impacting.* Our analysis identified 4 variants in autosomal recessive genes that were present in a carrier state in individuals in GTEx v7 and had observable impacts on splicing in these individuals. Metrics were calculated from aligned RNA-seq datasets from tissues with a transcript per million value > 5 for the gene of interest. *Cases,* individuals within the GTEx dataset carrying prioritized variant. *Controls*, a group of 10 randomly selected individuals within the GTEx dataset that do not carry the prioritized variant. NRC, normalized read count (see **Box 1**); * indicates a shift in the usage of two canonical exon junctions, corresponding to different transcript isoforms.

2.5. Discussion

The incorporation of the prioritization and functional strategies described in this study for variants impacting splicing significantly improved molecular diagnostic services. However, we expect that the true impact of such analysis strategies will be more profound. Targeted next generation sequencing approaches employed within this large cohort ignore deeply intronic regions of genes, which, as shown here and in other studies (den Hollander et al., 2006, Montalban et al., 2019, Sangermano et al., 2019), can harbor variants which result in aberrant splicing through the production of novel cryptic exons. The recent availability of genomic datasets within healthcare amplifies the current limitations in interpreting variation within the non-coding genome, particularly in large genome sequencing cohorts. Our findings demonstrate the opportunity to expand bioinformatics analysis to the pre-mRNA regions of known disease genes and provide immediate increases to diagnostic yield. Further, a wide variety of bioinformatics prediction tools continue to be developed, as seen with the recent release of CADD-Splice, (Rentzsch et al., 2021) and SQUIRLS (Danis et al., 2021). As such tools continue to become available, careful analysis of their utility using a framework as described here will allow integration with maximum effect. Future approaches may expand on the consensus model described here through integration of probabilistic models, for example based on Bayesian statistics. Importantly, we demonstrate a requirement to functionally assess variant impact on pre-mRNA splicing as the delineation of the precise effects may be important in considerations for variant pathogenicity. The prioritization and identification of pathogenic variants impacting splicing is therefore an important consideration for diagnostic services and for the development of new targeted treatments.

Chapter 3

Identification of a pathogenic BBS1 branchpoint variant through bespoke bioinformatics analysis

3. Identification of a pathogenic BBS1 branchpoint variant through bespoke bioinformatics analysis

3.1. Abstract

Recent years have seen the development of a wide variety of computational strategies for the prediction of SNV impact on splicing. Previous work has shown SpliceAI to most often be the best-performing individual strategy for splice variant prioritisation. However, it remains to be seen whether the efficacy of SpliceAI holds across different variant types and locations. One such subtype are variants impacting the intronic branchpoint (BP), which are ostensibly rare, numbering only the tens in the existing scientific literature. We thus sought to evaluate the ability of SpliceAI to prioritise BP variants, and develop a novel approach for their prioritisation from genomic datasets.

Collating 30 known pathogenic branchpoints from the literature, we demonstrate that SpliceAI (with a threshold of 0.2) is correctly able to prioritise only 50% of branchpoint variants. We thus developed a meta-analytical tool combining the output of two existing branchpoint predictor tools and an empirically derived branchpoint dataset.

Applying this approach to a cohort of intronic variants in 2021 patients referred for retinal dystrophy gene panel testing, we show that putatively BP-impacting variants are present in 5.9% of surveyed patients. While many of these are unlikely to be pathogenic, we are able to identify a BP-impacting variant in intron 7 of the *BBS1* gene, which is shown through downstream functional studies to be both splice-impacting and causative, in a compound heterozygous state, of the patient phenotype of retinitis pigmentosa.

We thereby demonstrate the diagnostic value of evaluating the efficacy of predictive tools across different variant subtypes and the development of bespoke bioinformatics approaches where this efficacy is sub-optimal.

3.2. Introduction

Variation at any of the essential *cis*-acting splicing sequence elements is liable to cause pathogenic mis-splicing of the respective transcript. Two such elements are the branchpoint sequence (BPS) and poly-pyrimidine tract (PPT), which demarcate the 3’ end of all human introns, serving as binding sites for spliceosomal proteins. Specifically, the BPS and PPT are believed to primarily recruit the U2 snRNP auxiliary proteins U2AF65 and U2AF35, respectively (Zorio and Blumenthal, 1999, Merendino et al., 1999, Wu et al., 1999). The assembled U2 complex then interacts with local and distal splicing complexes, including the 5’SS-recognising U1 snRNP, to initiate the splicing reaction (Shao et al., 2012).

The spatial relationship between the BP and other sequence features facilitates this process of binding and interaction. Between the BP and 3’SS generally lies an *AG-exclusion zone* (*AGEZ*), in which no AG dinucleotides are present. Cryptic AGs introduced by variants into the AGEZ are liable to act as competing splice sites, as they lie downstream of required *cis*-acting sequence features, i.e. the BP and PPT themselves (Smith et al., 1993, Wimmer et al., 2020). The distance between the BP and 3’SS also has a significant on splicing patterns, with greater BP-3’SS distances being associated with alternatively spliced exons (Corvelo et al., 2010).

The adenosine residue in the canonical branchpoint TNA motif is itself an active agent of the splicing reaction, serving as a nucleophile that attacks the 3’ splice acceptor in the first of the two steps of the central transesterification reaction. The splicing reaction results in the formation of a *lariat*, a lasso-like structure consisting of the intron, with the branchpoint adenosine acting as the “junction” of the lariat, being bonded to its neighbouring nucleotides, as well as the 3’ splice site that it has attacked. Identifying this distinctive point in the lariat can thus allow direct confirmation of the position and identity of the branchpoint residue.

This experimental identification of the BPS is hindered by the rapid degradation of lariat sequences after completion of splicing. Empirical approaches to identify BP sequences have therefore relied on the mining of large RNA-seq datasets (Mercer et al., 2015, Taggart et al., 2012, Taggart et al., 2017) in an effort to detect low-level lariat sequences. Sequencing reads mapping to intronic lariats are easily discernible, as traversing the BP-5’SS junction results in a split and inverted read (Taggart et al., 2012, Mercer et al., 2015). By designing bespoke approaches that identify reads that map to known introns but with this, putative branchpoints can be empirically identified *en masse*. The traversal of the often also resulting in the incorporation of an incorrect nucleotide in place of the BP adenosine, providing another indicator of branchpoint identity (Taggart et al., 2017).

The difficulty in direct observation of functional branchpoints has led to the development of bioinformatics tools based on theoretical models. Two such tools are SVM-BPfinder (Corvelo et al., 2010) and Branchpoint Predictor, or BPP (Zhang et al., 2017). Both models take as input an intronic sequence, and scan the sequence in windows of 9 or 7 bp (for SVM-BPfinder and BPP, respectively).

The similarity to the canonical BPS is evaluated using either an SVM (for SVM-BPfinder) or mixture model (for BPP). Both tools also incorporate consideration of the length, composition and relative position of the PPT in evaluating the potential strength of a BPS. In the case of SVM-BPfinder, predictions are returned for every canonical TNA motif present in the supplied sequence, and can later be filtered to identify the "optimal" branchpoint, while, for BPP, every position in the sequence is evaluated, and only the position most likely to be the branchpoint is returned for a given intron.

Pathogenic branchpoint variants are ostensibly rare in current literature, currently numbering in the order of only tens of reported variants (see **Table 5**, below). Whether such variants are truly rare, or whether they are simply under-analysed, remains to be seen. As with many intronic variants, the ability to reliably call BP variants is dependent on the sequencing approach used: while the majority of BPs should be well-covered in WGS data, they are not within the targeted enrichment region for gene panels or WES. However, while the intended coverage of panel and exome sequencing is primarily exonic, both approaches may in fact show some degree of intronic coverage, usually lower than that of the target region. This is a result of the enrichment process, which encompasses small flanking regions to ensure adequate coverage of the target regions.

Leveraging this intronic coverage may allow the investigation of variation at the BPS and PPT through re-analysis of existing data. In this study, we developed a bespoke bioinformatics approach that aimed to evaluate the frequency of SNVs at putative branch points for a cohort of patients. We further sought to identify cases where variants at the BPS may be responsible – either wholly or partially – for a patient's phenotype.

3.3. Materials and methods

Curation of existing branchpoint variants

Pathogenic variants shown to impact the function of the human BPS were curated from existing literature. We retained only SNVs for compatibility with empirical datasets and our selected predictive tools. SpliceAI scores (Jaganathan et al., 2019) were retrieved for all identified variants.

Retinal dystrophy gene transcript selection

For each of the genes present in the Manchester Centre for Genomic Medicine retinal dystrophy gene panel, we selected the transcript(s) assigned to that gene in-house for downstream analysis and extracted their intronic sequences via the Ensembl API (Howe et al., 2021). Selected transcripts are listed in **Supplementary Table 10**.

Bioinformatics analysis

The co-ordinates of experimentally derived BPSs were accessed via primary publication (Mercer et al., 2015). Source code for SVM-BPfinder and BPP was obtained through online repositories (https://bitbucket.org/regulatorygenomicsupf/svm-bpfinder/src/master/ and https://github.com/zhqingit/BPP, respectively). No customisable parameters were available for BPP; for

SVM-BPfinder, default parameters were used except for the minimum 3'SS-BP parameter, which was decreased to 10 bp to allow identification of any BPs located an abnormally short distance from the splice acceptor. Intronic sequences were retrieved for all retinal panel gene introns using the Ensembl API and bespoke scripts used to cross-reference intronic variants from retinal dystrophy panel patients with the output of SVM-BPfinder, BPP and the list of empirical branchpoints.

Variant filtering strategy to identify putatively BP-impacting SNVs

We filtered putatively BP-impacting variants for rarity by excluding those with an AF frequency > 1% in the gnomAD database (Karczewski et al., 2020). We further excluded any variants present in 10 in-house individuals or more. Additionally, we filtered out heterozygous variants for which there was substantial imbalance in variant calling read support - namely, where one allele was supported by 80% or more of the total reads at that base position. Such variants were hypothesised to be artefactual findings generated through erroneous variant calling.

Midigene-based investigation of putative BP variant

Midigene assays were conducted externally, as described in (Fadaie et al., 2021).

3.4. Results

3.4.1. SpliceAI is underpowered to detect branchpoint variants

To evaluate the accuracy of SpliceAI in predicting BP-related splice disruption, we curated a list of 20 high-confidence pathogenic branchpoint variants from existing literature (**Table 5**). Of these, 55% (11/20) affected the canonical branchpoint adenosine, while 45% (9/20) affected the highly conserved thymidine two bases upstream. The remaining variant affected a thymidine five bases upstream of the canonical adenosine. Cross-referencing with predictive scores revealed that, for this cohort of variants, SpliceAI (with a threshold score of significance > 0.2, as described above) had a sensitivity of 50%. Recent studies have consistently demonstrated that, at similar score thresholds, SpliceAI has an overall sensitivity of between 78-91% (Jaganathan et al., 2019, Wai et al., 2020, Riepe et al., 2021). Despite the small sample size, this preliminary analysis suggests that SpliceAI may be underpowered to detect the splicing impact of some BP variants. We thus aimed to develop a bespoke approach to highlight genomic variants that may overlap branchpoint sites.

3.4.2. Analysis of branchpoint frequency and distribution

We aimed to re-analyse the intronic variants identified in a cohort of 2021 patients who had previously undergone clinical diagnostic sequencing against a retinal dystrophy gene panel (see *3.3.*). We generated a map of both predicted and empirically identified branchpoints for all introns in 188 transcripts annotated to the 176 genes associated with the retinal dystrophy panel, encompassing a total of 2560 unique introns. Introns ranged in length from 30-238,135 bp, with a median length of 1766 bp (**Figure 15a**).

Empirically identified BPs were sourced from a previously described dataset (Mercer et al., 2015). No validated BP from this dataset was identified for 86.8% (2221/2560) of retinal dystrophy panel gene introns (**Figure 15b**); this likely represents the difficulty in experimentally capturing lariats, rather than a true lack of BP in those introns. Of the 13.2% (339/2560) of introns in which one or more validated branchpoints were identified, the majority (71.7%, 243/339) contained just a single branchpoint, with two and three branchpoints being identified in 18.9% (64/339) and 6.8% (23/339) of introns, respectively. Nine experimentally validated branchpoints overlapped the co-ordinates of intron 30 of the *SNRNP200* gene, the greatest such number among the retinal dystrophy gene introns analysed (**Figure 15c**). Only one of the nine branchpoint residues was an adenosine, and was part of a CNA motif.

The predictive component of our analysis combined the use of both SVM-BPfinder and BPP to predict the optimum branchpoint sequence within the last 500 bp of the intron. Between both tools, a total of 3516 unique residues were predicted to be the branchpoint of at least one of the retinal dystrophy gene introns. For 60.1% (1539/2560) of surveyed introns, both tools returned the same predicted branchpoint, suggesting a relatively high concordance between their predictions (**Figure 15d**). We also compared the distance between the BP and 3'SS identified by each tool (**Figure 15e**). Both tools showed a median BP-3'SS distance of 27 bp across the surveyed introns; SVM-BPfinder, however, was more likely to predict residues very proximal and distal to the 3'SS as being BPs, with 17.7% of BPs predicted to be < 20 nt from the 3'SS, and 20.4% > 50 bp from the 3'SS (compared to 11.8% and 5.4%, respectively, for BPP). This suggests BPP prioritises a narrower range of genomic space when predicting BPs than does SVM-BPfinder.

SVM-BPfinder was unable to identify an optimal branchpoint for 1.5% (39/2560) of introns. 31 of these had at least one TNA motif in the final 500 bp of the intron, but with none predicted by the filtering algorithm to be sufficiently similar to the consensus BPS. The remaining 8 were all short introns of 75-242 bp lacking any TNA motifs; 7 of these were predicted by BPP to have branchpoints with the recognised non-canonical CNA motif, while the remaining branchpoint was predicted to consist of a rare TGC motif. Of the 2543 total unique branchpoints identified by BPP, 9.2% (233/2543) were predicted to consist of non-TNA motifs.

Using SVM-BPfinder, we also investigated the number of predicted branchpoints present in each intron (**Figure 15f**). Across all investigated genes, the number of unfiltered TNA motifs observed in the last 500 bp of each intron ranged from 1-73. Counting only those with a positive SVM output score (as stipulated by SVM-BPfinder when selecting the optimum branchpoint in an intronic sequence) reduced this to 0-22. When solely considering the final 100 bp of each intron, the unfiltered TNA motif counts ranged from 0-20, dropping to 0-9 when filtering for only positively-scoring motifs. The median number of positively-scoring TNA motifs in the final 100 bp of retinal dystrophy gene introns was just 2. This suggests that, while a large number of sites within introns may theoretically serve as branchpoints, the number of those resembling genuine BPSs is much lower.

Table 5. *Characteristics of 20 known pathogenic BP-impacting SNVs.*

Gene	HGVSc	HGVSg	Affected BP site	Phenotype	Citation	Max SpliceAI score
COL5A1	NM_000093.4:c.2701-25T>G	chr9:g.137686903T>G	-2 T	Ehlers-Danlos syndrome type II	Burrows et al., 1998	0.1634
DYSF	NM_003494.3:c.3443-33A>G	chr2:g.71817308A>G	A	Mild limb-girdle muscular dystrophy (LGMD)	Sinnreich et al., 2006	0.4426
F9	NM_000133.3:c.253-25A>G	chrX:g.138619496A>G	A	Haemophilia B	David et al., 1998; Ketterling et al., 1999	0.447
FBN2	NM_001999.3:c.3974-26T>G	chr5:g.127670562A>C	-2 T	Congenital contractural arachnodactyly (CCA)	Maslen et al., 1997	0.464
IKBKG/ NEMO	NM_001099857.1:c.519-23A>T	chrX:g.153788599A>T	A	Anhidrotic ectodermal dysplasia with immunodeficiency (EDA-ID)	Jørgensen et al., 2016	-
ITGB4	NM_000213.3:c.1762-25T>A	chr17:g.73732344T>A	-2 T	Epidermolysis bullosa with pyloric atresia (PA-JEB)	Masunaga et al., 2015	-
ITGB4	NM_000213.3:c.3977-19T>A	chr17:g.73748508T>A	-2 T	Epidermolysis bullosa with pyloric atresia (PA-JEB)	Chavanas et al., 1999	-
KCNH2	NM_000238.3:c.2399-28A>G	chr7:g.150646165T>C	A	Long QT syndrome	Crotti et al., 2009	-
L1CAM	NM_000425.3:c.2432-19A>C	chrX:g.153131293T>G	A	X-linked hydrocephalus	Rosenthal et al., 1992	0.1446
LCAT	NM_000229.1:c.524-22T>C	chr16:g.67976512A>G	-2 T	Fish-eye disease	Kuivenhoven et al., 1996	-
NPC1	NM_000271.4:c.882-28A>G	chr18:g.21137182T>C	A	Niemann-Pick disease type C (NPC)	Di Leo et al., 2004	-
NTRK1	NM_002529.3:c.851-33T>A	chr1:g.146853392T>A	-2 T	Congenital insensitivity to pain with anhidrosis (CIPA)	Miura et al., 2000	-
PC	NM_000920.3:c.1369-29A>G	chr11:g.66620883T>C	A	Type B pyruvate carboxylase deficiency	Ostergaard et al., 2013	0.4153
TH	NM_199292.2:c.1198-24T>A	chr11:g.2187017A>T	-2 T	Extrapyramidal movement disorder	Janssen et al., 2000	0.5052
TSC2	NM_001114382.1:c.5000-18A>G	chr16:g.2138031A>G	A	Tuberous sclerosis	Mayer et al., 2000	0.3819
UROS	NM_000375.2:c.661-31T>G	chr10:g.127477605A>C	-2 T	Congenital erythropoietic porphyria	Bishop et al., 2010	0.3114
USH2A	NM_206933.2:c.8682-17A>G	chr1:g.216040529T>C	A	Usher syndrome	Le Guédard-Méreuze et al., 2010	0.4978
VWF	NM_000552.3:c.6599-20A>T	chr12:g.6101204T>A	A	Type 1 von Willebrand disease (VWD1)	Identification: James et al., 2007; Functional testing: Hawke et al., 2016	-
XPC	NM_004628.4:c.413-9T>A	chr3:g.14209889A>T	-5 T	Xeroderma pigmentosum	Khan et al., 2004	0.974
XPC	NM_004628.4:c.413-24A>G	chr3:g.14209904T>C	A	Xeroderma pigmentosum	Khan et al., 2004	0.2742

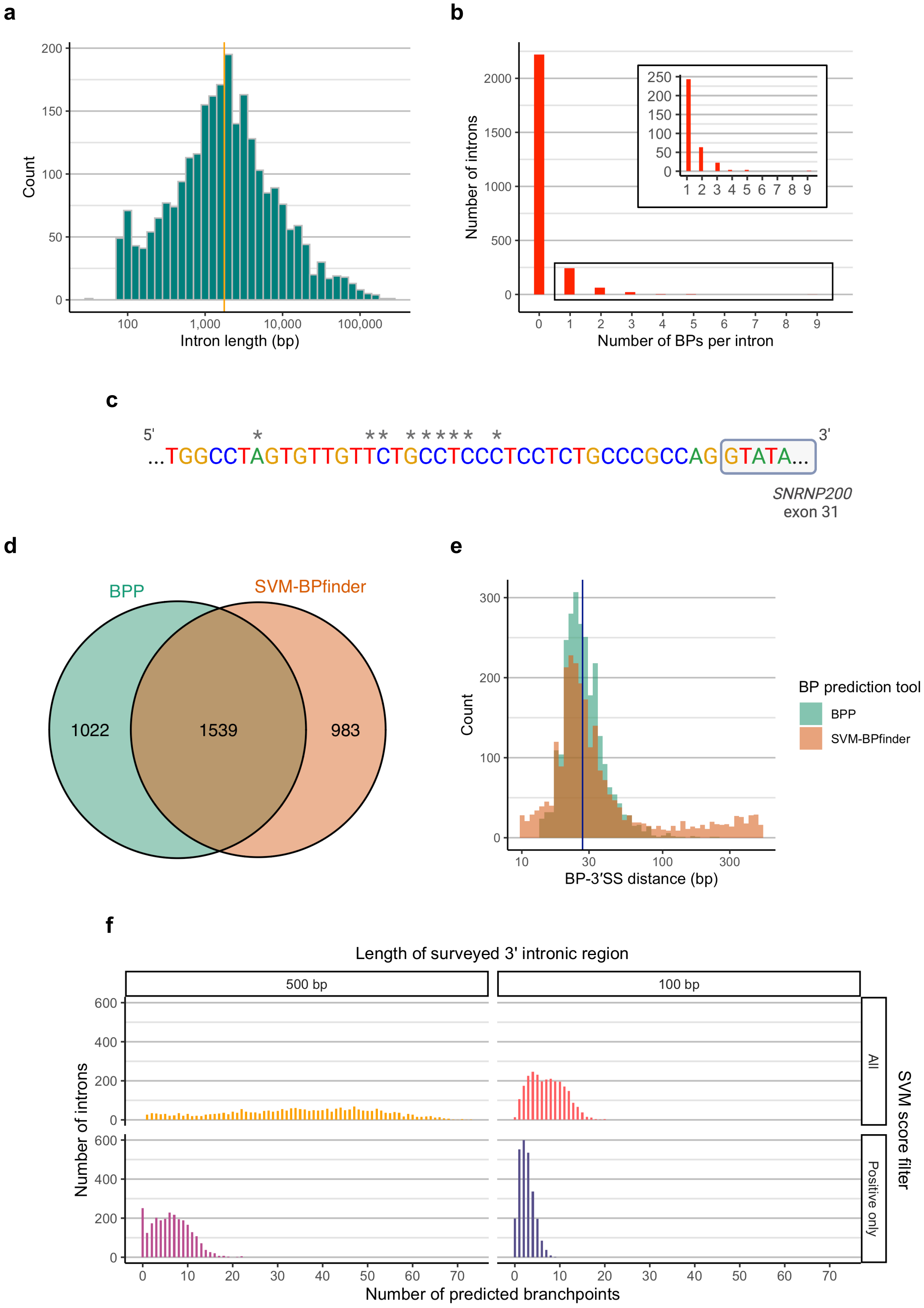

a
Count
200
150
100
50
0
100
1,000
10,000
100,000
Intron length (bp)
b
Number of introns
2000
1500
1000
500
0
250
200
150
100
50
0
1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
Number of BPs per intron
c
5'
3'
* ** ***** *
...TGGCCTAGTGTTGTTCTGCCTCCCTCCTCTGCCCGCCAG GTATA...
SNRNP200
exon 31
d
BPP
SVM-BPfinder
1022
1539
983
e
Count
300
200
100
0
10
30
100
300
BP-3′SS distance (bp)
BP prediction tool
BPP
SVM-BPfinder
f
Length of surveyed 3' intronic region
500 bp
100 bp
Number of introns
600
400
200
0
All
Positive only
SVM score filter
0 10 20 30 40 50 60 70
Number of predicted branchpoints

Figure 15. *Branchpoint distribution and frequency in retinal dystrophy gene introns.* We evaluated the branchpoint characteristics of 176 retinal dystrophy (RD) genes featured on the MCGM gene panel (**a**) Introns varied widely in length from 30-238,135 bp (median = 1776 bp, orange line). (**b**) Most retinal dystrophy panel gene introns do not contain an experimentally validated branchpoint; 9.5% of introns contained a single validated BPS, with higher numbers of BPs identified in a much lower number of introns. (**c**) Intron 30 of the *SNRNP200* gene contains nine experimentally validated BPs, the most of any surveyed intron. Only one has a canonical adenosine as the branchpoint residue. (**d**) The two bioinformatics tools used to predict BP residues showed fairly high concordance, with analysis of 60.1% of introns returning the same predicted branchpoint using both tools. (**e**) Across all surveyed introns, the median distance between the BPS and 3'SS was the same regardless of bioinformatics tool used (median = 27, blue line). SVM-BPfinder, however, showed a stronger tendency to predict more proximal and distal residues as being BPs. (**f**) Using SVM-BPfinder, which evaluates the resemblance of each TNA motif in a sequence to the canonical BPS, we observed that the final 500 and 100 bp of surveyed introns had between 0-73 and 0-20 TNA motifs, respectively. Introns had a median of two TNA motifs with positive SVM scores in the final 100 bp of the intron (bottom-right), which is used by SVM-BPfinder as an indicator of stronger BPS similarity.

3.4.3. Investigation of branchpoint variation in retinal dystrophy gene panel patients

We next cross-referenced the list of predicted branchpoints, derived from both the bioinformatics tools and experimentally-derived dataset, against a set of 6731 intronic variants called during sequencing of 2021 patients against the retinal dystrophy gene panel described above. Considering cases where variants were predicted to overlap either the BP adenine or the constrained residue two bases upstream, we observed that 1.5% (101/6731) of the variants overlapped with putative branchpoint residues from one of the three datasets, of which 78.2% (79/101) were rare (gnomAD AF < 1% and present in fewer than 10 in-house samples). Although more rigorous filters may ordinarily be applied to filtered for rare variants in dominant-acting genes, we opted to enforce the same in-house count filter for all variants, in case of the same pathogenic variant causing disease in multiple in-house patients. We additionally excluded five variants that displayed substantial allelic imbalance (allelic read count ratio of 80:20 or greater, see *3.3.*), which were speculated to result from artefactual errors in variant calling. The final variant set consisted of 74 putatively BP-impacting variants present across 50 genes in a total of 119 individuals; 56.7% (42/74) of these variants were predicted to affect the branchpoint residue, while the remainder (43.3%; 32/74) overlapped with the constrained residue two bases upstream. In seven individuals, two independent BP-overlapping variants were identified.

We again investigated the concordance of predictions between SVM-BPfinder and BPP, as well as the set of empirically validated BPs (**Figure 16a**). Although only 6.8% (5/74) of the highlighted variants were predicted by all three approaches to overlap a BPS, there was higher concordance when considering SVM-BPfinder and BPP alone: 72 variants were predicted to overlap a BPS by at least one of the two tools, of which 41.7% (30/72) were predicted by both. Of the 15 variants affecting residues identified as BPs in the empirical dataset, the majority (86.7%; 13/15) were also highlighted

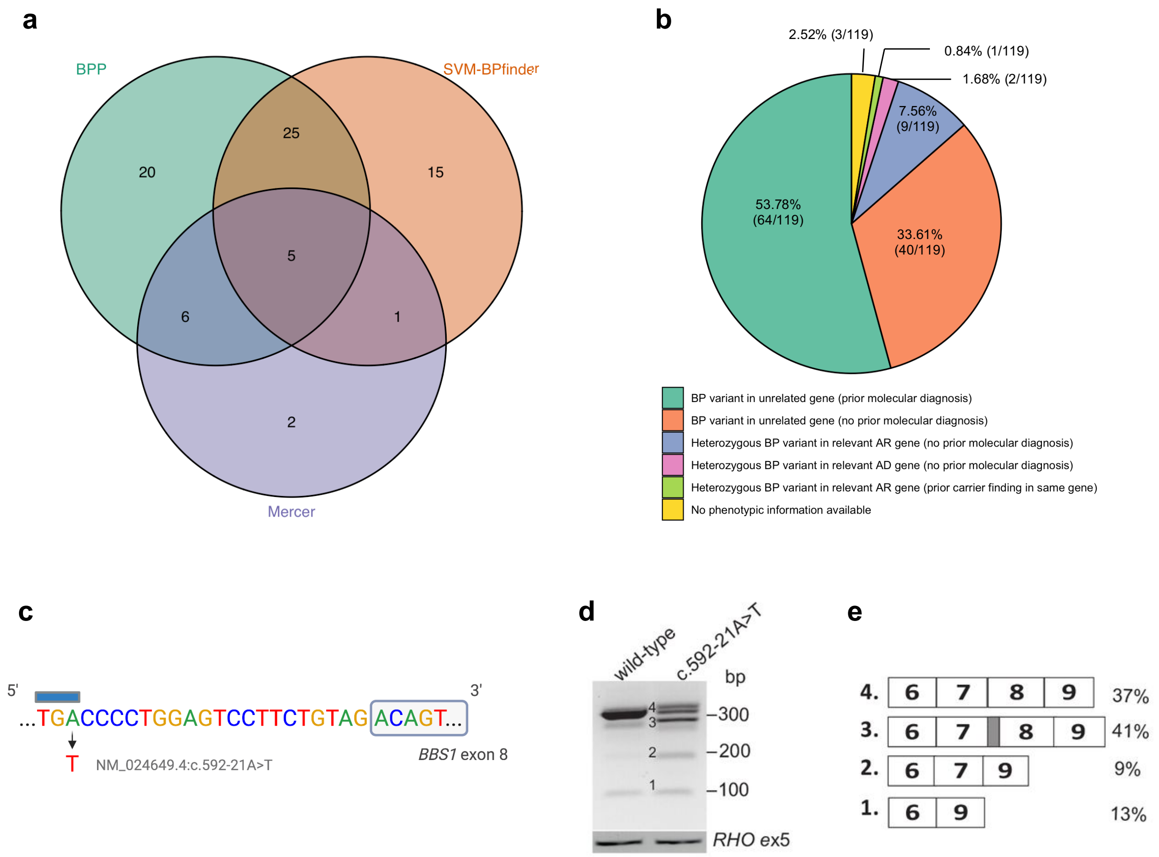


Figure 16. *Investigation of putative clinical BP-impacting variants.* We cross-referenced 6731 intronic variants called during retinal dystrophy (RD) gene panel testing of 2021 individuals against an empirically generated BP dataset (Mercer et al., 2015) and two computationally predicted lists of retinal dystrophy gene branchpoints. (**a**) Of the 74 high-quality SNVs predicted to overlap a BP residue, 50.0% were predicted by more than one tool. (**b**) The majority of patients harbouring putative BP variants had phenotypes not conventionally associated with loss of function of the affected gene. 13/118 predicted BP-impacting variants were in genes potentially relevant to patient phenotype. (**c**) A putative BP variant was identified in intron 7 of the *BBS1* gene, and was predicted to affect the canonical branchpoint adenosine residue 21 bases from the 3'ss. (**d**) Introduction of the variant into a midigene assay results in multiple aberrant bands when run on gel; (**e**) The majority of transcripts exhibited a 30-bp deletion of exon 8 (band 3 in (**d**)), with 22% of transcripts containing either a single- or multi-exon skip. (**d**) and (**e**) taken from Fadaie et al. (2021).

by at least one of the two bioinformatics approaches, suggesting that, for this subset of variants, the selected predictive tools effectively recapitulated empirical findings.

We re-analysed diagnostic reports for the 119 individuals harbouring putatively BPS-impacting variants to investigate the potential contribution of the variants to patient phenotype (**Figure 16b**, **Supplementary Table S11**). Phenotypic information was unavailable for 2.5% (3/119) of patients. 53.8% (64/119) of investigated patients had already received a molecular diagnosis, defined as having a pathogenic or likely pathogenic variant(s) in a disease-causing state identified through prior gene panel testing; the genes harbouring putative BP variants were not associated with disease phenotype in these patients. Of the 52 unsolved patients with phenotypic data available, the putative BP variants in 76.9% (40/52) were not in genes typically associated with the respective phenotypes or inheritance patterns. In two patients, heterozygous BP variants were identified in phenotype-relevant genes associated with autosomal dominant inheritance. These variants, however, were present in a heterozygous state in multiple other patients with unrelated phenotypes, and so were deemed unlikely to be pathogenic. In ten patients, putative BP variants were identified in phenotype-relevant genes associated with autosomal recessive inheritance, and in each patient constituted the first potentially pathogenic variant identified in the given gene.

In one patient, we identified a putatively BP-impacting variant in intron 7 of the *BBS1* gene (c.592-21A>T). In this patient, a known pathogenic variant, c.1169T>G (p.Met390Arg; Mykytyn et al., 2002), had previously been identified in a heterozygous state through upstream gene panel testing, and had been reported as a carrier finding. The putative BP variant was predicted to affect the canonical adenosine of the branchpoint intron (**Figure 16c**). One other TNA motif was identified within 100 bp of the 3’ end of intron 7, but was not predicted by either SVM-BPfinder or BPP to be the optimal branchpoint sequence in intron 7.

The patient had presented with non-syndromic retinitis pigmentosa (RP), which may be caused by homozygous loss-of-function variants in *BBS1* (Estrada-Cuzcano et al., 2012), and so the identified variant was hypothesised to contribute to the RP phenotype in this patient. Notably, this variant was predicted to have a SpliceAI score of 0.204, and so would have been missed if applying the SpliceAI threshold of 0.25 recommended by the creators of the tool, but would have been identified if using the threshold of 0.2 identified in our earlier analysis of splice variants (see *2.4.2.*).

Initial investigation of the impact of the c.592-21A>T variant using a minigene assay did not result in significant disruption to splicing. However, following variant sharing, further downstream functional analysis of the effect of this variant was conducted elsewhere using a larger midigene assay incorporating a 7.3 kb insert, encompassing six exons of *BBS1* (Fadaie et al., 2021). Wild-type splicing of the transcript was observed to be reduced by 63% in the presence of the c.592-21A>T variant. Three distinct mis-splicing events were identified (**Figure 16d-e**): 41% of all transcripts showed evidence of the use of an alternative acceptor in exon 8, leading to omission of 30 nucleotides from the 5’ end of exon 8. Skipping of exon 8 was observed in 9% of transcripts, while 13% of transcripts showed a multi-exon skip encompassing both exons 7 and 8. Due to the large scale of the disruption, this variant was deemed likely to cause disease when in compound heterozygosity with p.Met390Arg.

3.5. Discussion

Through our bespoke pipeline, we were able to identify a single pathogenic variant in *BBS1* that had escaped consideration during upstream diagnostic analysis. We also identified a further ten variants in autosomal recessive genes that may represent a first pathogenic variant in the respective patients; further work to corroborate the effect of these variants will be necessary to evaluate the true frequency of splice-impacting BP variants in patient populations and the accuracy of our diagnostic approach. Despite this uncertainty, these findings support the notion that the identification of pathogenic intronic variation is not well-served by current diagnostic pipelines. It should also be noted that, although genotype-phenotype correlations were not strong for most of the putative BP variants, they may still constitute reportable incidental findings if shown to significantly impact splicing.

Recent years have seen the release of several other BP prediction tools, including the machine learning-based tools Branchpointer (Signal et al., 2018), LaBranchoR (Paggi and Bejerano, 2018) and RNABPS (Nazari et al., 2018). A recent comparison of six available BP prediction tools – including those described here – demonstrated that Branchpointer showed the greatest accuracy in identifying genuine branchpoints from falsely simulated ones (Leman et al., 2020), and so its incorporation, with other tools, into the approach described here may highlight putatively BP-impacting variants missed when using SVM-BPfinder and BPP alone. However, the same study also demonstrated that variants predicted by BPP to affect the BP adenosine or -2 residue were more likely to impact splicing than when predicted to do so by other tools, and recommended primarily using BPP in clinical contexts.

A notable limitation of our study is the absence of consideration of indels. A small number of BP-impacting indels have been described previously in the literature (Agrawal et al., 2005, Aten et al., 2013, Bosch et al., 2005). However, it has long been known that deletions at the 3’ end of the intron are also liable to impact splicing through the shortening of the PPT, such that assembly of the spliceosome is sterically unfeasible (Frendewey and Keller, 1985, Reed, 1989, Ruskin and Green, 1985). Such variants are seemingly rare in Mendelian disease, though this may represent an under-analysis of intronic indels analogous to that of BP-impacting variants. An extension of our approach to identify significant changes in PPT length may further boost diagnostic yield.

One complexity in the investigation of mis-splicing is the presence of tissue-specific splicing differences that may not be apparent in the surveyed tissue or vector system. This is also true for BPs, which show extensive tissue specificity, with up to 75% of human introns identified by one study as exhibiting different branchpoint usage between tissues (Pineda and Bradley, 2018). These findings are supported by the multiple branchpoints identified in single introns in empirical datasets (Mercer et al., 2015, Taggart et al., 2012, Taggart et al., 2017), as in the case of *SNRNP200* intron 30 described above. Accordingly, it may be hypothesised that investigation of the effect of BP variants should, where possible, be carried out in a disease-relevant tissue sample. The ability of non-human cell-based assays to recapitulate biological findings observed in human tissues will be an important area of study to ensure that corroborative methodologies are accurately portraying biological impact.

Relatedly, the failure of a minigene assay to demonstrate the splicing impact of the c.591-21A>T variant illustrates the importance of intelligent selection of functional assay. Multi-exon mis-splicing events, for instance, are not captured using a single-exon approach (Sangermano et al., 2019), and so our findings suggest that, where possible, assaying larger genomic regions may provide a more accurate picture of splicing dynamics.

Although we only identified one BP variant thus far confirmed to be pathogenic through our approach, this still constitutes a significant increase in the number of these ostensibly rare variants in the literature. It is promising that re-analysis of variants from a single gene panel cohort has led to a confirmed molecular diagnosis, and it can be speculated that re-analysis of further datasets may increase diagnostic yield further still.

Chapter 4

MRSD：a novel quantitative approach for assessing suitability of RNA-seq in the clinical investigation of mis-splicing in Mendelian disease

4. MRSD: a novel quantitative approach for assessing suitability of RNA-seq in the clinical investigation of mis-splicing in Mendelian disease

4.1. Abstract

RNA-seq of patient biosamples is a promising approach to delineate the impact of genomic variants on splicing, but variable gene expression between tissues complicates selection of appropriate tissues. Relative expression level is often used as a metric to predict RNA-sequencing utility. Here, we describe a gene- and tissue-specific metric to inform the feasibility of RNA-sequencing, overcoming some issues with using expression values alone.

We derive a novel metric, *Minimum Required Sequencing Depth* (MRSD), for all genes across three human biosamples (whole blood, lymphoblastoid cell lines (LCLs) and skeletal muscle). MRSD estimates the depth of sequencing required from RNA-sequencing to achieve user-specified sequencing coverage of a gene, transcript or group of genes of interest. MRSD predicts levels of splice junction coverage with high precision (90.1-98.2%) and overcomes transcript region-specific sequencing biases. Applying MRSD scoring to established disease gene panels shows that LCLs are the optimum source of RNA, of the three investigated biosamples, for 69.3% of gene panels. Our approach demonstrates that up to 59.4% of variants of uncertain significance in ClinVar predicted to impact splicing could be functionally assayed by RNA-sequencing in at least one of the investigated biosamples.

We demonstrate the power of MRSD as a metric to inform choice of appropriate biosamples for the functional assessment of splicing aberrations. We apply MRSD in the context of Mendelian genetic disorders and illustrate its benefits over expression-based approaches. We anticipate that the integration of MRSD into clinical pipelines will improve variant interpretation and, ultimately, diagnostic yield.

4.2. Introduction

Pinpointing disease-causing genomic variation informs diagnosis, treatment and management for a wide range of rare disorders. An underappreciated group of pathogenic variants is those that lie outside of canonical splice sites but act through disruption of pre-mRNA splicing, the process whereby introns are removed from nascent pre-mRNA to produce mature and functional transcripts (**Supplementary Figure 4a**). The ways through which genomic variants can disrupt pre-mRNA splicing are diverse (**Supplementary Figures 4b-g**), including both protein-coding and intronic variants that are well described as causes of rare disorders (Anna and Monika, 2018, Scotti and Swanson, 2016, Wai et al., 2020). However, the omission of intronic regions in targeted sequencing approaches (Sangermano et al., 2019, Khan et al., 2020), discordance between *in silico* variant prioritization tools (Rowlands et al., 2020) and the lack of availability of the appropriate tissue from which to survey RNA for splicing disruption (Aicher et al., 2020, Marston et al., 2009) limit effective identification of pathogenic splice-impacting variants.

RNA sequencing (RNA-seq) offers a potential route to overcome issues of variant interpretation (Wai et al., 2020, Mertes et al., 2021, Kremer et al., 2017, Byron et al., 2016, Marco-Puche et al., 2019, Cummings et al., 2017). The complex impacts of variants on splicing can be fully characterized through RNA-seq. Moreover, aberrant splicing events can be identified from RNA-seq datasets without prior knowledge of genomic variants driving their impact. Whilst targeted analyses, such as RT-PCR, also enable detection of splicing aberrations (Wai et al., 2020), such approaches are designed to test the presence of specific disruptions and may not identify the complete spectrum of splicing disruption caused by a single genomic variant.

There is growing evidence that RNA-seq can substantially improve diagnostic yield across a variety of disease subtypes (Wai et al., 2020, Kremer et al., 2017, Cummings et al., 2017, Frésard et al., 2019, Lee et al., 2020) through identification of variants impacting splicing, or leading to impairment of transcript expression or stability (Abdrabo et al., 2020). However, there remain several hurdles to the effective and routine integration of RNA-seq into diagnostic pipelines. For example, surveying a whole transcriptome identifies a large number of splicing events – in the order of hundreds of thousands. Despite a recent increase in the number of tools designed to scrutinize RNA-seq data for so-called “splicing outliers” (Mertes et al., 2021, Ferraro et al., 2020, Jenkinson et al., 2020, Cummings et al., 2017), there is little consensus regarding the best approach to filter true positive and pathogenic events from harmless or artefactual findings. Furthermore, diagnostic analysis using RNA-seq is only effective when sufficient levels of sequence coverage of a relevant gene transcript are present in the sampled tissue.

In this study, we develop an informatics approach to quantify the likelihood that a transcript, or a defined set of transcripts, can be appropriately surveyed using RNA-seq. We name our framework the *minimum required sequencing depth* (MRSD), which can be utilized in a flexible and customized manner to assess the suitability of RNA-seq derived from different tissues to identify pathogenic splicing aberrations in specific genes of interest (**Supplementary Figure 5**). MRSD scores (available at: https://mcgm-mrsd.github.io/) can be utilized to select the most appropriate biosample to detect splicing aberrations for a candidate set of transcripts, or to guide the amount of sequencing reads from a specific biosample required to generate appropriate transcriptomic datasets for a transcript of interest. We apply these techniques to the study of monogenic disease genes, and assess four clinically accessible biosamples for their appropriateness to survey all known monogenic disease genes.

4.3. Materials and methods

Minimum required sequencing depth (MRSD) score

We generated a collated map of splice junction coverage for GTEx samples from four tissues (whole blood, *n* = 150; LCLs, *n* = 91; skeletal muscle, *n* = 184; cultured fibroblasts, *n* = 150; see *Control RNA-seq data acquisition*, below), using established methods (Cummings et al., 2017). These samples were designated as *reference sets.* Our model considers the level of sequencing coverage for splice junctions in each tissue-specific reference set and calculates the minimum required sequencing depth (MRSD), in millions of uniquely mapping 75 bp reads, that would be required for the desired proportion of splice junctions in a given transcript to be covered by a desired number of sequencing reads. The model is dynamic, and can be adjusted by the user to account for customized levels of desired sequencing coverage per splicing junction, the proportion of splicing junctions covered, and the so-called "MRSD parameter", representing the proportion of control samples for which the returned MRSD holds true (suggested usage of 95 or 99%).

MRSD is defined for an individual transcript in a given sample as:

$$MRSD_m = r/\left(\frac{R_p}{d}\right)$$

Where r is the desired level of read coverage across desired proportion p of splice junctions, R is the set of read counts supporting each of the splice junctions in the transcript of interest, ordered from lowest to highest, and R_p is the read count at the position in R at which proportion p of read counts values in R are greater than or equal to it. d represents the total number of sequencing reads, in millions of reads, in the RNA-seq sample (by default, the number of uniquely mapping sequencing reads).

For instance, suppose a sample sequenced to a depth (d) of 40 M uniquely mapping sequencing reads generates coverage of 14, 16, 8 and 10 reads across the splice junctions of a five-exon transcript. Suppose we wish 75% of splice junctions to be covered by a minimum of 6 reads (i.e. p = 0.75 and r = 6). Here, R = (6, 10, 14, 16) and R_p = 10, as 3/4 (75%, i.e. p) of all values in R are greater than or equal to 10. Inserting these values into the formula shows that this transcript has an MRSD of $\frac{6}{10/_{40}} = 24\,M$ uniquely mapping sequencing reads in this sample.

The set of MRSD scores for the given transcript are then collated across all control samples and ordered from lowest to highest. The score at the m-th percentile position in the collated list of sample-specific MRSDs is returned as the overall MRSD for that transcript, where m is termed the "MRSD parameter" and is customizable by the user (default = 0.95). The $\text{MRSD}_{0.99}$ of a transcript, for example, represents the sequencing depth that would be required for 99% of control samples to achieve the specified coverage for that transcript. The MRSD parameter therefore approximately represents the likelihood that a sequencing run at the returned depth will yield the desired coverage level. An illustrated example of MRSD generation is provided in **Supplementary Methods 1**.

Transcript selection

MRSD can be calculated for any transcript sets of interest. To extend this to the gene level for the analyses shown here, we generated a single transcript model for each gene in the GENCODE v19 human genome annotation (**Supplementary Methods 2**). We utilized a hierarchical approach for transcript selection, whereby we prioritized transcripts in the MANE v0.7 curated transcript list, providing that all splicing junctions for a given transcript were supported in the GENCODE v19 annotation. Genes without MANE transcripts were assigned composite transcripts, consisting of the union of all junctions found in transcripts for the given gene in NCBI RefSeq. For genes that matched neither criteria, the union of all junctions present in all GENCODE v19-listed transcripts for that gene were used as the transcript model.

Genomics England PanelApp data collection

Tabulated versions of 295 gene panels were downloaded from the Genomics England PanelApp repository on June 28th 2021. Each panel was filtered to retain only genes assigned a “green” classification for that panel, representing the highest level of confidence of a real genotype-phenotype association.

Control RNA-seq data acquisition

FASTQs were downloaded from the Database of Genotypes and Phenotypes (dbGaP) under the project accessions phs000424.v8.p2 and phs000655.v3.p1.c1 for GTEx control individuals and neuromuscular disease patients, respectively. GTEx controls were selected for LCLs (*n* = 91), skeletal muscle (*n* = 184), whole blood (*n* = 150) and cultured fibroblasts (*n* = 150) according to tissue-specific criteria (**Supplementary Methods 3**) to ensure use of only high-quality samples in generating control splicing datasets.

In-house RNA-seq generation

RNA-seq datasets used to evaluate model performance were accessed from previously published datasets (Cummings et al., 2017), under dbGaP project accession phs000655.v3.p1.c1, through international consortia (Osborne et al., 2000), or from individuals in whom written informed consent was obtained and ethical approval for the study granted by Scotland A (refs: 06/MRE00/76 and 16/SS/0201), South Central-Hampshire A (ref: 17/SC/0026), South Central-Oxford B (ref:11/SC/0269) or South Manchester (ref: 11/H10003/3).

For in-house peripheral blood samples, RNA was extracted from PAXgene Blood RNA Kits and underwent poly-A enrichment library preparation using the TruSeq Stranded mRNA assay (Illumina) followed by 76 bp paired end sequencing using an Illumina HiSeq 4000 sequencing platform. For in-house LCL samples, RNA was extracted from pelleted LCLs thawed directly into TRIzol reagent

(Invitrogen, 15596-026) using chloroform, and treated with TURBO DNase (Invitrogen, AM1907), following the manufacturers' instructions. RNA was prepared using the NEBNEXT Ultra II Directional RNA Library Prep kit (NEB #7760) with the Poly-A mRNA magnetic isolation module (NEB #E7490), according to manufacturer's instructions, and 75bp paired end sequencing was performed using the Illumina NextSeq 550 sequencing platform. Ribosomal RNA depleted datasets were generated using RNA extracted via the PAXgene Blood RNA system, and 150bp paired end sequencing performed via Novogene (Hong Kong) using the NEBNext Globin and rRNA Depletion and NEBNext Ultra Directional RNA Library Prep Kits on a HiSeq 2000 instrument (Illumina). RNA samples from 20 LCLs were obtained from the kConFab consortium. Poly(A)-selected RNA was generated using the TruSeq Stranded mRNA Library Prep Kit (Illumina), and 150bp paired end reads created using the NextSeq 500 instrument (Illumina).

Splice event identification

All FASTQs were aligned and processed as previously described (Cummings et al., 2017). Briefly, this analysis consisted of two-pass alignment using STAR v2.4.2 (Dobin et al., 2013), marking of suspected PCR duplicates, and processing of the resulting alignments to generate tissue-by-tissue lists of read support counts for splice junctions present within the samples in the cohort. Metrics for each splicing event were collected (Box 1), and splicing junctions were filtered to retain only those events that were unique to single samples (singletons) or that were present in multiple samples (non-singletons) but with an increased usage in the sample of interest, that is, with a higher normalized read count (NRC), than any control. The resulting list was ranked according to NRC fold change, with singletons with high read counts considered the most significant events. The resulting junctions were considered "events of interest".

Factors influencing the likelihood of aberrant splicing identification

To calculate how the level of background splicing aberrations was altered by sample size, each individual in the three control splicing datasets was processed using the above pipeline (Cummings et al., 2017) and compared against 2000 bootstraps of 30, 60 and 90 controls each from their respective control tissue dataset with replacement. Events were then filtered to retain only those events for which the NRC was higher in the given individual than in any controls, and then counted for each bootstrap. Median counts for singleton and non-singleton events were collated for each control group size. We selected 31 splicing events identified in neuromuscular patient RNA-seq data that were either unique to, or highly increased in prevalence in, the individual. From the genes in which we identified these variants, samtools was used to remove random subsets of reads in 10% intervals from each of these events to simulate variability in the number of reads generated for the gene of interest. The resulting datasets, exhibiting variable expression of a single gene, were then rerun through the splice analysis pipeline and the above metrics gathered for these simulated datasets.

Curation of ClinVar variants of uncertain significance

A tabulated version of the comprehensive ClinVar variant listing (Landrum et al., 2018) for January 2021 was downloaded and filtered to retain only those variants that were annotated as either "Uncertain significance" or "Conflicting interpretations of pathogenicity". SpliceAI scores (v1.2.1, Jaganathan et al., 2019) were generated for these variants and those with a score of 0.5 or greater retained for downstream analysis.

4.4 Results

4.4.1. Minimum required sequencing depth (MRSD) scores differ across biosamples

We first curated a list of disease-related genes, comprising 3417 unique genes listed in at least one of the 295 disease panels in the Genomics England PanelApp repository. 95 single-exon genes were removed from this set, leaving 3322 genes for downstream analysis. MRSD scores were generated for each of these genes, corresponding to the required sequencing depth (in millions of uniquely mapping sequencing reads) for a specified level of coverage of the corresponding transcript, in four clinically relevant tissues: whole blood, cultured fibroblasts, LCLs and skeletal muscle. As MRSD is a transcript-level metric, each gene was assigned a single transcript using a hierarchical approach (see *4.3.*). Three parameters can be altered for the MRSD model; we observed that MRSD differed dependent on the values chosen for these parameters, which comprised the number of reads desired to cover each splice junction, the proportion of splicing junctions for each gene that must meet this coverage threshold (75% or 95%), and the proportion of sequencing runs for which the predicted depth is predicted to achieve the desired level of coverage (the "MRSD parameter" of either 95% or 99%, denoted $\mathrm{MRSD}_{0.95}$ and $\mathrm{MRSD}_{0.99}$, respectively; **Figure 17a-b**). For example, when specifying a desired read coverage level of eight reads per splicing junction, we observed that increases in the desired proportion of covered splice junctions from 75-95% was associated with an increase in median MRSD of between 0.27% (in skeletal muscle, $\mathrm{MRSD}_{0.99}$) to 55.95% (in LCLs, $\mathrm{MRSD}_{0.95}$; **Figure 17b**, top). For all but one parameter combination, moving from $\mathrm{MRSD}_{0.95}$ to $\mathrm{MRSD}_{0.99}$ resulted in an increase in median MRSD of between 26.19-155.40%. However, when stipulating 95% splice junction coverage for skeletal muscle samples , we observed a decrease of 4.66% in MRSD scores when the MRSD parameter was increased from 95% (*n* = 1323, median = 42.52) to 99% (*n* = 973, median = 40.54); this was accounted for by an increase in the number of genes that were considered "unfeasible" for surveillance, i.e. those for which zero reads cover the given proportion of junctions (*n* unfeasible, $\mathrm{MRSD}_{0.95}$ = 1999, *n* unfeasible, $\mathrm{MRSD}_{0.99}$ = 2349). This definition of feasibility is limited by the sequencing depth of the control models on which the predictions are based. For example, no coverage of splice junctions in a particular transcript may have been observed simply due to low sequencing depth; with ultra-deep sequencing of the same sample, we may have observed coverage of splice junctions and so have been able to generate a feasible MRSD prediction.

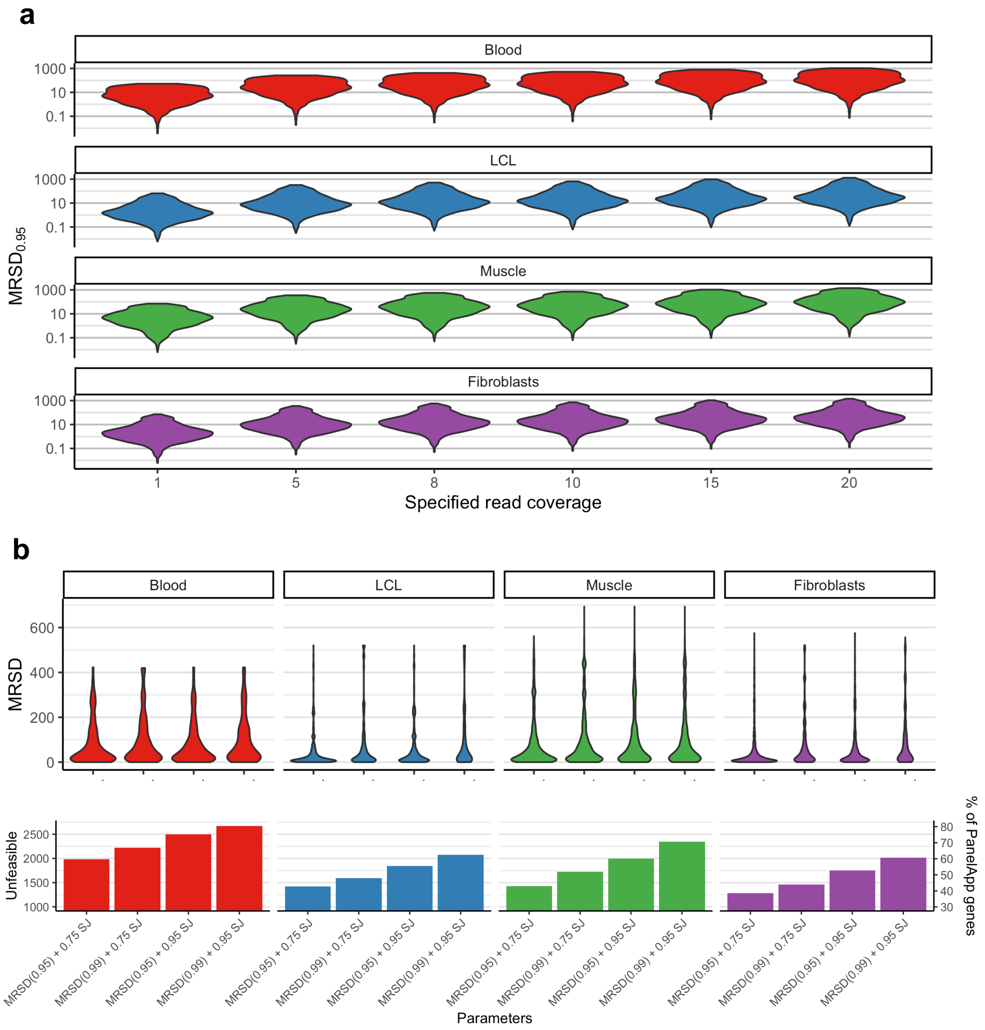


Figure 17. *Minimum required sequencing depth (MRSD) predictions vary with changes in model parameters and across tissues.* (**a**) When all other parameters are constant (default parameters used here), increasing the desired level of read coverage of a gene results in a proportional increase in MRSD. The distribution of MRSD scores for 3322 PanelApp genes in lymphoblastoid cell lines (LCLs) appears to be the lowest of the 3 tissues (median = 15.975 M at 10 reads), while whole blood exhibits the highest overall MRSD scores (median = 50.95 M at 10 reads), suggesting coverage of disease genes is generally poorer in blood. (**b,** top) In most cases, for a given level of splice junction (SJ) coverage, increasing the desired confidence level (the proportion of RNA-seq runs for which the MRSD prediction is expected to be sufficient) results in an increase in median MRSD score. (**b**, bottom) The number of PanelApp genes for which no amount of sequencing is predicted to yield the specified level of coverage increases gradually as parameter stringency increases. At the highest level of stringency, the specified coverage was predicted unfeasible for between 60.7% (2017/3322, in fibroblasts) and 80.3% (2668/3322, in blood) of PanelApp genes.

Overall, these analyses suggested that, of the four investigated biosamples, fibroblasts would enable investigation of the most comprehensive set of genes for aberrant splicing. Although LCLs displayed, across all four parameter combinations, the lowest median MRSDs (range = 12.86-33.77, **Figure 17b**, top), the difference in median MRSDs compared to fibroblasts was small (range = 14.44-35.06), while a greater number of genes were predicted “unfeasible” for analysis in LCLs than fibroblasts (42.8-62.5% vs. 38.6-60.7% of PanelApp genes, respectively). On the other hand, whole blood exhibited the highest number of unfeasible genes across the different parameter combinations (59.7-80.3%).

4.4.2. Accuracy of minimum required sequencing depth (MRSD) calculations

We next obtained independent RNA-seq datasets for 68 samples from the three investigated tissues (blood, *n* = 12; LCLs, *n* = 4; muscle, *n* = 52), with a range of sequencing depths (**Supplementary Figure S6**). Across a variety of parameter combinations, we considered all transcripts for which the parameter-specific MRSD was lower than that of the sequencing depth of the sample, i.e. those genes which are expected to be covered to at least the specified level. The positive predictive value (PPV) of a sample was defined as the proportion of these transcripts that did achieve the specified level of coverage. Conversely, the negative predictive value (NPV) was calculated as the proportion of transcripts with an MRSD greater than the sample sequencing depth which, as predicted did not achieve the specified level of coverage. Across all investigated MRSD parameters, we observed 96% PPV and 79% NPV, on average, for the 68 samples (**Figure 18a**). We observed a general trend that the PPV and NPV of MRSD decreased and increased, respectively, as higher levels of required coverage were imposed (**Figure 18b-c**). Across all parameter combinations, PPV values ranged from 90.1-98.2%, while NPV ranged from 56.4-94.7%, suggesting MRSD is a fairly conservative model that primarily returns positive results with high certainty.

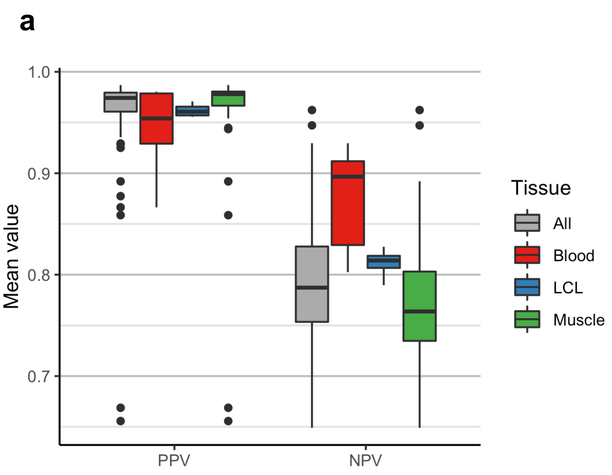
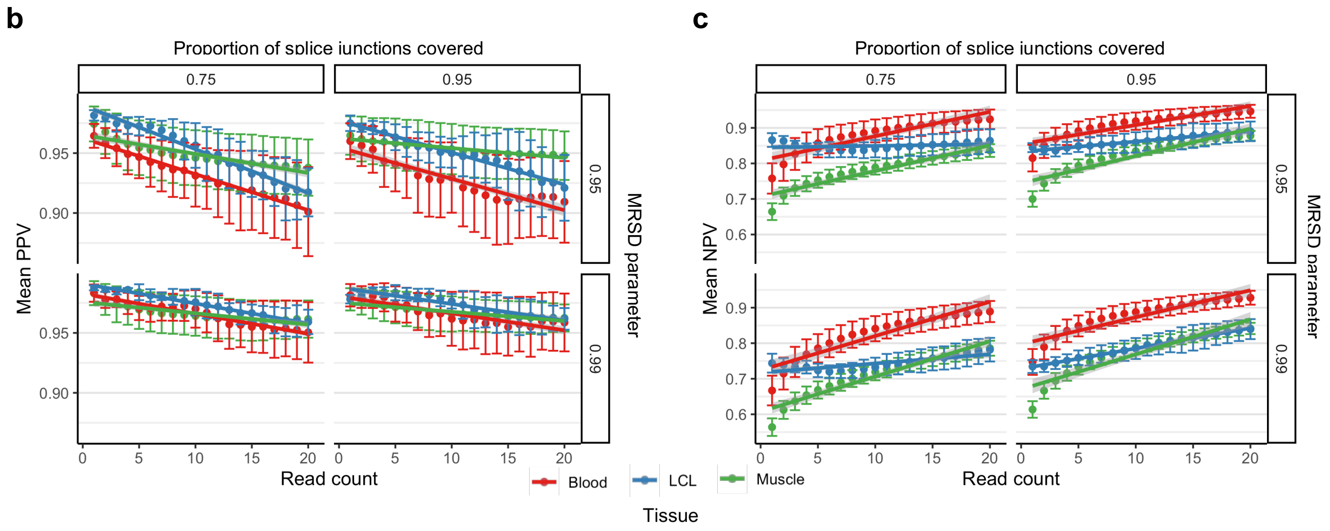


Figure 18. *Performance metrics of the MRSD model.* The ability of MRSD to accurately predict levels of PanelApp disease gene coverage based on sequencing depth was tested on unseen RNA-seq datasets from blood (*n* = 12), LCLs (*n* = 4) and muscle (*n* = 52). (**a**) The mean positive predictive values (PPVs) and negative predictive values (NPVs) averaged across all parameter combinations for each RNA-seq dataset show that the median PPV is slightly lower, and the median NPV slightly higher, for whole blood than for LCLs and skeletal muscle. Breakdown of (**b**) PPVs and (**c**) NPVs for the MRSD model by parameters shows that specifying an increasing required read coverage results in a gradual decrease in PPV and increase in NPV across all tissues and parameter combinations. Dependent on parameter stringency, and limiting analysis to a maximum specification of 20-read coverage, PPV predictions range from 90.1-98.2%, while NPV ranges from 56.4-94.7%. Overall, the model is fairly conservative and returns positive predictions only when they are deemed likely to be true.



Although MRSD scores were derived from 75 bp paired-end RNA-seq data, we evaluated the ability of the model to predict transcript coverage in 150 bp paired-end data (LCLs, *n* = 20), and observed higher median PPV across samples than with 75 bp data for half of the four parameter combinations tested, while NPV was only slightly lower for all combinations (**Supplementary Figure S7**). While MRSD scores should ideally be applied to datasets generated using the same experimental approach, these data suggest that MRSDs may be loosely applicable between related methodologies.

We additionally generated MRSD scores for the 3322 disease genes *de novo* based on the 150 bp dataset, and compared these to scores derived from a trimmed version of the same dataset, with all reads trimmed to 75 bp or fewer (**Supplementary Figure S8**). We observed that coverage was too poor to allow MRSD score generation regardless of read length for 45.8% (1520/3322) of disease-associated genes. However, of the remaining 1802 genes, 13.5% (243/1802) counter-intuitively exhibited a higher MRSD in the 150 bp dataset, suggesting that fewer 75 bp reads than 150 bp reads were required to adequately cover these transcripts. In many cases, this was found to be due to a decrease in mapping quality of longer reads such that the reads did not pass the quality filters of the employed pipeline (Cummings et al., 2017). Further work is needed to ascertain whether this discarding of longer reads is a harmful artefact of the filtering process, or a genuine removal of uninformative reads.

4.4.3. Comparison of MRSD and TPM as a guide for appropriate surveillance

We compared MRSD to the use of relative expression level (in transcripts per million, TPM) as a possible indicator of RNA-seq suitability for the detection of aberrant splicing events. We compared the expression levels, in TPM, of the 3322 disease-associated genes against tissue-specific MRSD predictions, finding a negative correlation between the level of gene expression and its predicted MRSD across all four tissues (r^2 = 0.613-0.714; Figure **19a-d**). This confirms that more highly-expressed genes are associated with lower MRSD scores. However, we noted significant overlap between genes grouped into low-MRSD (< 100 M reads) and high-MRSD (≥ 100 M reads) brackets. For example, among genes considered low-MRSD, TPM values ranged from 0.99-246,600, while feasible genes with high-MRSD values had TPM values between 0.14-8644 (Figure **19e**). We quantified the overlap between these distributions, demonstrating that, depending on the tissue, between 93.0% and 99.3% of high-MRSD genes had higher TPM values than at least one low-MRSD gene. We also observed that, in their respective tissues, the TPMs of 44.0-60.0%, 8.5-16.7% and 3.4-6.6% of high-MRSD genes exceeded those of the 5%, 30% and 50% least-expressed low-MRSD genes, respectively (**Figure 19e**). The substantial overlap in the TPM values for low and high MRSD genes suggests that relative expression does not provide a wholly accurate representation of transcript coverage in RNA-seq data. Such inconsistencies may arise from bias in the regions of genes that are sequenced, for example, genes with high degrees of 3’ bias in RNA-seq datasets (**Supplementary Figure S9**).

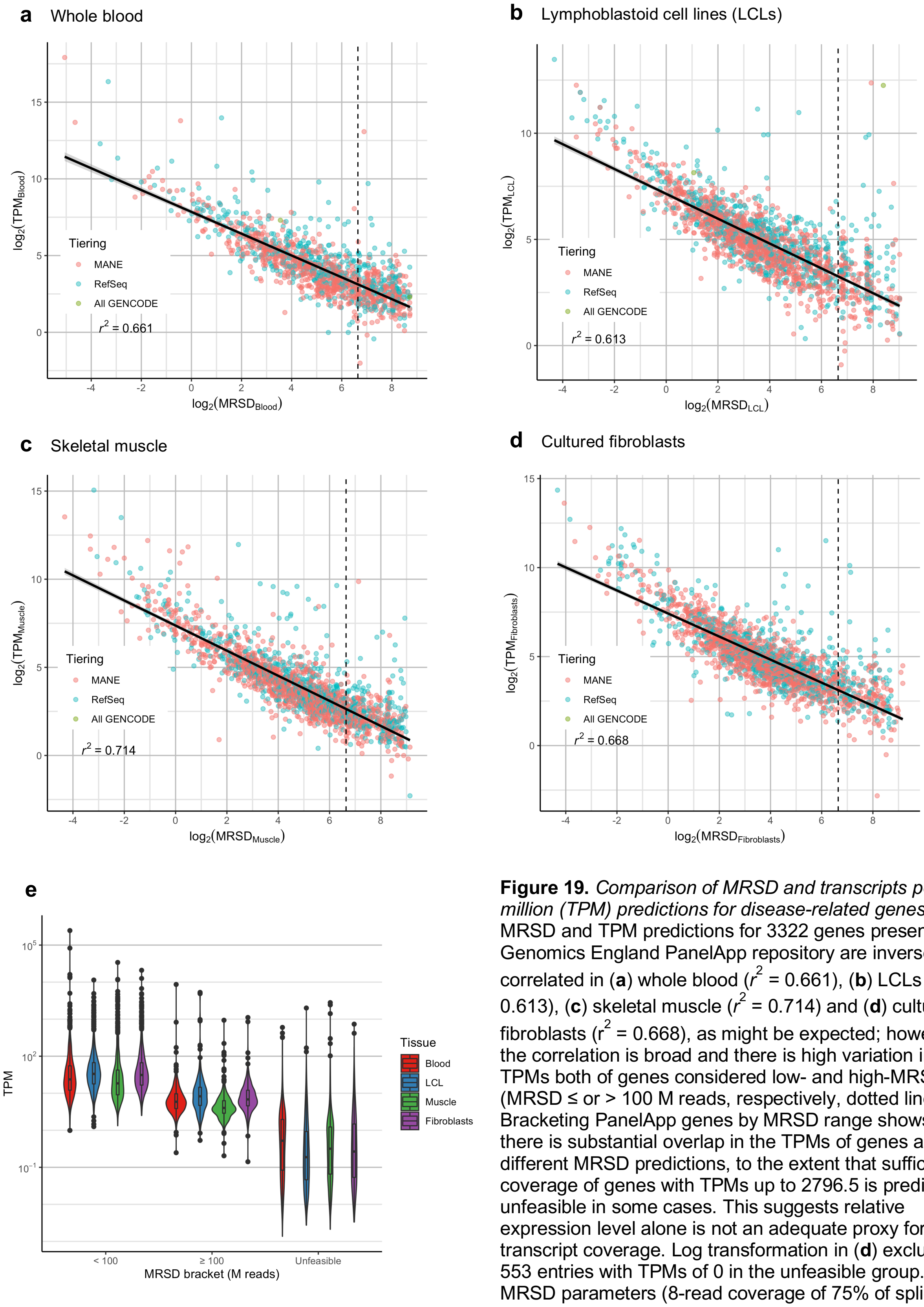


Figure 19. *Comparison of MRSD and transcripts per million (TPM) predictions for disease-related genes.* MRSD and TPM predictions for 3322 genes present in the Genomics England PanelApp repository are inversely correlated in (**a**) whole blood (r^2 = 0.661), (**b**) LCLs (r^2 = 0.613), (**c**) skeletal muscle (r^2 = 0.714) and (**d**) cultured fibroblasts (r^2 = 0.668), as might be expected; however, the correlation is broad and there is high variation in the TPMs both of genes considered low- and high-MRSD (MRSD ≤ or > 100 M reads, respectively, dotted line). (**d**) Bracketing PanelApp genes by MRSD range shows that there is substantial overlap in the TPMs of genes across different MRSD predictions, to the extent that sufficient coverage of genes with TPMs up to 2796.5 is predicted unfeasible in some cases. This suggests relative expression level alone is not an adequate proxy for transcript coverage. Log transformation in (**d**) excludes 553 entries with TPMs of 0 in the unfeasible group. Default MRSD parameters (8-read coverage of 75% of splice junctions, confidence level of 95%) used throughout.

4.4.4. Traits of pathogenic splicing variation vary widely between genes and events

We next aimed to determine the optimal MRSD parameters for detection of aberrant splicing. This required a deeper understanding of the proportion of transcripts likely impacted by pathogenic splicing events. We investigated 21 RNA-seq samples from patients harboring pathogenic mis-splicing events using a previously described analysis pipeline (Cummings et al., 2017). These included a wide variety of mis-splicing effects (**Supplementary Figure S10**). We calculated median TPM and MRSD values for the genes in which the mis-splicing events were present (**Supplementary Table S12**). The method employed for aberrant splicing detection pooled read support counts for splicing junctions from reference RNA-seq datasets to generate tissue-specific models of "healthy" splicing. We then generated splice junctions counts from the 21 patient RNA-seq datasets and merged these with those in the healthy splicing models (datasets summarized in **Supplementary Table S13**), collecting pipeline-specific metrics indicative of aberrant splicing events (**Box 1**). We observed high variability in all metrics associated with pathogenic aberrant splicing events (**Table 6**). All patients harbored at least one pathogenic splicing event supported by two or more reads and with normalized read counts (NRCs) ⩾ 0.19, and 80% of these events had a relative fold change in NRC > 19x relative to controls (**Table 6**). While a blanket set of parameters for all aberrant splicing events may be unsuitable, our data suggests that 90% of pathogenic events could be retained if filtering for events that were singletons (evident only in a single sample), or were non-singletons with an NRC > 0.25.

We further investigated the ability of three recent splice prediction tools to identify the 21 pathogenic events; these were FRASER (Mertes et al., 2021), SPOT (Ferraro et al., 2020) and LeafCutterMD (Jenkinson et al., 2020). We observed that the performance of these tools was mixed: LeafCutterMD performed the worst, identifying 57.1% (12/21) of the events, while SPOT and FRASER identified 76.2% (16/21) and 81.0% (17/21) events, respectively. Of the 17 events identified by FRASER, only one was not identified as being a statistically significant splicing outlier ($p < 0.05$), likely due to this variant (in BRCA1) being present in three members of the LCL patient cohort investigated. FRASER was still able to assign significance to a pathogenic mis-splicing event supported by just 3 reads; however, while the pipeline used for our initial analysis (Cummings et al., 2017) is able to identify pathogenic splicing events with a lower number of supporting reads, it is likely that significance-based tools, such as FRASER, LeafCutterMD and SPOT, may require a deeper amount of sequencing to highlight these events as significant.

Box 1. *Metrics collated during splice event analysis*

- Read count – Number of split reads supporting the existence of a given splice junction
- Normalized read count (NRC) – Ratio of the number of reads supporting a given junction to the numbers of reads supporting adjoining canonical junction with the highest supporting read count
- NRC fold change – fold difference in NRC for a given event between an individual and the control individual with the next-highest NRC for that event
- Number of samples – the number of individuals, across both case and controls, in which an event is present
- Rank – position of a given event in a list of significant events, when ordered by decreasing read count (for singleton events) or fold change (for non-singleton events)

	Tissue		
Metric	**Whole blood (n=3)**	**LCLs (n=7)**	**Skeletal muscle (n=11)**
Read count	2-40	4-38	2-462
NRC	0.48-1.25	0.19-1.52	0.34-3.19
NRC fold change	Singletons	3.7-8.2 + singletons	19.6-442 + singletons
Number of samples	1	1-48	1-110
Rank	2-5	10-232	1-342
FRASER events identified	3/3	4/7	10/11
FRASER p-values	7.97 x 10^{-11} - 0.0022	2.36 x 10^{-5} - 0.13182	4.27 x10-13 - 0.0160
LeafCutterMD events identified	3/3	2/7	7/11
LeafCutterMD p-values	6.19 x 10^{-11} - 0.00936	7.66 x 10^{-6} - 0.586	2.2 x 10^{-15} - 1.35 x 10^{-3}
SPOT events identified	3/3	6/7	7/11
SPOT p-values	0.000181 - 0.0426	1 x 10^{-6} - 0.13582	0.00469 - 0.0159

Table 6. *Range of splice-related metrics observed in known pathogenic splicing events*

4.4.5. Factors influencing the likelihood of pathogenic splicing variation identification & MRSD predictions

To further define the most informative parameters for use in the MRSD model, we investigated the impact of a variety of metrics on the capability to identify pathogenic splicing events, including number of samples within the healthy reference set, the degree of read support for splicing junctions, and the relative expression of genes of interest. Namely, we aimed to quantify the effect of changes in these metrics on both the total number of events of interest and the position within the list of events (see *4.3.* for filtering and ranking strategy). Overall, our analyses suggested that two supporting reads for an aberrant splicing event that is novel or has an NRC > 0.25 would reliably highlight pathogenic aberrations amongst transcriptome-wide splicing variation. We acknowledge that, for variant reporting purposes, a higher read support may be desired for improved certainty; however, for the purposes of highlighting pathogenic splicing variation in the first instance, we have found these filtering parameters to be robust.

We first identified how the number of control samples used as a reference set for “healthy splicing” impacted our ability to identify aberrant splicing events. For all samples within our healthy splicing set, we iteratively selected groups of control samples at sizes of 30, 60 or 90. We observed that moving from 30 to 60 controls is associated with a mean reduction in event count of 19.3% (28.1% of non-singleton events, 17.1% of singleton events) across the three tissues, while increasing the control size to 90 results in a further reduction of 10.2% of events (16.5% of non-singleton events, 9.5% of singleton events; **Figure 20**); this effect was consistent across tissue types.

We next investigated how read count filters impacted the number of events observed for a given individual (**Figure 20**). Filtering out all splicing events supported by just a single read against a background of 90 control samples removes, on average, 91.2% of events (60.4% of non-singleton events, 97.3% of singleton events). Increasing read support thresholds to 10 unique sequencing reads results in a total of 99.4% of events being excluded on average (96.2% of non-singleton events, 99.99% of singleton events), while retaining only those events supported by 100 reads or more removes an average of 99.97% of events (99.8% of non-singleton events, 100.0% of singleton events).

To understand how the level of read support impacted the ability to identify specific events, we collated 31 aberrant splicing events across 22 muscle-derived RNA-seq samples, and downsampled

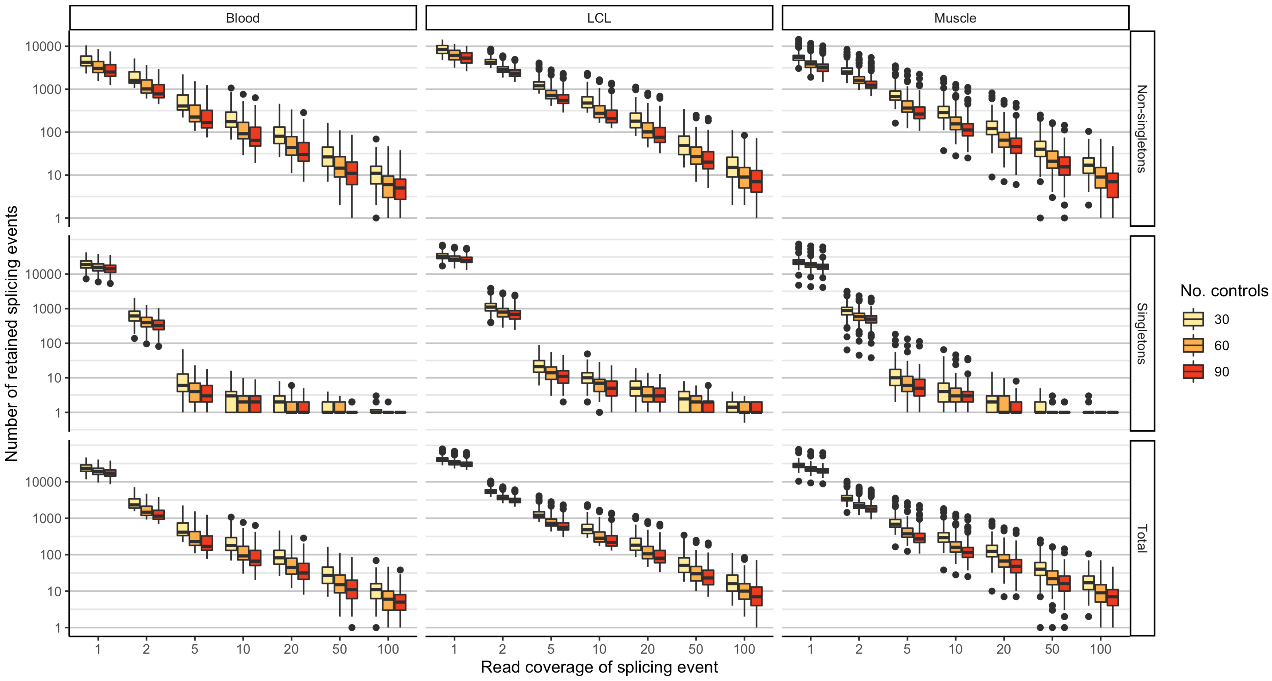


Figure 20. *Bootstrapping reveals the filtering power of increasing control dataset size and enforcing read filter thresholds in splice event analysis.* Counting the significant events identified in each individual in a control splicing dataset when analysed against 2000 bootstraps each of 30, 60 and 90 other individuals from within the control dataset for the same tissue reveals a small decrease in the number of total events identified as control dataset size increases, predominantly from non-singleton events. Enforcing a read coverage threshold has a more significant effect on event counts, particularly for singleton events, where filtering out events supported by a single read removes up to 95% of singletons. LCLs appear to exhibit the greatest number of splicing events regardless of filter, although this may be due to differences in sequencing depth between tissues.

reads in the genes containing these events to simulate reduction in expression. We observed that we could identify the same aberrant splicing events at reduced relative expression levels, and, while read support decreased (**Figure 21a**), the ranked position of the event within the rank-ordered output remained approximately the same in most cases (**Figure 21b**). However, the weakened read support increased the risk of eliminating the variant from consideration when read count filters were applied (**Figure 21c**). This analysis further emphasized that TPM values alone may not be a reliable measure of ability to survey all splicing junctions within a gene; we observed that splice junctions in different samples covered by the same number of sequencing reads belonged to genes with widely ranging TPM values (**Supplementary Figure S11**). For example, splice junctions covered by eight reads were identified in genes with TPMs ranging between 0.17 and 52.

Based on these investigations, we selected an eight-read coverage value for downstream analyses; as we observed that the majority of pathogenic mis-splicing events have an NRC ⩾ 0.25; stipulating

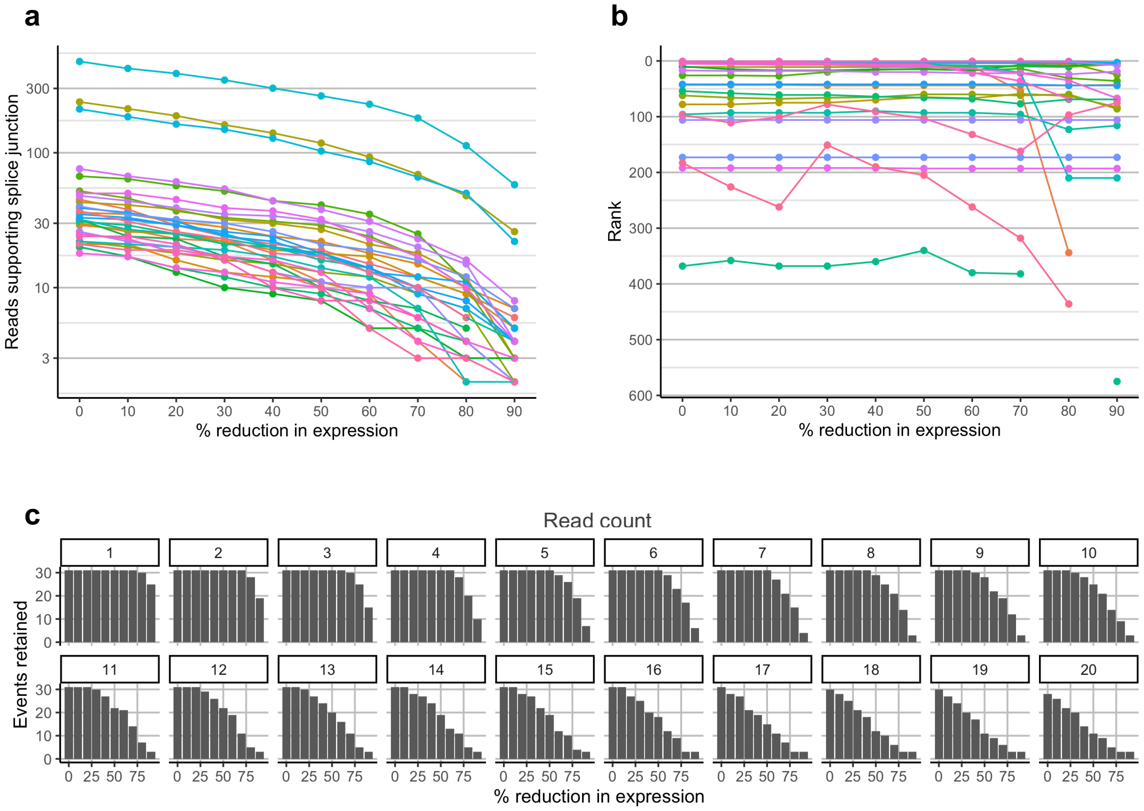


Figure 21. *Variability in expression level influences the capacity to identify mis-splicing events.* Genes harboring a selection of 31 splicing events that were identified during analysis of 52 muscle-based RNA-seq datasets (and which would be identified as events of interest using a filter of normalized read count (NRC) > 0.19) were artificially downsampled to simulate variation in expression. (**a**) Reduction in expression leads to an intuitive and proportional reduction in the number of reads supporting each mis-splicing event. (**b**) The rank position – where the event appears in a list of all splicing events in its respective sample, ordered by decreasing NRC fold change relative to controls, and – is generally consistent as expression of the gene decreases; however, for a subset of events, reduction in expression is sufficient to cause stochastic changes in the NRC value, and so cause movement of the event down the prioritized list. (**c**) Variation in expression impacts our ability to identify events of interest when filters of read count supporting the events are enforced. When the 31 events experience a 50% reduction in expression, for instance, the application of a minimum 15-read filter leads to the exclusion of 41.9% (13/31) of events**.**

an eight-read coverage requirement means that aberrant events should be covered by at least two reads, and so be retained when filtering single-read events from the list of splicing events. We appreciate that the use of more stringent parameters may be preferable in some use cases, such as to generate sufficient corroboration to support the reporting of a diagnostic finding to a patient or when using significance-based tools such as FRASER, LeafCutterMD and SPOT. However, our investigations have shown this approach to be robust for the initial highlighting of aberrant splicing events for downstream analysis.

4.4.6. Implications for investigation of variants in known disease-causing genes

We next sought to investigate the disease. Based on our above investigations, we generated MRSD scores for the cohort of 3322 multi-exon disease genes using the following parameters: read coverage = 8; proportion of junctions = 75%; MRSD parameter = 95% across whole blood, LCLs and skeletal muscle. We acknowledge that these parameters may be too lenient for some use cases, but

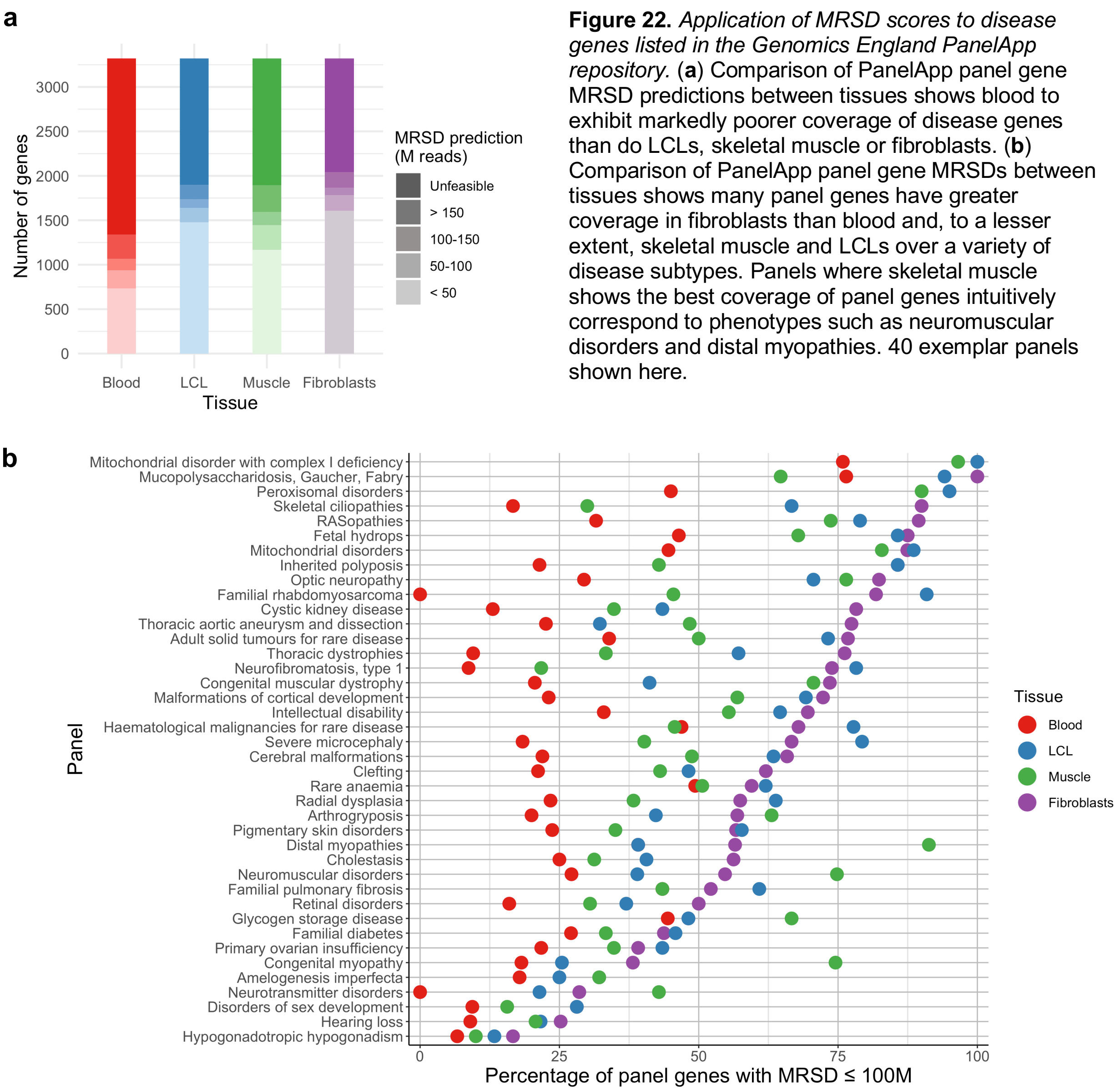


Figure 22. *Application of MRSD scores to disease genes listed in the Genomics England PanelApp repository.* (**a**) Comparison of PanelApp panel gene MRSD predictions between tissues shows blood to exhibit markedly poorer coverage of disease genes than do LCLs, skeletal muscle or fibroblasts. (**b**) Comparison of PanelApp panel gene MRSDs between tissues shows many panel genes have greater coverage in fibroblasts than blood and, to a lesser extent, skeletal muscle and LCLs over a variety of disease subtypes. Panels where skeletal muscle shows the best coverage of panel genes intuitively correspond to phenotypes such as neuromuscular disorders and distal myopathies. 40 exemplar panels shown here.

anticipate that the evidence for the majority of mis-splicing events can be reliably identified using these parameters. Evaluation of model performance was not possible for fibroblast MRSDs due to a lack of independently generated samples. Using this approach, and with expected PPV = 0.936-0.974, NPV = 0.776-0.880 across the three tissues, we observed that 64.2% (2133/3322) of PanelApp genes were predicted to be low-MRSD (< 100 M reads required) in at least one of the four tissues (**Figure 22a**). At the individual tissue level, 28.2% (936/3322) of PanelApp genes in whole blood, 49.4% (1641/3322) in LCLs, 43.6% (1447/3322) in skeletal muscle, and 53.7% (1784/3322) in cultured fibroblasts were predicted to be low-MRSD (**Figure 22a**). Of note, fibroblasts were observed to have the highest (or joint-highest) proportion of low-MRSD panel genes in 186/295 disease gene panels (63.1%, **Figure 22b**). This was the case for LCLs in 126/295 panels (42.7%), and skeletal muscle in 70/295 panels (23.7%). In only 21/295 panels (7.1%) did whole blood exhibit the highest proportion of low-MRSD genes.

MRSD predictions revealed many use cases for specific tissues: in the familial rhabdomyosarcoma panel, for example, none of the 11 genes were predicted to be low-MRSD in blood, while 10/11 were predicted low-MRSD in LCLs (**Figure 22**), of which nine were actually assigned an MRSD < 50 M reads.

Overall, this analysis suggests both that whole blood may often represent the poorest choice of RNA source tissue in terms of disease gene coverage; in contrast, fibroblasts appear to show robustly high coverage of splice junctions in disease gene transcripts across diverse disease subtypes, and so may represent a more reliable source of RNA for clinical transcriptomic investigations.

4.4.7. Quantifying the resolving power of RNA-seq for variants of uncertain significance

To analyze the possible impact of diagnostic RNA-seq integration on variant interpretation, we curated variants of uncertain significance (VUSs) from the ClinVar variant database (Landrum et al., 2018) that were predicted by SpliceAI (Jaganathan et al., 2019) to impact splicing (score ≥ 0.5; see *4.3.*). Of a total of 352,011 ClinVar variants, 185,119 (52.6%) were identified as VUSs, and 7,507 (2.1%) were retained after filtering based on SpliceAI score. Cross-referencing the MRSDs of the genes harboring SpliceAI-prioritized variants across tissues revealed that, at a specified read coverage of 8 reads, between 25.8% and 67.8% of these variants may lie in genes that are low-MRSD in at least one of the four tissues (**Figure 23a**), dependent on the stringency of the model. This range lies between 24.9-64.0% when specifying 10 reads, and 18.7-52.0% when specifying a coverage of 20 reads (**Supplementary Figure S12**), suggesting just under one in seven VUSs may be investigated using RNA-seq when using very high levels of stringency.

Further, among the 30 genes in which the greatest number of predicted splice-impacting VUSs were identified, 23 were predicted to be low-MRSD in at least one tissue (**Figure 23b**). Interestingly, raising the specified read coverage from 8 to 10 reads removes only one further gene, ATM, from the low-

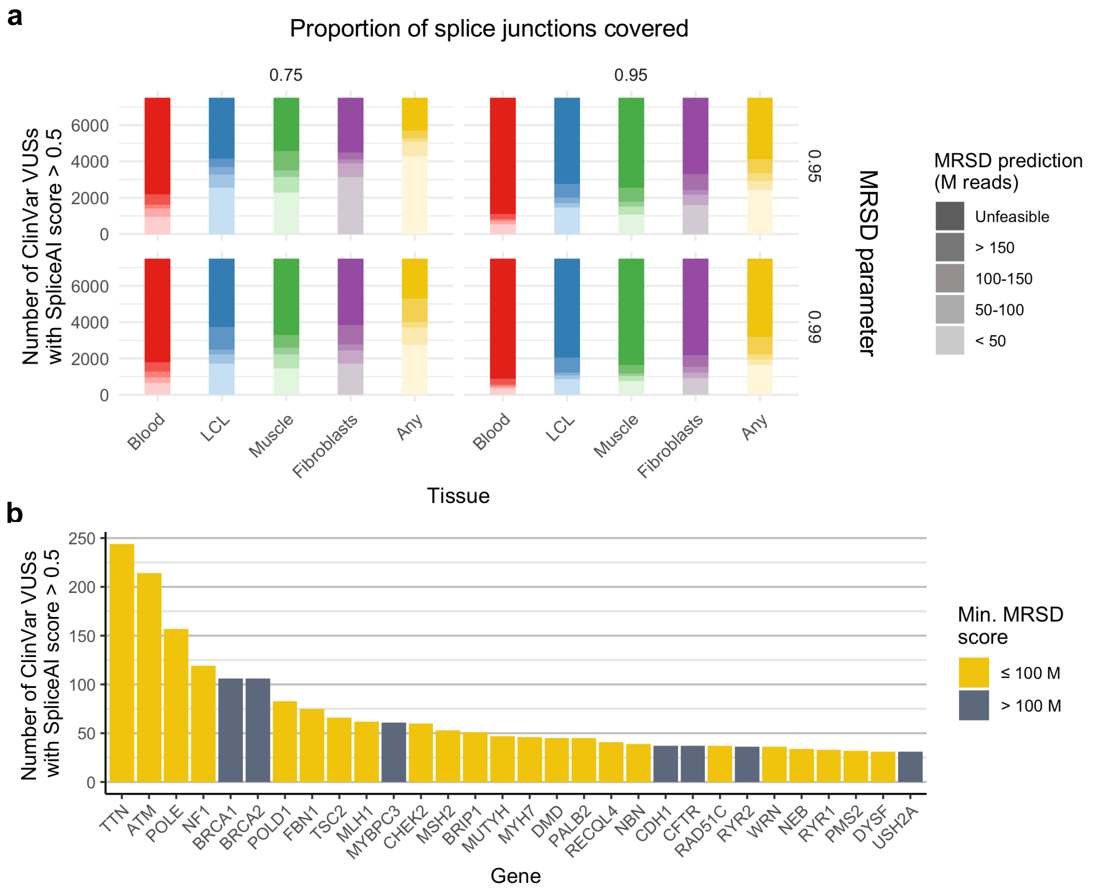


Figure 23. *The scope for resolution of variants of uncertain significance (VUSs) using RNA-seq-based analysis.* MRSD scores were derived for the genes harbouring VUSs present in ClinVar if the variants were predicted by the predictive tool SpliceAI to impact splicing (score ≥ 0.5; Jaganathan et al., 2019) (**a**) Depending on the stringency of the MRSD model parameters, between 25.8% (1940/7507) and 67.8% (5086/7507) of variants predicted to impact splicing are expected to be adequately covered by 100 M uniquely mapping reads or fewer in at least one of the four tissues (whole blood, LCLs, skeletal muscle and fibroblasts). Variants were most likely to be found to be in low-MRSD genes (MRSD ≤ 100 M) in fibroblasts, irrespective of model parameters. (**b**) Among the 30 genes with the greatest number of predicted splice-impacting VUSs, 23 were predicted to be adequately covered (using default parameters) with 100 M uniquely mapping reads or fewer in at least one of the four tissues. An 8-read junction support parameter was used throughout.

MRSD category (**Supplementary Figure S13**). However, when specifying a deeper level of read coverage (20 reads), only 18 (60%) of the top 30 genes remain low-MRSD. This includes increases in MRSD such that three of the four genes with the greatest number of predicted splice-impacting VUSs have MRSDs above 100 M reads. Regardless, the guided application of RNA-seq to functionally investigate the splicing impact of VUSs holds promise to improve diagnostic yield.

4.5. Discussion

The recent development of machine learning approaches has underpinned improvements to the prioritization of variants that impact splicing and cause rare disease (Rowlands et al., 2019). Despite these advances, corroboration of the effect of such variants remains a major obstacle to improving diagnostic yield for Mendelian disorders. This obstacle is amplified by the unexpected functional impact of some variants on splicing, which may change the way the variant is classified in accordance with current guidelines (Rowlands et al., 2020). The MRSD-based approach described here allows the informed selection of biosample(s) for bulk RNA-seq, based on the required number of sequencing reads that need to be generated for appropriate surveillance of genes of interest. This approach enables the effective identification of patients, disease groups and genomic variants that are amenable to functional assessment of mis-splicing through RNA-seq, and may help to improve the efficiency and accuracy of genomic diagnostic approaches.

The primary purpose of MRSD is to predict the likelihood of observing pathogenic splicing defects in a given transcript and tissue, and we quantify the utility of four distinct biosamples in this manner for known monogenic disease genes (**Figure 21**). Through this analysis, we are able to highlight biosamples that may be most informative for RNA-seq based analysis datasets for specific disease subsets. Although our model is conservative (**Figure 17**), we demonstrate through MRSD-guided re-inspection of VUSs in ClinVar that it may be possible to use RNA-seq to clarify the effect of up to 2.4% of variants of uncertain significance (**Figure 23a**).

Other approaches to select genes amenable to functional analysis through RNA-seq include leveraging relative gene expression metrics (Frésard et al., 2019, Murdock et al., 2021), or tools which assess the similarity of transcript isoforms between tissues, e.g. MAGIQ-CAT (Aicher et al., 2020). We show that, whilst TPM values are well correlated with MRSD scores (**Figure 19a-c**), uneven sequencing coverage across the length of the transcript may, in some cases, falsely identify specific genes or splice junctions as being amenable to RNA-seq-based analysis (**Supplementary Figure S9**). 3' sequencing bias, which is a known artefact of poly-A enriched mRNA sequencing (Finotello et al., 2014, Nagalakshmi et al., 2008, Wang et al., 2009), may elevate the risk of inaccurately selecting genes that could be surveyed through RNA-seq when considering TPM alone. Additionally, the normalization against sequencing depth that occurs during the calculation of TPM obscures information about raw read count, which is important when analyzing the utility of RNA-seq for clinical diagnostics. MRSD scoring, conversely, leverages variation in sample read depth to provide quantitative predictions about optimal sequencing depths.

Some novel bioinformatics tools may complement the utility of MRSD. The aforementioned tool MAGIQ-CAT (Aicher et al., 2020), for instance, assesses the degree to which transcript isoforms in a sampled tissue accurately resemble those in the primary disease-affected tissue. However, MAGIQ-CAT primarily captures the degree of similarity between isoform structure and does not aim to provide a quantitative readout to guide the diagnostic route. Thus, a proxy tissue may be described as suitable for RNA-seq-based analysis despite having poor coverage of splice junctions. We envision that the use of both MAGIQ-CAT and MRSD could comprehensively capture information about the utility of RNA-seq, both in terms of similarity of isoform structure relative to the disease-affected tissue and in terms of the likelihood of observing disruptions to this structure.

There are several limitations of the current MRSD model, which could be incorporated into future work. Firstly, the MRSD model cannot directly be extended to predict the suitability of datasets to detect allele-specific expression biases and differential gene expression, which have been demonstrated to be evidence of pathogenic mechanisms in known disease-causing genes (Kremer et al., 2017, Byron et al., 2016, Frésard et al., 2019, Kukurba et al., 2014). Although further investigations are required to quantify and prove this suitability, it is likely that genes with low MRSD scores are also amenable to investigations of differential gene expression and isoform imbalance.

Secondly, further extensions to the model could incorporate genomic background which influences gene expression profiles. For example, interferonopathies are a class of genomic immune disorders (Rodero and Crow, 2016, Volpi et al., 2016) that are characterized by the aberrant upregulation of large numbers of transcripts belonging to so-called “interferon-stimulated genes” (Rodero and Crow, 2016, Schneider et al., 2014). As a result of these wide-ranging impacts on their transcriptomes, MRSD predictions, which ostensibly represent the “normal” transcriptomic landscape, may not accurately reflect the degree of sequencing coverage for certain transcripts in patients with interferonopathies, or indeed other disease groups where disrupted expression of many transcripts is characteristic, such as disorders where chromatin structure (Bélanger et al., 2018, Liu et al., 2009) or the function of the spliceosome (Wood et al., 2019, Wood et al., 2020, Buskin et al., 2018) is disrupted. Moreover, the current MRSD model does not explicitly account for the presence of expression quantitative trait loci (eQTLs) or splicing quantitative trait loci (sQTLs) which are known to influence gene expression profiles (Richards et al., 2012, Takata et al., 2017, Westra and Franke, 2014). We have demonstrated that modulation in expression levels may disrupt our ability to reliably highlight pathogenic splicing events (**Figure 21c**). As a greater number of paired transcriptome and genomic datasets become available, we expect that MRSD scores can be generated in a dynamic manner to account for the presence of eQTLs, sQTLs or other modifiers of gene expression profiles.

Thirdly, our approach is built for a specific cohort of RNA-seq-based analyses; namely, the analysis of a selection of tissues by bulk short-read poly-A enrichment RNA-seq, processed using a specific bioinformatics analysis pipeline (Cummings et al., 2017). This experimental RNA-seq approach currently remains widespread (Cummings et al., 2017, Frésard et al., 2019, Lee et al., 2020); however, our model may be readily applicable to RNA-seq generated using alternative methodologies, such as increased read length, with only minor variations in model performance (**Supplementary Figure 8**). As other technologies, such as long-read (Mantere et al., 2019, Merker et al., 2018, Pauper et al., 2020), single-cell (Del-Aguila et al., 2019, Nomura, 2021) and spatially resolved RNA-seq (Crosetto et al., 2015, Larsson et al., 2021, Marx, 2021, Navarro et al., 2020), become more prevalent in a clinical setting, appropriate control datasets must be generated to develop corresponding MRSD models. Similarly, recent research has shown noticeable improvements to diagnostic yield for neuromuscular disorders by conducting RNA-seq on *in vitro* myofibrils generated by a fibroblast-to-myofibril transdifferentiation protocol (Gonorazky et al., 2019). Such patient-derived cell line approaches represent a promising avenue to scrutinize transcripts not otherwise observable in proxy tissues (Wood et al., 2020, Lin et al., 2011). As these protocols gain wider use, generation of control RNA-seq data from healthy individuals using these approaches will be vital both to allow the generation of MRSD scores and to accurately assess pathogenicity of any identified mis-splicing events.

In summary, the novel MRSD model presented here offers a gene-specific readout to predict the most suitable biosample for interrogation of splicing disruption at the transcript level. This may uncover previously unintuitive choices of biosample, as discussed above in the case of familial rhabdomyosarcoma (**Figure 22c;** see *4.4.6.*). The use of different biosamples is associated with different costs: while whole blood is routinely taken in the clinic, cell-based RNA-seq requires harvesting and culturing of patient cells, and muscle biopsy is an invasive procedure that is generally only undertaken if deemed necessary. Our tool may allow clinical staff to make informed decisions about the likely cost-benefit balance of RNA-seq analysis to ensure such costs are not incurred unnecessarily. We expect that the use of MRSD will allow effective and appropriate integration of RNA-seq into diagnostic genomic services, and ultimately improve variant interpretation and diagnostic yield.

Chapter 5

Discussion

5. Discussion

5.1. Summary of aims and key findings

The clinical analysis of variants impacting splicing is hindered by numerous challenges at the levels of identification, prioritization and functional analysis. The work described here aimed to develop and apply novel bioinformatics analyses to address these challenges, and offer potential avenues to facilitate a more comprehensive approach to splice variant analysis.

Despite the development of several machine learning-based tools for splice prediction in recent years, the clinical interpretation of splice-impacting variants still relies in many cases on the use of older tools that may not provide the most accurate predictions of splice variant impact. The first aim of this work was thus to facilitate the integration of recent bioinformatics tools for the prediction of mis-splicing. Through comparison of seven of these tools, as well as a consensus-based approach, against a large set of clinical variants functionally validated for splice impact, I identified the use of SpliceAI as the most effective individual strategy for splice variant prioritisation. However, through the design of a novel approach that weighted individual predictions by the maximum predictive score possible for the respective tool, it was possible to achieve a small improvement on the performance of any one individual model.

Through the application of SpliceAI scores to a cohort of retinal dystrophy gene panel-derived intronic variants, I was able to identify 758 intronic variants that were predicted to impact splicing, including 379 cases in which the respective variant had not yet been noted on an existing diagnostic report (or in which the variant was identified but splice impact not considered). These constituted a potentially new set of pathogenic variants for clinical consideration.

Despite the efficacy of SpliceAI in predicting splice impact among our selected cohort of variants, it remained to be seen whether this efficacy would hold equally across different types of splice variant. When applying SpliceAI scores to 30 known pathogenic branchpoint-impacting variants, I was able to demonstrate that the SpliceAI splice impact threshold of 0.2 recommended by the tool's creators would have successfully predicted only 50% of the true positive variants as being splice-impacting.

My second aim was therefore to develop a novel bioinformatics pipeline to uniquely predict the likelihood of a variant potentially impacting a branchpoint residue. To this end, I designed a meta-analytical tool that, for a given variant, cross-referenced an empirically derived branchpoint dataset and the predictions of two existing branch point predictive models to highlight variants at potentially critical residues within the branchpoint sequence. Applying this to a cohort of intronic retinal dystrophy gene panel variants revealed a set of ten putatively branchpoint-impacting variants for which zygosity and genotype-phenotype correlation supported pathogenicity, including one variant in the *BBS1* gene that was shown through downstream functional analysis to be pathogenic and causative (in a compound heterozygous state) of the patient phenotype of retinitis pigmentosa.

While predictive models, such as those described in Chapters 2 & 3, serve a valuable function as tools for variant prioritisation, functional evidence is nonetheless required for pathogenicity to be reliably assigned to a variant. One of the most promising technologies to do so is direct analysis of the transcriptome through RNA-seq; however, integration of RNA-seq requires consideration, among other factors, of the tissue-specific expression of transcripts, which limits the set of genes which can be consistently surveyed for splicing dysfunction in clinically derived biosamples. This formed the basis of the final aim of the project: to devise a novel predictive pipeline to inform clinical staff of the likelihood of being able to investigate mis-splicing of a given transcript or gene in a given tissue.

I thus developed the minimum required sequencing depth (MRSD) metric, based on control transcriptomic data from the GTEx project, across four different clinical biosamples of interest. I demonstrated the high positive and negative predictive values of the model, and showed the ability of the metric to capture the uneven nature of splice junction coverage across the lengths of specific transcripts, in a way that is not captured by conventional “total expression” metrics.

Using our selected bioinformatics pipeline for identification of mis-splicing events in RNA-seq data (Cummings et al., 2017), we quantified the characteristics of a selection of known pathogenic mis-splicing events. From this, a minimum parameter set was inferred consisting of 8 reads covering 75% of splice junctions (and with an MRSD parameter of 0.95), which was sufficient to identify 90% of the cohort of investigated pathogenic events.

Finally, I applied these scores to real sets of disease-associated genes and clinical variants of uncertain significance. When considering the numbers of low-MRSD (MRSD ≤100 M reads) genes present in panels listed in the Genomics England PanelApp repository, we observed that fibroblasts were the optimum choice (out of four selected tissues) for 63.1% of gene panels. Applying MRSD scores to predicted splice-impacting variants of uncertain significance in ClinVar showed that, depending on model stringency, the impact of 25.8-67.8% of these variants may be investigated using RNA-seq in one or more of the investigated tissues.

In summary, this work has provided new insights into both the prediction of splice impact, and its investigation at a functional level. However, there are nonetheless several aspects of splicing biology and research that may have implications for the findings presented here. While potentially challenging the results described here, these may also serve as the starting points for promising new avenues of research.

5.2. Clinical bioinformatics in a rapidly evolving computational landscape

Throughout the work conducted by others and described here, bioinformatics tools have proven an invaluable asset in both the prediction and confirmation of the splicing impact of variants. Although our analysis of splicing prediction tools has highlighted SpliceAI as perhaps the most effective predictor of splicing impact, even high-performing models such as SpliceAI may exhibit biases and weakness, as demonstrated by its seeming decrease in power when identifying known pathogenic branchpoint variants (see *3.4.1.*). Similarly, a recent study investigating three distinct cohorts of splice donor GT>GC variants – a class of canonical splice variant that does *not* invariably impact splicing – showed that, while SpliceAI was by some margin the best performing tool in predicting GT>GC splice impact, it still had a ROC-AUC of just 0.79 (Chen et al., 2020). Together, these results suggest that certain subtypes of variants may be inherently more difficult for current models to assess accurately, and consequently that negative splicing predictions for these variants should not necessarily be blindly trusted without further investigation. Additionally, the continually expanding repertoire of predictive tools available to researchers means that comparative analyses of such variant subtypes should be repeated periodically to incorporate the latest developments.

Since conducting the work in Chapter 2, for instance, several major splicing prediction tools have been released, including CADD-Splice (Rentzsch et al., 2021), an extension of the CADD framework specifically tailored to prediction of splicing variants, and Super Quick Information-content Random-forest Learning of Splice variants (SQUIRLS; Danis et al., 2021). In the paper describing CADD-Splice, a similar analysis to that detailed in Chapter 2 was conducted to compare the relative performances of existing machine learning models and revealed that a composite model reminiscent of our approach incorporating predictions from both MMSplice and SpliceAI into the CADD framework substantially improved upon the poor ability of CADD to predict splicing impact.

A similar trend has emerged for the field of branchpoint prediction, in which, as mentioned above (see *3.5.*), the development of novel tools like LaBranchoR (Paggi and Bejerano, 2017) and Branchpointer (Signal et al., 2016) have merited comprehensive comparative testing. While the results of this testing (Leman et al., 2020) have suggested that BPP – one of our selected tools – is the optimal choice for clinical use, it may still be subject to biases that can be compensated for by other tools or approaches, in much the same way that SpliceAI, despite its high accuracy, has difficulty in correctly identifying branchpoint variants impacting splicing. As with splice prediction tools, iterative conduction of comparative studies as subsequent tools are released will be crucial to ensure that any BP prediction tools integrated into clinical practice are the most effective choice.

5.3. Tissue-specific transcripts in clinical analysis of mis-splicing

Alongside differential gene expression, a key hallmark of tissue specificity at the genetic level (alongside differential gene expression) is the existence of diverse transcript species that may vary in proportion between tissues, or may even be completely unique to a single tissue. Such transcripts may allow diversification of function or regulation for a single gene, and so hold substantial biological importance. In the context of clinical examinations of splicing, however, this rich diversity may confound diagnostics at multiple levels.

When considering predictive approaches to splicing analysis, for example, it may be necessary to ensure that tissue-specific splicing features are accounted for in predictive models if greater transcript diversity is known to be associated with a particular gene(s) and/or tissue(s) of interest. Where the use of tissue-associated splicing features is governed by a small number of highly specific factors, this is particularly pertinent; in humans, for instance, the splicing factor *SRRM4* is involved in the inclusion of microexons across a number of neural cell types (Calarco et al., 2009; Irimia et al., 2014), and its family member *SRRM3* is further required for inclusion of such microexons in the retina for around 75 genes (Ciampi et al., 2021). As such events only occur in a minority of transcripts when compared to all annotated transcripts, there is a likelihood that many models may be underpowered to detect changes affecting tissue-specific isoforms. Further, where training of a machine learning-based is based on provision of a known transcript model, as in the cases of CryptSplice (Lee et al., 2017), a bias will likely exist in favour of constitutive transcript features, as the supplied model may list tissue-specific transcripts with lower confidence, or omit them entirely, particularly where the supplied transcript models are older.

Taken together, these observations imply that some of the “negative” findings in Chapters 2 may, in fact, have constituted pathogenic variants for which the retina-specific impact was simply not predicted. A further extension to my comparative analysis that simulates likely disruptive changes (for example, canonical splice site variants) to tissue-specific isoforms may reveal whether, and to what extent, an underpower exists in current predictive models.

Some predictive tools have indeed begun to incorporate tissue-specific considerations. Of note, the tool MTSplice, an extension of the MMSplice framework that integrates a so-called “tissue module” to reflect tissue-specific splicing patterns, including the peripheral retina, was recently described by Cheng et al. (2021). While most surveyed variants did not differentially impact splicing across tissues, the authors did note a handful of highly tissue-specific splicing disruptions that suggested the model accurately recapitulates biology for at least some transcript species. This may prove a useful starting point for re-analysis of the splice impact of our retinal dystrophy variant cohort.

Tissue-specific transcript diversity may also have significant implications for the design of gene panels and clinical exomes, which risk overlooking the capture of biologically important sequences if not updated to included novel features identified in successive transcript annotations. A limit to this process may be the predominance of bulk RNA-seq as a tool to capture transcriptome diversity: as most human tissues are highly heterogeneous mixtures of large numbers of cell types, our current understanding of their transcriptomic landscapes may overrepresent the transcript species identified in the most numerous cell types, while underserving those that may exist in smaller number, such as stem cells.

As transcriptomic data becomes publicly available across a wider range of cell types, and with a gradual move away from bulk tissue RNA-seq - perhaps even at the level of single-cell RNA-seq for individual cell types, it may be anticipated that exciting new layers of splicing regulation may be identified. Consequently, there is a need to regularly monitor literature and incorporate novel transcript features in clinically relevant genes into enrichment strategies.

The design of the MRSD approach in Chapter 4 has thus far been based around transcripts selected not so much for their biological relevance as for the confidence in their existence. While suitable for proof of principle, use of our selected transcripts may not necessarily reflect the most clinically relevant transcript. An extension to the MRSD framework and web portal that allows user selection of transcripts of choice will therefore give clinicians the ability to tailor MRSD analyses and predictions to their particular use case based on the relevant disease subtype.

Increasing numbers of lines of study show that tissue-specific transcript features may be sources of pathogenic variation: besides the *DYNC2H1* example described above (see *1.3.5.1.*), the retina-specific ORF15 variant of the *RPGR* gene is already well-established as a prevalent cause of X-linked retinal degeneration (Vervoort et al., 2000), and variants in the retina-specific second exon of the *COL2A1* gene, typically associated with Stickler syndrome, have also been shown to cause a uniquely or predominantly ocular phenotype (Richards et al., 2000). There will doubtless exist such effects in other as-yet-unidentified transcripts, particularly for disease subtypes affecting tissues with high numbers of tissue-specific transcripts, such as cardiac (Zhu et al., 2021) and nervous tissue (Raj and Blencowe, 2015; Ray et al., 2020;). Thus, a directed effort to identify novel transcript features will be an important next step in improving diagnostic yield for several disease subtypes.

5.4. Proxy tissues and the selection of valid functional assays

Our development of MRSD has highlighted the critical importance of selecting tissues with suitable coverage of transcripts of interest. Despite the clinical value we predict MRSD scoring will bring, it does also reveal that RNA-seq of clinically accessible tissues is not sufficient for adequate gene coverage for a large number of disease subtypes. As previously alluded to (see *4.5.*), one of the most intriguing developments is in the transdifferentiation of patient cells to clinically relevant cell types, as conducted by Gonorazky et al. (2019) in the case of fibroblast-derived myofibrils in the study of neuromuscular disorders.

However, this approach, while ostensibly very effective in the case of myofibrils, may have varying efficacy and economic feasibility in other disease and tissue types. One experimental system used to invest retinal biology and pathogenesis is retinal organoids. Generated from patient pluripotent stem cells (hPSCs), retinal organoids recapitulate much of the *in vivo* structure of the retina, including even neurite outgrowths (Fligor et al., 2018). These systems would likely be an incredibly beneficial source of RNA for the investigation of undiagnosed Mendelian retinal dystrophy patients. However, the protocol to culture these organoids is substantially time-consuming, taking around 100 days for a fully-formed organoid to develop from initial culture (Li et al., 2021). Such cultures are also very expensive, both in terms of reagents and the training of personnel to be able to culture them, limiting their clinical applicability

For other cell types, the transdifferentiation and culture process may be yet more time-consuming. In the case of one protocol for the culture of interneuron cells, for example, a protocol of 20-30 weeks is recommended to differentiate hPSCs into fully mature neural cells (Nicholas et al., 2013). However, other transdifferentiation approaches may not be as time-consuming: with the addition of small molecule inhibitors, for example, one study demonstrated the high-efficacy generation of nociceptors from hPSCs in just 10 days (Chambers et al., 2012). Considering the lack of clinically accessible tissues for the investigation of neurodevelopmental (or otherwise neurological) disorders, these may serve a particularly valuable diagnostic role. However, there remains an additional time constraint for the initial generation of the hPSCs, and the growth of sufficient numbers of cells to yield adequate RNA for transcriptomic analysis.

The incorporation of such culture techniques into clinical practice is therefore limited in many cases by their economic and infrastructural feasibility. Further, although they may serve as proxies for tissues of clinical interest, their transcriptomic landscapes may not wholly reflect those of tissues in a native biological context. For MRSD scores to be useful in these contexts, development of culture methodologies must be accompanied by the release of associated RNA-seq datasets, to allow effective generation of control splice junction counts. MRSD may then serve as a useful guiding metric for clinical integration of RNA-seq for these promising transdifferentiation methodologies.

5.5. Reanalysis of existing data

Decreasing sequencing costs have allowed whole genome sequencing to become an increasingly adopted sequencing methodology for Mendelian disease, particularly for panel- or exome-negative cases (Mattick et al., 2018; Palmer et al., 2021). It is notable that the work conducted in Chapters 2 and 3 did not rely on the generation of novel patient datasets: rather, through analysis of the variation in the ~50 bp intronic regions covered by standard gene panel enrichment kits, it was possible to identify substantial numbers of non-coding variants not highlighted through upstream diagnostic analyses, and these were identified in substantial numbers (331/2783; 12.1%) of investigated patients in Chapter 2 (see *2.4.4.*). While the majority of these are unlikely to represent causes of Mendelian disease in the respective patients, my work demonstrated a number of potentially pathogenic variants. In addition, for identified intronic variants classified as likely benign or benign, reporting of these incidental findings is nonetheless important to ensure accuracy of diagnostic reporting.

Further, my work in Chapter 3 demonstrates that, while gene panels may not seem an intuitive diagnostic methodology for investigation of intronic sequence elements, such as branchpoints, there is in fact sufficient intronic coverage to conclusively identify pathogenic intronic variants affecting such elements. It may be expected that this observation holds true for other near-splice site elements, such as the poly-pyrimidine tract and extended 5′ splice site, as well as for other sequencing methodologies, such as exomes, that rely on the targeted enrichment of exonic regions.

5.6. Future prospects for interpretation of pathogenic splicing

The incorporation of the ACMG guidelines (Richards et al., 2015) into routine clinical practice has greatly improved consistency in variant interpretation both within and between diagnostic centres, and has thereby improved the quality of variant reporting returned to patients and their families. As discussed in *1.3.5.1.*, provision in these guidelines for non-canonical splice variants is lacking, and extension of existing guidelines to cover these (or the development of novel and bespoke guidance for them) is likely to refine their interpretation and may lead to new diagnoses.

Beyond the level of the individual variant, however, an intriguing possibility is the development of guidelines for the interpretation of pathogenicity of mis-splicing events themselves. While I have made initial attempts to conduct such an analysis in Chapter 4, the sample numbers in my analysis remain low, and a collaborative pooling of RNA-seq data for patients harbouring known mis-splicing events between centres would allow more robust discernment of the precise characteristics of pathogenic splicing variation. There is a substantial unmet need to provide a platform for data sharing in this regard: a database displaying the reads supporting pathogenic and benign events, as well as accompanying splice metrics, may aid the interpretation of mis-splicing events in a manner similar to that of the variant sharing platform GeneMatcher (Sobreira et al., 2015). Further, comparison of pathogenic splicing characteristics between different RNA-seq analytical methodologies, such as FRASER (Mertes et al., 2021), LeafCutterMD (Jenkinson et al., 2020) and SPOT (Ferraro et al., 2020) will allow prediction of mis-splicing impact to be more consistent across analytical approaches.

The development of such guidelines may prove especially valuable where an interpreted variant is incompletely penetrant: for such variants, we may anticipate from current guidelines that interpretation would yield an identical classification between individuals. With the incorporation of splicing event-level metrics, however, it may be demonstrated that a single variant can result in mis-splicing in one individual, and a lack thereof in another (perhaps due to underlying genetic background), suggesting incomplete penetrance of the variant and resulting in different variant classifications according to the individual in question. Thus, guidelines can begin to encompass inter-patient variation and ensure that returned variant classifications are pertinent to the patient in question.

A likely consideration for the development of mis-splicing guidelines is the necessity of gene-specific regulations: such variant interpretation frameworks, most often based on the ACMG guidelines, have already been developed for genes such as *PTEN* (Mester et al., 2019) and multiple RASopathy genes (Gelb et al., 2018). Concerted efforts to characterise the impact of mis-splicing for individual genes, or individual exons within genes, are lacking in existing literature. It is likely that genes highly constrained against loss of function may require much lower levels of splice disruption to result in disease presentation, while less constrained genes may require all or almost all alleles to harbour pathogenic splicing changes for disease phenotypes to present.

Variably expressive splice variants may also underlie the phenotypic spectrums seen for individual genes, as seen in one study demonstrating the milder osteogenesis imperfecta phenotype observed in patients with less penetrant COL1A1 and COL1A2 splice variants (Li et al., 2019). Gene-specific splicing guidelines may allow severity of phenotype, or the presence/absence of specific phenotypic features, to guide the interpretation of mis-splicing events. Again, the sharing of RNA-seq and/or other functional data will be crucial to shed light on exactly what constitutes a pathogenic mis-splicing event for a given gene.

5.7. Concluding remarks

Splicing variation has long proven difficult to identify and interpret. However, the landscape of clinical bioinformatics is now well-placed to begin to unpick the complexities of mis-splicing in disease contexts. In this work, I have demonstrated the importance of re-analysis of existing genomic data through the lens of mis-splicing, and the value of developing novel approaches to account for shortfalls in the performance of bioinformatics models.

With the increasing diversity in computational and experimental frameworks to examine splicing impact, as well as the increasing sharing of patient data, a more holistic approach to prediction and interpretation of splice impact is fast becoming possible. This will serve only to improve diagnostic yield and allow a greater number of Mendelian disease patients to receive a molecular diagnosis that may improve diagnosis, management and, ultimately, treatment of their disorders.

References

ABDRABO, L. S., WATKINS, D., WANG, S. R., LAFOND-LAPALME, J., RIVIERE, J. B. & ROSENBLATT, D. S. 2020. Genome and RNA sequencing in patients with methylmalonic aciduria of unknown cause. *Genet Med,* 22**,** 432-436.

ADAMSON, S. I., ZHAN, L. & GRAVELEY, B. R. 2018. Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biol,* 19**,** 71.

ADZHUBEI, I. A., SCHMIDT, S., PESHKIN, L., RAMENSKY, V. E., GERASIMOVA, A., BORK, P., KONDRASHOV, A. S. & SUNYAEV, S. R. 2010. A method and server for predicting damaging missense mutations. *Nat Methods,* 7**,** 248-9.

AGRAWAL, S., PILARSKI, R. & ENG, C. 2005. Different splicing defects lead to differential effects downstream of the lipid and protein phosphatase activities of PTEN. *Hum Mol Genet,* 14**,** 2459-68.

AICHER, J. K., JEWELL, P., VAQUERO-GARCIA, J., BARASH, Y. & BHOJ, E. J. 2020. Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. *Genet Med,* 22**,** 1181-1190.

AL HAFID, N. & CHRISTODOULOU, J. 2015. Phenylketonuria: a review of current and future treatments. *Transl Pediatr,* 4**,** 304-17.

ALEXANDROVA, E. A., OLOVNIKOV, I. A., MALAKHOVA, G. V., ZABOLOTNEVA, A. A., SUNTSOVA, M. V., DMITRIEV, S. E. & BUZDIN, A. A. 2012. Sense transcripts originated from an internal part of the human retrotransposon LINE-1 5' UTR. *Gene,* 511**,** 46-53.

ALIOTO, T. S. 2007. U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res,* 35**,** D110-5.

ANDERSON, S., BANKIER, A. T., BARRELL, B. G., DE BRUIJN, M. H., COULSON, A. R., DROUIN, J., EPERON, I. C., NIERLICH, D. P., ROE, B. A., SANGER, F., SCHREIER, P. H., SMITH, A. J., STADEN, R. & YOUNG, I. G. 1981. Sequence and organization of the human mitochondrial genome. *Nature,* 290**,** 457-65.

ANDERSSON, R. & SANDELIN, A. 2020. Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet,* 21**,** 71-87.

ANNA, A. & MONIKA, G. 2018. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genet,* 59**,** 253-268.

ATEN, E., SUN, Y., ALMOMANI, R., SANTEN, G. W., MESSEMAKER, T., MAAS, S. M., BREUNING, M. H. & DEN DUNNEN, J. T. 2013. Exome sequencing identifies a branch point variant in Aarskog-Scott syndrome. *Hum Mutat,* 34**,** 430-4.

ATWAL, P. S., BRENNAN, M. L., COX, R., NIAKI, M., PLATT, J., HOMEYER, M., KWAN, A., PARKIN, S., SCHELLEY, S., SLATTERY, L., WILNAI, Y., BERNSTEIN, J. A., ENNS, G. M. & HUDGINS, L. 2014. Clinical whole-exome sequencing: are we there yet? *Genet Med,* 16**,** 717-9.

AUTON, A., BROOKS, L. D., DURBIN, R. M., GARRISON, E. P., KANG, H. M., KORBEL, J. O., MARCHINI, J. L., MCCARTHY, S., MCVEAN, G. A., ABECASIS, G. R. & CONSORTIUM, G. P. 2015. A global reference for human genetic variation. *Nature,* 526**,** 68-74.

AVSEC, Ž., KREUZHUBER, R., ISRAELI, J., XU, N., CHENG, J., SHRIKUMAR, A., BANERJEE, A., KIM, D. S., BEIER, T., URBAN, L., KUNDAJE, A., STEGLE, O. & GAGNEUR, J. 2019. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat Biotechnol,* 37**,** 592-600.

BAUWENS, M., GARANTO, A., SANGERMANO, R., NAESSENS, S., WEISSCHUH, N., DE ZAEYTIJD, J., KHAN, M., SADLER, F., BALIKOVA, I., VAN CAUWENBERGH, C., ROSSEEL, T., BAUWENS, J., DE LEENEER, K., DE JAEGERE, S., VAN LAETHEM, T., DE VRIES, M., CARSS, K., ARNO, G., FAKIN, A., WEBSTER, A. R., DE RAVEL DE L'ARGENTIERE, T. J. L., SZNAJER, Y., VUYLSTEKE, M., KOHL, S., WISSINGER, B., CHERRY, T., COLLIN, R. W. J., CREMERS, F. P. M., LEROY, B. P. & DE BAERE, E. 2019. ABCA4-associated disease as a model for missing heritability in autosomal recessive disorders: novel noncoding splice, cis-regulatory, structural, and recurrent hypomorphic variants. *Genet Med,* 21**,** 1761-1771.

BEECROFT, S. J., YAU, K. S., ALLCOCK, R. J. N., MINA, K., GOODING, R., FAIZ, F., ATKINSON, V. J., WISE, C., SIVADORAI, P., TRAJANOSKI, D., KRESOJE, N., ONG, R., DUFF, R. M., CABRERA-SERRANO, M., NOWAK, K. J., PACHTER, N., RAVENSCROFT, G., LAMONT, P. J., DAVIS, M. R. & LAING, N. G. 2020. Targeted gene panel use in 2249 neuromuscular patients: the Australasian referral center experience. *Ann Clin Transl Neurol,* 7**,** 353-362.

BÉLANGER, C., BÉRUBÉ-SIMARD, F. A., LEDUC, E., BERNAS, G., CAMPEAU, P. M., LALANI, S. R., MARTIN, D. M., BIELAS, S., MOCCIA, A., SRIVASTAVA, A., SILVERSIDES, D. W. & PILON, N. 2018. Dysregulation of cotranscriptional alternative splicing underlies CHARGE syndrome. *Proc Natl Acad Sci U S A,* 115**,** E620-E629.

BEYTER, D., INGIMUNDARDOTTIR, H., ODDSSON, A., EGGERTSSON, H. P., BJORNSSON, E., JONSSON, H., ATLASON, B. A., KRISTMUNDSDOTTIR, S., MEHRINGER, S., HARDARSON, M. T., GUDJONSSON, S. A., MAGNUSDOTTIR, D. N., JONASDOTTIR, A., KRISTJANSSON, R. P., SVERRISSON, S. T., HOLLEY, G., PALSSON, G., STEFANSSON, O. A., EYJOLFSSON, G., OLAFSSON, I., SIGURDARDOTTIR, O., TORFASON, B., MASSON, G., HELGASON, A., THORSTEINSDOTTIR, U., HOLM, H., GUDBJARTSSON, D. F., SULEM, P., MAGNUSSON, O. T., HALLDORSSON, B. V. & STEFANSSON, K. 2021. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet,* 53**,** 779-786.

BHATIA, S., BENGANI, H., FISH, M., BROWN, A., DIVIZIA, M. T., DE MARCO, R., DAMANTE, G., GRAINGER, R., VAN HEYNINGEN, V. & KLEINJAN, D. A. 2013. Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia. *Am J Hum Genet,* 93**,** 1126-34.

BINDER, J., HOFMANN, S., KREISEL, S., WÖHRLE, J. C., BÄZNER, H., KRAUSS, J. K., HENNERICI, M. G. & BAUER, M. F. 2003. Clinical and molecular findings in a patient with a novel mutation in the deafness-dystonia peptide (DDP1) gene. *Brain,* 126**,** 1814-20.

BISHOP, D. F., SCHNEIDER-YIN, X., CLAVERO, S., YOO, H. W., MINDER, E. I. & DESNICK, R. J. 2010. Congenital erythropoietic porphyria: a novel uroporphyrinogen III synthase branchpoint mutation reveals underlying wild-type alternatively spliced transcripts. *Blood,* 115**,** 1062-9.

BOSCH, A. M., IJLST, L., OOSTHEIM, W., MULDERS, J., BAKKER, H. D., WIJBURG, F. A., WANDERS, R. J. & WATERHAM, H. R. 2005. Identification of novel mutations in classical galactosemia. *Hum Mutat,* 25**,** 502.

BOURGEOIS, P., BOLCATO-BELLEMIN, A. L., DANSE, J. M., BLOCH-ZUPAN, A., YOSHIBA, K., STOETZEL, C. & PERRIN-SCHMITT, F. 1998. The variable expressivity and incomplete penetrance of the twist-null heterozygous mouse phenotype resemble those of human Saethre-Chotzen syndrome. *Hum Mol Genet,* 7**,** 945-57.

BOYLE, A. P., DAVIS, S., SHULHA, H. P., MELTZER, P., MARGULIES, E. H., WENG, Z., FUREY, T. S. & CRAWFORD, G. E. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell,* 132**,** 311-22.

BRETSCHNEIDER, H., GANDHI, S., DESHWAR, A. G., ZUBERI, K. & FREY, B. J. 2018. COSSMO: predicting competitive alternative splice site selection using deep learning. *Bioinformatics,* 34**,** i429-i437.

BRYEN, S. J., JOSHI, H., EVESSON, F. J., GIRARD, C., GHAOUI, R., WADDELL, L. B., TESTA, A. C., CUMMINGS, B., ARBUCKLE, S., GRAF, N., WEBSTER, R., MACARTHUR, D. G., LAING, N. G., DAVIS, M. R., LÜHRMANN, R. & COOPER, S. T. 2019. Pathogenic Abnormal Splicing Due to Intronic Deletions that Induce Biophysical Space Constraint for Spliceosome Assembly. *Am J Hum Genet,* 105**,** 573-587.

BURROWS, N. P., NICHOLLS, A. C., RICHARDS, A. J., LUCCARINI, C., HARRISON, J. B., YATES, J. R. & POPE, F. M. 1998. A point mutation in an intronic branch site results in aberrant splicing of COL5A1 and in Ehlers-Danlos syndrome type II in two British families. *Am J Hum Genet,* 63**,** 390-8.

BURSET, M., SELEDTSOV, I. A. & SOLOVYEV, V. V. 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res,* 28**,** 4364-75.

BUSKIN, A., ZHU, L., CHICHAGOVA, V., BASU, B., MOZAFFARI-JOVIN, S., DOLAN, D., DROOP, A., COLLIN, J., BRONSTEIN, R., MEHROTRA, S., FARKAS, M., HILGEN, G., WHITE, K., PAN, K. T., TREUMANN, A., HALLAM, D., BIALAS, K., CHUNG, G., MELLOUGH, C., DING, Y., KRASNOGOR, N., PRZYBORSKI, S., ZWOLINSKI, S., AL-AAMA, J., ALHARTHI, S., XU, Y., WHEWAY, G., SZYMANSKA, K., MCKIBBIN, M., INGLEHEARN, C. F., ELLIOTT, D. J., LINDSAY, S., ALI, R. R., STEEL, D. H., ARMSTRONG, L., SERNAGOR, E., URLAUB, H., PIERCE, E., LÜHRMANN, R., GRELLSCHEID, S. N., JOHNSON, C. A. & LAKO, M. 2018. Disrupted alternative splicing for genes implicated in splicing and ciliogenesis causes PRPF31 retinitis pigmentosa. *Nat Commun,* 9**,** 4234.

BYRON, S. A., VAN KEUREN-JENSEN, K. R., ENGELTHALER, D. M., CARPTEN, J. D. & CRAIG, D. W. 2016. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet,* 17**,** 257-71.

CALARCO, J. A., SUPERINA, S., O'HANLON, D., GABUT, M., RAJ, B., PAN, Q., SKALSKA, U., CLARKE, L., GELINAS, D., VAN DER KOOY, D., ZHEN, M., CIRUNA, B. & BLENCOWE, B. J. 2009. Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell,* 138**,** 898-910.

CALVO, S. E., PAGLIARINI, D. J. & MOOTHA, V. K. 2009. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci U S A,* 106**,** 7507-12.

CARTEGNI, L., WANG, J., ZHU, Z., ZHANG, M. Q. & KRAINER, A. R. 2003. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res,* 31**,** 3568-71.

CASCINO, I., FIUCCI, G., PAPOFF, G. & RUBERTI, G. 1995. Three functional soluble forms of the human apoptosis-inducing Fas molecule are produced by alternative splicing. *J Immunol,* 154**,** 2706-13.

CASTEL, S. E., AGUET, F., MOHAMMADI, P., CONSORTIUM, G. T., ARDLIE, K. G. & LAPPALAINEN, T. 2020. A vast resource of allelic expression data spanning human tissues. *Genome Biol,* 21**,** 234.

CASTEL, S. E., CERVERA, A., MOHAMMADI, P., AGUET, F., REVERTER, F., WOLMAN, A., GUIGO, R., IOSSIFOV, I., VASILEVA, A. & LAPPALAINEN, T. 2018. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat Genet,* 50**,** 1327-1334.

CAVALOC, Y., BOURGEOIS, C. F., KISTER, L. & STÉVENIN, J. 1999. The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *RNA,* 5**,** 468-83.

CHAMBERS, S. M., QI, Y., MICA, Y., LEE, G., ZHANG, X. J., NIU, L., BILSLAND, J., CAO, L., STEVENS, E., WHITING, P., SHI, S. H. & STUDER, L. 2012. Combined small-molecule inhibition accelerates developmental timing and converts human pluripotent stem cells into nociceptors. *Nat Biotechnol,* 30**,** 715-20.

CHAOUCH, L., SELLAMI, H., KALAI, M., DARRAGI, I., BOUDRIGUA, I., CHAOUACHI, D., ABBES, S. & MNIF, S. 2020. New Deletion at Promoter of HBG1 Gene in Sickle Cell Disease Patients With High HbF Level. *J Pediatr Hematol Oncol,* 42**,** 20-22.

CHAVANAS, S., GACHE, Y., VAILLY, J., KANITAKIS, J., PULKKINEN, L., UITTO, J., ORTONNE, J. & MENEGUZZI, G. 1999. Splicing modulation of integrin beta4 pre-mRNA carrying a branch point mutation underlies epidermolysis bullosa with pyloric atresia undergoing spontaneous amelioration with ageing. *Hum Mol Genet,* 8**,** 2097-105.

CHEN, J. M., LIN, J. H., MASSON, E., LIAO, Z., FEREC, C., COOPER, D. N. & HAYDEN, M. 2020. The Experimentally Obtained Functional Impact Assessments of 5' Splice Site GT'GC Variants Differ Markedly from Those Predicted. *Curr Genomics,* 21**,** 56-66.

CHEN, J., LI, X., ZHONG, H., MENG, Y. & DU, H. 2019. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Sci Rep,* 9**,** 9345.

CHEN, X., SCHULZ-TRIEGLAFF, O., SHAW, R., BARNES, B., SCHLESINGER, F., KALLBERG, M., COX, A. J., KRUGLYAK, S. & SAUNDERS, C. T. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics,* 32**,** 1220-2.

CHENG, C., YAFFE, M. B. & SHARP, P. A. 2006. A positive feedback loop couples Ras activation and CD44 alternative splicing. *Genes Dev,* 20**,** 1715-20.

CHENG, J., CELIK, M. H., KUNDAJE, A. & GAGNEUR, J. 2021. MTSplice predicts effects of genetic variants on tissue-specific splicing. *Genome Biol,* 22**,** 94.

CHENG, J., NGUYEN, T. Y. D., CYGAN, K. J., CELIK, M. H., FAIRBROTHER, W. G., AVSEC, Z. & GAGNEUR, J. 2019. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol,* 20**,** 48.

CHIANG, J. P., LAMEY, T., MCLAREN, T., THOMPSON, J. A., MONTGOMERY, H. & DE ROACH, J. 2015. Progress and prospects of next-generation sequencing testing for inherited retinal dystrophy. *Expert Rev Mol Diagn,* 15**,** 1269-75.

CHO, S., HOANG, A., SINHA, R., ZHONG, X. Y., FU, X. D., KRAINER, A. R. & GHOSH, G. 2011. Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. *Proc Natl Acad Sci U S A,* 108**,** 8233-8.

CIAMPI, L., MANTICA, F., LOPEZ-BLANCH, L., RODRÍGUEZ-MARIN, C., CIANFERONI, D., ZANG, J., PERMANYER, J., JIMÉNEZ-DELGADO, S., BONNAL, S., MIRAVET-VERDE, S., RUPRECHT, V., NEUHAUSS, S. C. F., BANFI, S., CARRELLA, S., SERRANO, L., HEAD, S. A. & IRIMIA, M. 2021. Specialization of the photoreceptor transcriptome by Srrm3-dependent microexons is required for outer segment maintenance and vision. *bioRxiv***,** 2021.09.08.459463.

CINGOLANI, P., PLATTS, A., WANG, L. L., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. & RUDEN, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin),* 6**,** 80-92.

COOLIDGE, C. J., SEELY, R. J. & PATTON, J. G. 1997. Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res,* 25**,** 888-96.

COOPER, G. M., STONE, E. A., ASIMENOS, G., GREEN, E. D., BATZOGLOU, S., SIDOW, A. & PROGRAM, N. C. S. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res,* 15**,** 901-13.

CORTÉS-CIRIANO, I., LEE, J. J., XI, R., JAIN, D., JUNG, Y. L., YANG, L., GORDENIN, D., KLIMCZAK, L. J., ZHANG, C. Z., PELLMAN, D. S., PARK, P. J., GROUP, P. S. V. W. & CONSORTIUM, P. 2020. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet,* 52**,** 331-341.

CORVELO, A., HALLEGGER, M., SMITH, C. W. & EYRAS, E. 2010. Genome-wide association between branch point properties and alternative splicing. *PLoS Comput Biol,* 6**,** e1001016.

CROSETTO, N., BIENKO, M. & VAN OUDENAARDEN, A. 2015. Spatially resolved transcriptomics and beyond. *Nat Rev Genet,* 16**,** 57-66.

CROTTI, L., LEWANDOWSKA, M. A., SCHWARTZ, P. J., INSOLIA, R., PEDRAZZINI, M., BUSSANI, E., DAGRADI, F., GEORGE, A. L. & PAGANI, F. 2009. A KCNH2 branch point mutation causing aberrant splicing contributes to an explanation of genotype-negative long QT syndrome. *Heart Rhythm,* 6**,** 212-8.

CUMMINGS, B. B., KARCZEWSKI, K. J., KOSMICKI, J. A., SEABY, E. G., WATTS, N. A., SINGER-BERK, M., MUDGE, J. M., KARJALAINEN, J., SATTERSTROM, F. K., O'DONNELL-LURIA, A. H., POTERBA, T., SEED, C., SOLOMONSON, M., ALFÖLDI, J., DALY, M. J., MACARTHUR, D. G., TEAM, G. A. D. P. & CONSORTIUM, G. A. D. 2020. Transcript expression-aware annotation improves rare variant interpretation. *Nature,* 581**,** 452-458.

CUMMINGS, B. B., MARSHALL, J. L., TUKIAINEN, T., LEK, M., DONKERVOORT, S., FOLEY, A. R., BOLDUC, V., WADDELL, L. B., SANDARADURA, S. A., O'GRADY, G. L., ESTRELLA, E., REDDY, H. M., ZHAO, F., WEISBURD, B., KARCZEWSKI, K. J., O'DONNELL-LURIA, A. H., BIRNBAUM, D., SARKOZY, A., HU, Y., GONORAZKY, H., CLAEYS, K., JOSHI, H., BOURNAZOS, A., OATES, E. C., GHAOUI, R., DAVIS, M. R., LAING, N. G., TOPF, A., KANG, P. B., BEGGS, A. H., NORTH, K. N., STRAUB, V., DOWLING, J. J., MUNTONI, F., CLARKE, N. F., COOPER, S. T., BONNEMANN, C. G. & MACARTHUR, D. G. 2017. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med,* 9.

CVITKOVIC, I. & JURICA, M. S. 2013. Spliceosome database: a tool for tracking components of the spliceosome. *Nucleic Acids Res,* 41**,** D132-41.

DAGONEAU, N., GOULET, M., GENEVIÈVE, D., SZNAJER, Y., MARTINOVIC, J., SMITHSON, S., HUBER, C., BAUJAT, G., FLORI, E., TECCO, L., CAVALCANTI, D., DELEZOIDE, A. L., SERRE, V., LE MERRER, M., MUNNICH, A. & CORMIER-DAIRE, V. 2009. DYNC2H1 mutations cause asphyxiating thoracic dystrophy and short rib-polydactyly syndrome, type III. *Am J Hum Genet,* 84**,** 706-11.

DANIS, D., JACOBSEN, J. O. B., CARMODY, L. C., GARGANO, M. A., MCMURRY, J. A., HEGDE, A., HAENDEL, M. A., VALENTINI, G., SMEDLEY, D. & ROBINSON, P. N. 2021. Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *Am J Hum Genet.*

DAVID, D., MOREIRA, I., MORAIS, S. & DE DEUS, G. 1998. Five novel factor IX mutations in unrelated hemophilia B patients. *Hum Mutat,* Suppl 1**,** S301-3.

DAVYDOV, E. V., GOODE, D. L., SIROTA, M., COOPER, G. M., SIDOW, A. & BATZOGLOU, S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol,* 6**,** e1001025.

DE CONTI, L., BARALLE, M. & BURATTI, E. 2013. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA,* 4**,** 49-60.

DE LIMA, R. L., HOPER, S. A., GHASSIBE, M., COOPER, M. E., RORICK, N. K., KONDO, S., KATZ, L., MARAZITA, M. L., COMPTON, J., BALE, S., HEHR, U., DIXON, M. J., DAACK-HIRSCH, S., BOUTE, O., BAYET, B., REVENCU, N., VERELLEN-DUMOULIN, C., VIKKULA, M., RICHIERI-COSTA, A., MORETTI-FERREIRA, D., MURRAY, J. C. & SCHUTTE, B. C. 2009. Prevalence and nonrandom distribution of exonic mutations in interferon regulatory factor 6 in 307 families with Van der Woude syndrome and 37 families with popliteal pterygium syndrome. *Genet Med,* 11**,** 241-7.

DE ROECK, A., VAN DEN BOSSCHE, T., VAN DER ZEE, J., VERHEIJEN, J., DE COSTER, W., VAN DONGEN, J., DILLEN, L., BARADARAN-HERAVI, Y., HEEMAN, B., SANCHEZ-VALLE, R., LLADO, A., NACMIAS, B., SORBI, S., GELPI, E., GRAU-RIVERA, O., GOMEZ-TORTOSA, E., PASTOR, P., ORTEGA-CUBERO, S., PASTOR, M. A., GRAFF, C., THONBERG, H., BENUSSI, L., GHIDONI, R., BINETTI, G., DE MENDONCA, A., MARTINS, M., BORRONI, B., PADOVANI, A., ALMEIDA, M. R., SANTANA, I., DIEHL-SCHMID, J., ALEXOPOULOS, P., CLARIMON, J., LLEO, A., FORTEA, J., TSOLAKI, M., KOUTROUMANI, M., MATEJ, R., ROHAN, Z., DE DEYN, P., ENGELBORGHS, S., CRAS, P., VAN BROECKHOVEN, C., SLEEGERS, K. & EUROPEAN EARLY-ONSET DEMENTIA, C. 2017. Deleterious ABCA7 mutations and transcript rescue mechanisms in early onset Alzheimer's disease. *Acta Neuropathol,* 134**,** 475-487.

DEL-AGUILA, J. L., LI, Z., DUBE, U., MIHINDUKULASURIYA, K. A., BUDDE, J. P., FERNANDEZ, M. V., IBANEZ, L., BRADLEY, J., WANG, F., BERGMANN, K., DAVENPORT, R., MORRIS, J. C., HOLTZMAN, D. M., PERRIN, R. J., BENITEZ, B. A., DOUGHERTY, J., CRUCHAGA, C. & HARARI, O. 2019. A single-nuclei RNA sequencing study of Mendelian and sporadic AD in the human brain. *Alzheimers Res Ther,* 11**,** 71.

DEN HOLLANDER, A. I., KOENEKOOP, R. K., YZER, S., LOPEZ, I., ARENDS, M. L., VOESENEK, K. E., ZONNEVELD, M. N., STROM, T. M., MEITINGER, T., BRUNNER, H. G., HOYNG, C. B., VAN DEN BORN, L. I., ROHRSCHNEIDER, K. & CREMERS, F. P. 2006. Mutations in the CEP290 (NPHP6) gene are a frequent cause of Leber congenital amaurosis. *Am J Hum Genet,* 79**,** 556-61.

DESMET, F. O., HAMROUN, D., LALANDE, M., COLLOD-BEROUD, G., CLAUSTRES, M. & BEROUD, C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res,* 37**,** e67.

DI LEO, E., PANICO, F., TARUGI, P., BATTISTI, C., FEDERICO, A. & CALANDRA, S. 2004. A point mutation in the lariat branch point of intron 6 of NPC1 as the cause of abnormal pre-mRNA splicing in Niemann-Pick type C disease. *Hum Mutat,* 24**,** 440.

DIONNET, E., DEFOUR, A., DA SILVA, N., SALVI, A., LÉVY, N., KRAHN, M., BARTOLI, M., PUPPO, F. & GOROKHOVA, S. 2020. Splicing impact of deep exonic missense variants in CAPN3 explored systematically by minigene functional assay. *Hum Mutat,* 41**,** 1797-1810.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29**,** 15-21.

DOETSCHMAN, T. 2009. Influence of genetic background on genetically engineered mouse phenotypes. *Methods Mol Biol,* 530**,** 423-33.

DRMANAC, R., SPARKS, A. B., CALLOW, M. J., HALPERN, A. L., BURNS, N. L., KERMANI, B. G., CARNEVALI, P., NAZARENKO, I., NILSEN, G. B., YEUNG, G., DAHL, F., FERNANDEZ, A., STAKER, B., PANT, K. P., BACCASH, J., BORCHERDING, A. P., BROWNLEY, A., CEDENO, R., CHEN, L., CHERNIKOFF, D., CHEUNG, A., CHIRITA, R., CURSON, B., EBERT, J. C., HACKER, C. R., HARTLAGE, R., HAUSER, B., HUANG, S., JIANG, Y., KARPINCHYK, V., KOENIG, M., KONG, C., LANDERS, T., LE, C., LIU, J., MCBRIDE, C. E., MORENZONI, M., MOREY, R. E., MUTCH, K., PERAZICH, H., PERRY, K., PETERS, B. A., PETERSON, J., PETHIYAGODA, C. L., POTHURAJU, K., RICHTER, C., ROSENBAUM, A. M., ROY, S., SHAFTO, J., SHARANHOVICH, U., SHANNON, K. W., SHEPPY, C. G., SUN, M., THAKURIA, J. V., TRAN, A., VU, D., ZARANEK, A. W., WU, X., DRMANAC, S., OLIPHANT, A. R., BANYAI, W. C., MARTIN, B., BALLINGER, D. G., CHURCH, G. M. & REID, C. A. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science,* 327**,** 78-81.

DUTTA, U. R., RAO, S. N., PIDUGU, V. K., V, S. V., BHATTACHERJEE, A., BHOWMIK, A. D., RAMASWAMY, S. K., SINGH, K. G. & DALAL, A. 2019. Breakpoint mapping of a novel de novo translocation t(X;20)(q11.1;p13) by positional cloning and long read sequencing. *Genomics,* 111**,** 1108-1114.

EASTON, D. F., PONDER, M. A., HUSON, S. M. & PONDER, B. A. 1993. An analysis of variation in expression of neurofibromatosis (NF) type 1 (NF1): evidence for modifying genes. *Am J Hum Genet,* 53**,** 305-13.

EDWARDS, C. R., RITCHIE, W., WONG, J. J., SCHMITZ, U., MIDDLETON, R., AN, X., MOHANDAS, N., RASKO, J. E. & BLOBEL, G. A. 2016. A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. *Blood,* 127**,** e24-e34.

ELLINGFORD, J. M., BARTON, S., BHASKAR, S., O'SULLIVAN, J., WILLIAMS, S. G., LAMB, J. A., PANDA, B., SERGOUNIOTIS, P. I., GILLESPIE, R. L., DAIGER, S. P., HALL, G., GALE, T., LLOYD, I. C., BISHOP, P. N., RAMSDEN, S. C. & BLACK, G. C. 2016a. Molecular findings from 537 individuals with inherited retinal disease. *J Med Genet*.

ELLINGFORD, J. M., BARTON, S., BHASKAR, S., WILLIAMS, S. G., SERGOUNIOTIS, P. I., O'SULLIVAN, J., LAMB, J. A., PERVEEN, R., HALL, G., NEWMAN, W. G., BISHOP, P. N., ROBERTS, S. A., LEACH, R., TEARLE, R., BAYLISS, S., RAMSDEN, S. C., NEMETH, A. H. & BLACK, G. C. 2016b. Whole Genome Sequencing Increases Molecular Diagnostic Yield Compared with Current Diagnostic Testing for Inherited Retinal Disease. *Ophthalmology,* 123**,** 1143-50.

ENCODE 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

ERICKSON, R. P., YATSENKO, S. A., LARSON, K. & CHEUNG, S. W. 2011. A Case of Agonadism, Skeletal Malformations, Bicuspid Aortic Valve, and Delayed Development with a 16p13.3 Duplication Including GNG13 and SOX8 Upstream Enhancers: Are Either, Both or Neither Involved in the Phenotype? *Mol Syndromol,* 1**,** 185-191.

ERKELENZ, S., MUELLER, W. F., EVANS, M. S., BUSCH, A., SCHÖNEWEIS, K., HERTEL, K. J. & SCHAAL, H. 2013. Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA,* 19**,** 96-102.

ESTRADA-CUZCANO, A., KOENEKOOP, R. K., SENECHAL, A., DE BAERE, E. B., DE RAVEL, T., BANFI, S., KOHL, S., AYUSO, C., SHARON, D., HOYNG, C. B., HAMEL, C. P., LEROY, B. P., ZIVIELLO, C., LOPEZ, I., BAZINET, A., WISSINGER, B., SLIESORAITYTE, I., AVILA-FERNANDEZ, A., LITTINK, K. W., VINGOLO, E. M., SIGNORINI, S., BANIN, E., MIZRAHI-MEISSONNIER, L., ZRENNER, E., KELLNER, U., COLLIN, R. W., DEN HOLLANDER, A. I., CREMERS, F. P. & KLEVERING, B. J. 2012. BBS1 mutations in a wide spectrum of phenotypes ranging from nonsyndromic retinitis pigmentosa to Bardet-Biedl syndrome. *Arch Ophthalmol,* 130**,** 1425-32.

EVANS, D. G. R., VAN VEEN, E. M., BYERS, H. J., WALLACE, A. J., ELLINGFORD, J. M., BEAMAN, G., SANTOYO-LOPEZ, J., AITMAN, T. J., ECCLES, D. M., LALLOO, F. I., SMITH, M. J. & NEWMAN, W. G. 2018. A Dominantly Inherited 5' UTR Variant Causing Methylation-Associated Silencing of BRCA1 as a Cause of Breast and Ovarian Cancer. *Am J Hum Genet,* 103**,** 213-220.

EVANS, D. G., BOWERS, N., BURKITT-WRIGHT, E., MILES, E., GARG, S., SCOTT-KITCHING, V., PENMAN-SPLITT, M., DOBBIE, A., HOWARD, E., EALING, J., VASSALO, G., WALLACE, A. J., NEWMAN, W., HUSON, S. M. & NETWORK, N. U. N. R. 2016. Comprehensive RNA Analysis of the NF1 Gene in Classically Affected NF1 Affected Individuals Meeting NIH Criteria has High Sensitivity and Mutation Negative Testing is Reassuring in Isolated Cases With Pigmentary Features Only. *EBioMedicine,* 7**,** 212-20.

FABRIZIO, P., DANNENBERG, J., DUBE, P., KASTNER, B., STARK, H., URLAUB, H. & LÜHRMANN, R. 2009. The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. *Mol Cell,* 36**,** 593-608.

FADAIE, Z., WHELAN, L., DOCKERY, A., LI, C. H. Z., VAN DEN BORN, L. I., HOYNG, C. B., GILISSEN, C., COROMINAS, J., ROWLANDS, C., MEGAW, R., LAMPE, A. K., CREMERS, F. P. M., FARRAR, G. J., ELLINGFORD, J. M., KENNA, P. F. & ROOSING, S. 2021. BBS1 branchpoint variant is associated with non-syndromic retinitis pigmentosa. *J Med Genet*.

FAIRBROTHER, W. G., YEH, R. F., SHARP, P. A. & BURGE, C. B. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science,* 297**,** 1007-13.

FARACH, L. S., LITTLE, M. E., DUKER, A. L., LOGAN, C. V., JACKSON, A., HECHT, J. T. & BOBER, M. 2018. The expanding phenotype of RNU4ATAC pathogenic variants to Lowry Wood syndrome. *Am J Med Genet A,* 176**,** 465-469.

FAUSTINO, N. A. & COOPER, T. A. 2003. Pre-mRNA splicing and human disease. *Genes & Development,* 17**,** 419-437.

FERRARO, N. M., STROBER, B. J., EINSON, J., ABELL, N. S., AGUET, F., BARBEIRA, A. N., BRANDT, M., BUCAN, M., CASTEL, S. E., DAVIS, J. R., GREENWALD, E., HESS, G. T., HILLIARD, A. T., KEMBER, R. L., KOTIS, B., PARK, Y., PELOSO, G., RAMDAS, S., SCOTT, A. J., SMAIL, C., TSANG, E. K., ZEKAVAT, S. M., ZIOSI, M., ARADHANA, ARDLIE, K. G., ASSIMES, T. L., BASSIK, M. C., BROWN, C. D., CORREA, A., HALL, I., IM, H. K., LI, X., NATARAJAN, P., LAPPALAINEN, T., MOHAMMADI, P., MONTGOMERY, S. B., BATTLE, A., GROUP, T. L. W. & CONSORTIUM, G. 2020. Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science,* 369.

FILICHKIN, S. A., PRIEST, H. D., GIVAN, S. A., SHEN, R., BRYANT, D. W., FOX, S. E., WONG, W. K. & MOCKLER, T. C. 2010. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res,* 20**,** 45-58.

FINOTELLO, F., LAVEZZO, E., BIANCO, L., BARZON, L., MAZZON, P., FONTANA, P., TOPPO, S. & DI CAMILLO, B. 2014. Reducing bias in RNA sequencing data: a novel approach to compute counts. *BMC Bioinformatics,* 15 Suppl 1**,** S7.

FOKKEMA, I. F., TASCHNER, P. E., SCHAAFSMA, G. C., CELLI, J., LAROS, J. F. & DEN DUNNEN, J. T. 2011. LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat,* 32**,** 557-63.

FORMENT, J. V., KAIDI, A. & JACKSON, S. P. 2012. Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nat Rev Cancer,* 12**,** 663-70.

FREDERICKS, A. M., CYGAN, K. J., BROWN, B. A. & FAIRBROTHER, W. G. 2015. RNA-Binding Proteins: Splicing Factors and Disease. *Biomolecules,* 5**,** 893-909.

FREEMAN, P. J., HART, R. K., GRETTON, L. J., BROOKES, A. J. & DALGLEISH, R. 2018. VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions. *Hum Mutat,* 39**,** 61-68.

FRENDEWEY, D. & KELLER, W. 1985. Stepwise assembly of a pre-mRNA splicing complex requires U-snRNPs and specific intron sequences. *Cell,* 42**,** 355-67.

FRÉSARD, L., SMAIL, C., FERRARO, N. M., TERAN, N. A., LI, X., SMITH, K. S., BONNER, D., KERNOHAN, K. D., MARWAHA, S., ZAPPALA, Z., BALLIU, B., DAVIS, J. R., LIU, B., PRYBOL, C. J., KOHLER, J. N., ZASTROW, D. B., REUTER, C. M., FISK, D. G., GROVE, M. E., DAVIDSON, J. M., HARTLEY, T., JOSHI, R., STROBER, B. J., UTIRAMERUR, S., LIND, L., INGELSSON, E., BATTLE, A., BEJERANO, G., BERNSTEIN, J. A., ASHLEY, E. A., BOYCOTT, K. M., MERKER, J. D., WHEELER, M. T., MONTGOMERY, S. B., NETWORK, U. D. & CONSORTIUM, C. R. C. 2019. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med,* 25**,** 911-919.

GABBAY, M., ELLARD, S., DE FRANCO, E. & MOISÉS, R. S. 2017. Pancreatic Agenesis due to Compound Heterozygosity for a Novel Enhancer and Truncating Mutation in the PTF1A Gene. *J Clin Res Pediatr Endocrinol,* 9**,** 274-277.

GAILDRAT, P., KILLIAN, A., MARTINS, A., TOURNIER, I., FRÉBOURG, T. & TOSI, M. 2010. Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. *Methods Mol Biol,* 653**,** 249-57.

GARRIDO-MARTÍN, D., BORSARI, B., CALVO, M., REVERTER, F. & GUIGÓ, R. 2021. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat Commun,* 12**,** 727.

GELB, B. D., CAVE, H., DILLON, M. W., GRIPP, K. W., LEE, J. A., MASON-SUARES, H., RAUEN, K. A., WILLIAMS, B., ZENKER, M., VINCENT, L. M. & CLINGEN, R. W. G. 2018. ClinGen's RASopathy Expert Panel consensus methods for variant interpretation. *Genet Med,* 20**,** 1334-1345.

GELFMAN, S., WANG, Q., MCSWEENEY, K. M., REN, Z., LA CARPIA, F., HALVORSEN, M., SCHOCH, K., RATZON, F., HEINZEN, E. L., BOLAND, M. J., PETROVSKI, S. & GOLDSTEIN, D. B. 2017. Annotating pathogenic non-coding variants in genic regions. *Nat Commun,* 8**,** 236.

GILLESPIE, R. L., O'SULLIVAN, J., ASHWORTH, J., BHASKAR, S., WILLIAMS, S., BISWAS, S., KEHDI, E., RAMSDEN, S. C., CLAYTON-SMITH, J., BLACK, G. C. & LLOYD, I. C. 2014. Personalized Diagnosis and Management of Congenital Cataract by Next-Generation Sequencing. *Ophthalmology,* 121**,** 2124-U302.

GONORAZKY, H. D., NAUMENKO, S., RAMANI, A. K., NELAKUDITI, V., MASHOURI, P., WANG, P., KAO, D., OHRI, K., VITHTHIYAPASKARAN, S., TARNOPOLSKY, M. A., MATHEWS, K. D., MOORE, S. A., OSORIO, A. N., VILLANOVA, D., KEMALADEWI, D. U., COHN, R. D., BRUDNO, M. & DOWLING, J. J. 2019. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am J Hum Genet,* 104**,** 1007.

GOODWIN, S., MCPHERSON, J. D. & MCCOMBIE, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet,* 17**,** 333-51.

GOTO, Y., HORAI, S., MATSUOKA, T., KOGA, Y., NIHEI, K., KOBAYASHI, M. & NONAKA, I. 1992. Mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes (MELAS): a correlative study of the clinical features and mitochondrial DNA mutation. *Neurology,* 42**,** 545-50.

GRANTHAM, R. 1974. Amino acid difference formula to help explain protein evolution. *Science,* 185**,** 862-4.

GRIESEMER, D., XUE, J. R., REILLY, S. K., ULIRSCH, J. C., KUKREJA, K., DAVIS, J., KANAI, M., YANG, D. K., MONTGOMERY, S. B., NOVINA, C. D., TEWHEY, R. & SABEDI, P. C. 2021. Genome-wide functional screen of 3' UTR variants uncovers causal variants for human disease and evolution. *bioRxiv*.

GTEX CONSORTIUM. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet,* 45**,** 580-5.

GTEX CONSORTIUM. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science,* 369**,** 1318-1330.

GUTIERREZ-RODRIGUES, F., DONAIRES, F. S., PINTO, A., VICENTE, A., DILLON, L. W., CLÉ, D. V., SANTANA, B. A., PIROOZNIA, M., IBANEZ, M. D. P. F., TOWNSLEY, D. M., KAJIGAYA, S., HOURIGAN, C. S., COOPER, J. N., CALADO, R. T. & YOUNG, N. S. 2019. Pathogenic TERT promoter variants in telomere diseases. *Genet Med,* 21**,** 1594-1602.

HAENDEL, M., VASILEVSKY, N., UNNI, D., BOLOGA, C., HARRIS, N., REHM, H., HAMOSH, A., BAYNAM, G., GROZA, T., MCMURRY, J., DAWKINS, H., RATH, A., THAXON, C., BOCCI, G., JOACHIMIAK, M. P., KOHLER, S., ROBINSON, P. N., MUNGALL, C. & OPREA, T. I. 2020. How many rare diseases are there? *Nat Rev Drug Discov,* 19**,** 77-78.

HAM, S. & LEE, S. V. 2020. Advances in transcriptome analysis of human brain aging. *Exp Mol Med,* 52**,** 1787-1797.

HARBOUR, J. W. 2001. Molecular basis of low-penetrance retinoblastoma. *Arch Ophthalmol,* 119**,** 1699-704.

HARROW, J., FRANKISH, A., GONZALEZ, J. M., TAPANARI, E., DIEKHANS, M., KOKOCINSKI, F., AKEN, B. L., BARRELL, D., ZADISSA, A., SEARLE, S., BARNES, I., BIGNELL, A., BOYCHENKO, V., HUNT, T., KAY, M., MUKHERJEE, G., RAJAN, J., DESPACIO-REYES, G., SAUNDERS, G., STEWARD, C., HARTE, R., LIN, M., HOWALD, C., TANZER, A., DERRIEN, T., CHRAST, J., WALTERS, N., BALASUBRAMANIAN, S., PEI, B., TRESS, M., RODRIGUEZ, J. M., EZKURDIA, I., VAN BAREN, J., BRENT, M., HAUSSLER, D., KELLIS, M., VALENCIA, A., REYMOND, A., GERSTEIN, M., GUIGO, R. & HUBBARD, T. J. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res,* 22**,** 1760-74.

HASTINGS, M. L., RESTA, N., TRAUM, D., STELLA, A., GUANTI, G. & KRAINER, A. R. 2005. An LKB1 AT-AC intron mutation causes Peutz-Jeghers syndrome via splicing at noncanonical cryptic splice sites. *Nat Struct Mol Biol,* 12**,** 54-9.

HAWKE, L., BOWMAN, M. L., POON, M. C., SCULLY, M. F., RIVARD, G. E. & JAMES, P. D. 2016. Characterization of aberrant splicing of von Willebrand factor in von Willebrand disease: an underrecognized mechanism. *Blood,* 128**,** 584-93.

HE, H., LIYANARACHCHI, S., AKAGI, K., NAGY, R., LI, J., DIETRICH, R. C., LI, W., SEBASTIAN, N., WEN, B., XIN, B., SINGH, J., YAN, P., ALDER, H., HAAN, E., WIECZOREK, D., ALBRECHT, B., PUFFENBERGER, E., WANG, H., WESTMAN, J. A., PADGETT, R. A., SYMER, D. E. & DE LA CHAPELLE, A. 2011. Mutations in U4atac snRNA, a component of the minor spliceosome, in the developmental disorder MOPD I. *Science,* 332**,** 238-40.

HE, K., ZHANG, X., REN, S. & SUN, J. 2016. Deep residual learning for image recognition. *In:* O'CONNER, L., ed. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 770-778.

HEBSGAARD, S. M., KORNING, P. G., TOLSTRUP, N., ENGELBRECHT, J., ROUZÉ, P. & BRUNAK, S. 1996. Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Res,* 24**,** 3439-52.

HOLMES, S. E., DOMBROSKI, B. A., KREBS, C. M., BOEHM, C. D. & KAZAZIAN, H. H. 1994. A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nat Genet,* 7**,** 143-8.

HOWE, K. L., ACHUTHAN, P., ALLEN, J., ALVAREZ-JARRETA, J., AMODE, M. R., ARMEAN, I. M., AZOV, A. G., 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, O. G., JUETTEMANN, T., KAIKALA, V., KAY, M., LAVIDAS, I., LE, T., LEMOS, D., GONZALEZ MARTINEZ, J., MARUGÁN, J. C., MAUREL, T., MCMAHON, A. C., MOHANAN, S., MOORE, B., MUFFATO, M., OHEH, D. N., PARASCHAS, D., PARKER, A., PARTON, A., PROSOVETSKAIA, I., SAKTHIVEL, M. P., SALAM, A. I. A., SCHMITT, B. M., 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, S. E., IISLEY, G. R., LANGRIDGE, N., LOVELAND, J. E., MARTIN, F.

J., MUDGE, J. M., MORALES, J., PERRY, E., RUFFIER, M., TATE, J., THYBERT, D., TREVANION, S. J., CUNNINGHAM, F., YATES, A. D., ZERBINO, D. R. & FLICEK, P. 2021. Ensembl 2021. *Nucleic Acids Res,* 49**,** D884-D891.

HYON, C., CHANTOT-BASTARAUD, S., HARBUZ, R., BHOURI, R., PERROT, N., PEYCELON, M., SIBONY, M., ROJO, S., PIGUEL, X., BILAN, F., GILBERT-DUSSARDIER, B., KITZIS, A., MCELREAVEY, K., SIFFROI, J. P. & BASHAMBOO, A. 2015. Refining the regulatory region upstream of SOX9 associated with 46,XX testicular disorders of Sex Development (DSD). *Am J Med Genet A,* 167A**,** 1851-8.

IRIMIA, M., WEATHERITT, R. J., ELLIS, J. D., PARIKSHAK, N. N., GONATOPOULOS-POURNATZIS, T., BABOR, M., QUESNEL-VALLIERES, M., TAPIAL, J., RAJ, B., O'HANLON, D., BARRIOS-RODILES, M., STERNBERG, M. J., CORDES, S. P., ROTH, F. P., WRANA, J. L., GESCHWIND, D. H. & BLENCOWE, B. J. 2014. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell,* 159**,** 1511-23.

ISHIURA, H., DOI, K., MITSUI, J., YOSHIMURA, J., MATSUKAWA, M. K., FUJIYAMA, A., TOYOSHIMA, Y., KAKITA, A., TAKAHASHI, H., SUZUKI, Y., SUGANO, S., QU, W., ICHIKAWA, K., YURINO, H., HIGASA, K., SHIBATA, S., MITSUE, A., TANAKA, M., ICHIKAWA, Y., TAKAHASHI, Y., DATE, H., MATSUKAWA, T., KANDA, J., NAKAMOTO, F. K., HIGASHIHARA, M., ABE, K., KOIKE, R., SASAGAWA, M., KUROHA, Y., HASEGAWA, N., KANESAWA, N., KONDO, T., HITOMI, T., TADA, M., TAKANO, H., SAITO, Y., SANPEI, K., ONODERA, O., NISHIZAWA, M., NAKAMURA, M., YASUDA, T., SAKIYAMA, Y., OTSUKA, M., UEKI, A., KAIDA, K. I., SHIMIZU, J., HANAJIMA, R., HAYASHI, T., TERAO, Y., INOMATA-TERADA, S., HAMADA, M., SHIROTA, Y., KUBOTA, A., UGAWA, Y., KOH, K., TAKIYAMA, Y., OHSAWA-YOSHIDA, N., ISHIURA, S., YAMASAKI, R., TAMAOKA, A., AKIYAMA, H., OTSUKI, T., SANO, A., IKEDA, A., GOTO, J., MORISHITA, S. & TSUJI, S. 2018. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat Genet,* 50**,** 581-590.

JAGADEESH, K. A., PAGGI, J. M., YE, J. S., STENSON, P. D., COOPER, D. N., BERNSTEIN, J. A. & BEJERANO, G. 2019. S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nat Genet,* 51**,** 755-763.

JAGANATHAN, K., KYRIAZOPOULOU PANAGIOTOPOULOU, S., MCRAE, J. F., DARBANDI, S. F., KNOWLES, D., LI, Y. I., KOSMICKI, J. A., ARBELAEZ, J., CUI, W., SCHWARTZ, G. B., CHOW, E. D., KANTERAKIS, E., GAO, H., KIA, A., BATZOGLOU, S., SANDERS, S. J. & FARH, K. K. 2019a. Predicting Splicing from Primary Sequence with Deep Learning. *Cell,* 176**,** 535-548.e24.

JAIN, M., KOREN, S., MIGA, K. H., QUICK, J., RAND, A. C., SASANI, T. A., TYSON, J. R., BEGGS, A. D., DILTHEY, A. T., FIDDES, I. T., MALLA, S., MARRIOTT, H., NIETO, T., O'GRADY, J., OLSEN, H. E., PEDERSEN, B. S., RHIE, A., RICHARDSON, H., QUINLAN, A. R., SNUTCH, T. P., TEE, L., PATEN, B., PHILLIPPY, A. M., SIMPSON, J. T., LOMAN, N. J. & LOOSE, M. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol,* 36**,** 338-345.

JAMES, P. D., NOTLEY, C., HEGADORN, C., LEGGO, J., TUTTLE, A., TINLIN, S., BROWN, C., ANDREWS, C., LABELLE, A., CHIRINIAN, Y., O'BRIEN, L., OTHMAN, M., RIVARD, G., RAPSON, D., HOUGH, C. & LILLICRAP, D. 2007. The mutational spectrum of type 1 von Willebrand disease: Results from a Canadian cohort study. *Blood,* 109**,** 145-54.

JANSSEN, R. J., WEVERS, R. A., HÄUSSLER, M., LUYTEN, J. A., STEENBERGEN-SPANJERS, G. C., HOFFMANN, G. F., NAGATSU, T. & VAN DEN HEUVEL, L. P. 2000. A branch site mutation leading to aberrant splicing of the human tyrosine hydroxylase gene in a child with a severe extrapyramidal movement disorder. *Ann Hum Genet,* 64**,** 375-82.

JENKINSON, G., LI, Y. I., BASU, S., COUSIN, M. A., OLIVER, G. R. & KLEE, E. W. 2020. LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. *Bioinformatics,* 36**,** 4609-4615.

JØRGENSEN, S. E., BØTTGER, P., KOFOD-OLSEN, E., HOLM, M., MØRK, N., ØRNTOFT, T. F., SØRENSEN, U. B. S., BERNTH-JENSEN, J. M., HERLIN, T., VEIRUM, J., LARSEN, C. S., ØSTERGAARD, L., HARTMANN, R., CHRISTIANSEN, M. & MOGENSEN, T. H. 2016. Ectodermal dysplasia with immunodeficiency caused by a branch-point mutation in IKBKG/NEMO. *J Allergy Clin Immunol,* 138**,** 1706-1709.e4.

KARCZEWSKI, K. J., FRANCIOLI, L. C., TIAO, G., CUMMINGS, B. B., ALFÖLDI, J., WANG, Q., COLLINS, R. L., LARICCHIA, K. M., GANNA, A., BIRNBAUM, D. P., GAUTHIER, L. D., BRAND, H., SOLOMONSON, M., WATTS, N. A., RHODES, D., SINGER-BERK, M., ENGLAND, E. M., SEABY, E. G., KOSMICKI, J. A., WALTERS, R. K., TASHMAN, K., FARJOUN, Y., BANKS, E., POTERBA, T., WANG, A., SEED, C., WHIFFIN, N., CHONG, J. X., SAMOCHA, K. E., PIERCE-HOFFMAN, E., ZAPPALA, Z., O'DONNELL-LURIA, A. H., MINIKEL, E. V., WEISBURD, B., LEK, M., WARE, J. S., VITTAL, C., ARMEAN, I. M., BERGELSON, L., CIBULSKIS, K., CONNOLLY, K. M., 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, M. E., NEALE, B. M., DALY, M. J., MACARTHUR, D. G. & CONSORTIUM, G. A. D. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature,* 581**,** 434-443.

KARNUTA, J. M. & SCACHERI, P. C. 2018. Enhancers: bridging the gap between gene control and human disease. *Hum Mol Genet,* 27**,** R219-R227.

KE, S. & CHASIN, L. A. 2010. Intronic motif pairs cooperate across exons to promote pre-mRNA splicing. *Genome Biol,* 11**,** R84.

KE, S., SHANG, S., KALACHIKOV, S. M., MOROZOVA, I., YU, L., RUSSO, J. J., JU, J. & CHASIN, L. A. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res,* 21**,** 1360-74.

KETTERLING, R. P., DROST, J. B., SCARINGE, W. A., LIAO, D. Z., LIU, J. Z., KASPER, C. K. & SOMMER, S. S. 1999. Reported in vivo splice-site mutations in the factor IX gene: severity of splicing defects and a hypothesis for predicting deleterious splice donor mutations. *Hum Mutat,* 13**,** 221-31.

KHAN, M., CORNELIS, S. S., POZO-VALERO, M. D., WHELAN, L., RUNHART, E. H., MISHRA, K., BULTS, F., ALSWAITI, Y., ALTALBISHI, A., DE BAERE, E., BANFI, S., BANIN, E., BAUWENS, M., BEN-YOSEF, T., BOON, C. J. F., VAN DEN BORN, L. I., DEFOORT, S., DEVOS, A., DOCKERY, A., DUDAKOVA, L., FAKIN, A., FARRAR, G. J., SALLUM, J. M. F., FUJINAMI, K., GILISSEN, C., GLAVAČ, D., GORIN, M. B., GREENBERG, J., HAYASHI, T., HETTINGA, Y. M., HOISCHEN, A., HOYNG, C. B., HUFENDIEK, K., JÄGLE, H., KAMAKARI, S., KARALI, M., KELLNER, U., KLAVER, C. C. W., KOUSAL, B., LAMEY, T. M., MACDONALD, I. M., MATYNIA, A., MCLAREN, T. L., MENA, M. D., MEUNIER, I., MILLER, R., NEWMAN, H., NTOZINI, B., OLDAK, M., PIETERSE, M., PODHAJCER, O. L., PUECH, B., RAMESAR, R., RÜTHER, K., SALAMEH, M., SALLES, M. V., SHARON, D., SIMONELLI, F., SPITAL, G., STEEHOUWER, M., SZAFLIK, J. P., THOMPSON, J. A., THUILLIER, C., TRACEWSKA, A. M., VAN ZWEEDEN, M., VINCENT, A. L., ZANLONGHI, X., LISKOVA, P., STÖHR, H., ROACH, J. N., AYUSO, C., ROBERTS, L., WEBER, B. H. F., DHAENENS, C. M. & CREMERS, F. P. M. 2020. Resolving the dark matter of ABCA4 for 1054 Stargardt disease probands through integrated genomics and transcriptomics. *Genet Med,* 22**,** 1235-1246.

KHAN, S. G., METIN, A., GOZUKARA, E., INUI, H., SHAHLAVI, T., MUNIZ-MEDINA, V., BAKER, C. C., UEDA, T., AIKEN, J. R., SCHNEIDER, T. D. & KRAEMER, K. H. 2004. Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: mutations result in reduced XPC mRNA levels that correlate with cancer risk. *Hum Mol Genet,* 13**,** 343-52.

KIM, D., PAGGI, J. M., PARK, C., BENNETT, C. & SALZBERG, S. L. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol,* 37**,** 907-915.

KIM, D., PERTEA, G., TRAPNELL, C., PIMENTEL, H., KELLEY, R. & SALZBERG, S. L. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol,* 14**,** R36.

KIM, S., SCHEFFLER, K., HALPERN, A. L., BEKRITSKY, M. A., NOH, E., KÄLLBERG, M., CHEN, X., KIM, Y., BEYTER, D., KRUSCHE, P. & SAUNDERS, C. T. 2018. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods,* 15**,** 591-594.

KIRCHER, M., WITTEN, D. M., JAIN, P., O'ROAK, B. J., COOPER, G. M. & SHENDURE, J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics,* 46**,** 310-+.

KISHORE, S., KHANNA, A. & STAMM, S. 2008. Rapid generation of splicing reporters with pSpliceExpress. *Gene,* 427**,** 104-10.

KRAGESTEEN, B. K., BRANCATI, F., DIGILIO, M. C., MUNDLOS, S. & SPIELMANN, M. 2019. promoter deletion causes. *J Med Genet,* 56**,** 246-251.

KRAWCZAK, M., THOMAS, N. S., HUNDRIESER, B., MORT, M., WITTIG, M., HAMPE, J. & COOPER, D. N. 2007. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum Mutat,* 28**,** 150-8.

KREMER, L. S., BADER, D. M., MERTES, C., KOPAJTICH, R., PICHLER, G., IUSO, A., HAACK, T. B., GRAF, E., SCHWARZMAYR, T., TERRILE, C., KONARIKOVA, E., REPP, B., KASTENMULLER, G., ADAMSKI, J., LICHTNER, P., LEONHARDT, C., FUNALOT, B., DONATI, A., TIRANTI, V., LOMBES, A., JARDEL, C., GLASER, D., TAYLOR, R. W., GHEZZI, D., MAYR, J. A., ROTIG, A., FREISINGER, P., DISTELMAIER, F., STROM, T. M., MEITINGER, T., GAGNEUR, J. & PROKISCH, H. 2017. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat Commun,* 8**,** 15824.

KUIJPER, E. C., BERGSMA, A. J., PIJNAPPEL, W. W. M. P. & AARTSMA-RUS, A. 2021. Opportunities and challenges for antisense oligonucleotide therapies. *J Inherit Metab Dis,* 44**,** 72-87.

KUIVENHOVEN, J. A., WEIBUSCH, H., PRITCHARD, P. H., FUNKE, H., BENNE, R., ASSMANN, G. & KASTELEIN, J. J. 1996. An intronic mutation in a lariat branchpoint sequence is a direct cause of an inherited human disorder (fish-eye disease). *J Clin Invest,* 98**,** 358-64.

KUKURBA, K. R., ZHANG, R., LI, X., SMITH, K. S., KNOWLES, D. A., HOW TAN, M., PISKOL, R., LEK, M., SNYDER, M., MACARTHUR, D. G., LI, J. B. & MONTGOMERY, S. B. 2014. Allelic expression of deleterious protein-coding variants across human tissues. *PLoS Genet,* 10**,** e1004304.

KUNDAJE, A., MEULEMAN, W., ERNST, J., BILENKY, M., YEN, A., HERAVI-MOUSSAVI, A., KHERADPOUR, P., ZHANG, Z., WANG, J., ZILLER, M. J., AMIN, V., WHITAKER, J. W., SCHULTZ, M. D., WARD, L. D., SARKAR, A., QUON, G., SANDSTROM, R. S., EATON, M. L., WU, Y. C., PFENNING, A. R., WANG, X., CLAUSSNITZER, M., LIU, Y., COARFA, C., HARRIS, R. A., SHORESH, N., EPSTEIN, C. B., GJONESKA, E., LEUNG, D., XIE, W., HAWKINS, R. D., LISTER, R., HONG, C., GASCARD, P., MUNGALL, A. J., MOORE, R., CHUAH, E., TAM, A., CANFIELD, T. K., HANSEN, R. S., KAUL, R., SABO, P. J., BANSAL, M. S., CARLES, A., DIXON, J. R., FARH, K. H., FEIZI, S., KARLIC, R., KIM, A. R., KULKARNI, A., LI, D., LOWDON, R., ELLIOTT, G., MERCER, T. R., NEPH, S. J., ONUCHIC, V., POLAK, P., RAJAGOPAL, N., RAY, P., SALLARI, R. C., SIEBENTHALL, K. T., SINNOTT-ARMSTRONG, N. A., STEVENS, M., THURMAN, R. E., WU, J., ZHANG, B., ZHOU, X., BEAUDET, A. E., BOYER, L. A., DE JAGER, P. L., FARNHAM, P. J., FISHER, S. J., HAUSSLER, D., JONES, S. J., LI, W., MARRA, M. A., MCMANUS, M. T., SUNYAEV, S., THOMSON, J. A., TLSTY, T. D., TSAI, L. H., WANG, W., WATERLAND, R. A., ZHANG, M. Q., CHADWICK, L. H., BERNSTEIN, B. E., COSTELLO, J. F., ECKER, J. R., HIRST, M., MEISSNER, A., MILOSAVLJEVIC, A., REN, B., STAMATOYANNOPOULOS, J. A., WANG, T., KELLIS, M. & CONSORTIUM, R. E. 2015. Integrative analysis of 111 reference human epigenomes. *Nature,* 518**,** 317-30.

LANDRUM, M. J., LEE, J. M., BENSON, M., BROWN, G. R., CHAO, C., CHITIPIRALLA, S., GU, B., HART, J., HOFFMAN, D., JANG, W., KARAPETYAN, K., KATZ, K., LIU, C., MADDIPATLA, Z., MALHEIRO, A., MCDANIEL, K., OVETSKY, M., RILEY, G., ZHOU, G., HOLMES, J. B., KATTMAN, B. L. & MAGLOTT, D. R. 2018. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res,* 46**,** D1062-D1067.

LARSSON, L., FRISÉN, J. & LUNDEBERG, J. 2021. Spatially resolved transcriptomics adds a new dimension to genomics. *Nat Methods,* 18**,** 15-18.

LE GUÉDARD-MÉREUZE, S., VACHÉ, C., BAUX, D., FAUGÈRE, V., LARRIEU, L., ABADIE, C., JANECKE, A., CLAUSTRES, M., ROUX, A. F. & TUFFERY-GIRAUD, S. 2010. Ex vivo splicing assays of mutations at noncanonical positions of splice sites in USHER genes. *Hum Mutat,* 31**,** 347-55.

LEE, B., VITALE, E., SUPERTI-FURGA, A., STEINMANN, B. & RAMIREZ, F. 1991. G to T transversion at position +5 of a splice donor site causes skipping of the preceding exon in the type III procollagen transcripts of a patient with Ehlers-Danlos syndrome type IV. *J Biol Chem,* 266**,** 5256-9.

LEE, H., HUANG, A. Y., WANG, L. K., YOON, A. J., RENTERIA, G., ESKIN, A., SIGNER, R. H., DORRANI, N., NIEVES-RODRIGUEZ, S., WAN, J., DOUINE, E. D., WOODS, J. D., DELL'ANGELICA, E. C., FOGEL, B. L., MARTIN, M. G., BUTTE, M. J., PARKER, N. H., WANG, R. T., SHIEH, P. B., WONG, D. A., GALLANT, N., SINGH, K. E., TAVYEV ASHER, Y. J., SINSHEIMER, J. S., KRAKOW, D., LOO, S. K., ALLARD, P., PAPP, J. C., PALMER, C. G. S., MARTINEZ-AGOSTO, J. A., NELSON, S. F. & NETWORK, U. D. 2020. Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genet Med,* 22**,** 490-499.

LEE, K., BERG, J. S., MILKO, L., CROOKS, K., LU, M., BIZON, C., OWEN, P., WILHELMSEN, K. C., WECK, K. E., EVANS, J. P. & GARG, S. 2015. High Diagnostic Yield of Whole Exome

Sequencing in Participants With Retinal Dystrophies in a Clinical Ophthalmology Setting. *American Journal of Ophthalmology,* 160**,** 354-363.

LEE, M., ROOS, P., SHARMA, N., ATALAR, M., EVANS, T. A., PELLICORE, M. J., DAVIS, E., LAM, A. N., STANLEY, S. E., KHALIL, S. E., SOLOMON, G. M., WALKER, D., RARAIGH, K. S., VECCHIO-PAGAN, B., ARMANIOS, M. & CUTTING, G. R. 2017. Systematic Computational Identification of Variants That Activate Exonic and Intronic Cryptic Splice Sites. *Am J Hum Genet,* 100**,** 751-765.

LEK, M., KARCZEWSKI, K. J., MINIKEL, E. V., SAMOCHA, K. E., BANKS, E., FENNELL, T., O'DONNELL-LURIA, A. H., WARE, J. S., HILL, A. J., CUMMINGS, B. B., TUKIAINEN, T., BIRNBAUM, D. P., KOSMICKI, J. A., DUNCAN, L. E., ESTRADA, K., ZHAO, F., ZOU, J., PIERCE-HOFFMAN, E., BERGHOUT, J., COOPER, D. N., DEFLAUX, N., DEPRISTO, M., DO, R., FLANNICK, J., FROMER, M., GAUTHIER, L., GOLDSTEIN, J., GUPTA, N., HOWRIGAN, D., KIEZUN, A., KURKI, M. I., MOONSHINE, A. L., NATARAJAN, P., OROZCO, L., PELOSO, G. M., POPLIN, R., RIVAS, M. A., RUANO-RUBIO, V., ROSE, S. A., RUDERFER, D. M., SHAKIR, K., STENSON, P. D., STEVENS, C., THOMAS, B. P., TIAO, G., TUSIE-LUNA, M. T., WEISBURD, B., WON, H. H., YU, D., ALTSHULER, D. M., ARDISSINO, D., BOEHNKE, M., DANESH, J., DONNELLY, S., ELOSUA, R., FLOREZ, J. C., GABRIEL, S. B., GETZ, G., GLATT, S. J., HULTMAN, C. M., KATHIRESAN, S., LAAKSO, M., MCCARROLL, S., MCCARTHY, M. I., MCGOVERN, D., MCPHERSON, R., NEALE, B. M., PALOTIE, A., PURCELL, S. M., SALEHEEN, D., SCHARF, J. M., SKLAR, P., SULLIVAN, P. F., TUOMILEHTO, J., TSUANG, M. T., WATKINS, H. C., WILSON, J. G., DALY, M. J., MACARTHUR, D. G. & CONSORTIUM, E. A. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* 536**,** 285-91.

LEK, M., KARCZEWSKI, K. J., MINIKEL, E. V., SAMOCHA, K. E., BANKS, E., FENNELL, T., O'DONNELL-LURIA, A. H., WARE, J. S., HILL, A. J., CUMMINGS, B. B., TUKIAINEN, T., BIRNBAUM, D. P., KOSMICKI, J. A., DUNCAN, L. E., ESTRADA, K., ZHAO, F., ZOU, J., PIERCE-HOFFMAN, E., BERGHOUT, J., COOPER, D. N., DEFLAUX, N., DEPRISTO, M., DO, R., FLANNICK, J., FROMER, M., GAUTHIER, L., GOLDSTEIN, J., GUPTA, N., HOWRIGAN, D., KIEZUN, A., KURKI, M. I., MOONSHINE, A. L., NATARAJAN, P., OROZCO, L., PELOSO, G. M., POPLIN, R., RIVAS, M. A., RUANO-RUBIO, V., ROSE, S. A., RUDERFER, D. M., SHAKIR, K., STENSON, P. D., STEVENS, C., THOMAS, B. P., TIAO, G., TUSIE-LUNA, M. T., WEISBURD, B., WON, H. H., YU, D., ALTSHULER, D. M., ARDISSINO, D., BOEHNKE, M., DANESH, J., DONNELLY, S., ELOSUA, R., FLOREZ, J. C., GABRIEL, S. B., GETZ, G., GLATT, S. J., HULTMAN, C. M., KATHIRESAN, S., LAAKSO, M., MCCARROLL, S., MCCARTHY, M. I., MCGOVERN, D., MCPHERSON, R., NEALE, B. M., PALOTIE, A., PURCELL, S. M., SALEHEEN, D., SCHARF, J. M., SKLAR, P., SULLIVAN, P. F., TUOMILEHTO, J., TSUANG, M. T., WATKINS, H. C., WILSON, J. G., DALY, M. J., MACARTHUR, D. G. & CONSORTIUM, E. A. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* 536**,** 285-91.

LEMAN, R., TUBEUF, H., RAAD, S., TOURNIER, I., DERAMBURE, C., LANOS, R., GAILDRAT, P., CASTELAIN, G., HAUCHARD, J., KILLIAN, A., BAERT-DESURMONT, S., LEGROS, A., GOARDON, N., QUESNELLE, C., RICOU, A., CASTERA, L., VAUR, D., LE GAC, G., KA, C., FICHOU, Y., BONNET-DORION, F., SEVENET, N., GUILLAUD-BATAILLE, M., BOUTRY-KRYZA, N., SCHULTZ, I., CAUX-MONCOUTIER, V., ROSSING, M., WALKER, L. C., SPURDLE, A. B., HOUDAYER, C., MARTINS, A. & KRIEGER, S. 2020. Assessment of branch point prediction tools to predict physiological branch points and their alteration by variants. *BMC Genomics,* 21**,** 86.

LEVINE, A. & DURBIN, R. 2001. A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res,* 29**,** 4006-13.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-1760.

LI, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & SUBGROUP, G. P. D. P. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LI, L., CAO, Y., ZHAO, F., MAO, B., REN, X., WANG, Y., GUAN, Y., YOU, Y., LI, S., YANG, T. & ZHAO, X. 2019. Validation and Classification of Atypical Splicing Variants Associated With Osteogenesis Imperfecta. *Front Genet,* 10**,** 979.

LIN, M., PEDROSA, E., SHAH, A., HRABOVSKY, A., MAQBOOL, S., ZHENG, D. & LACHMAN, H. M. 2011. RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. *PLoS One,* 6**,** e23356.

LINES, M. A., HUANG, L., SCHWARTZENTRUBER, J., DOUGLAS, S. L., LYNCH, D. C., BEAULIEU, C., GUION-ALMEIDA, M. L., ZECHI-CEIDE, R. M., GENER, B., GILLESSEN-KAESBACH, G., NAVA, C., BAUJAT, G., HORN, D., KINI, U., CALIEBE, A., ALANAY, Y., UTINE, G. E., LEV, D., KOHLHASE, J., GRIX, A. W., LOHMANN, D. R., HEHR, U., BÖHM, D., MAJEWSKI, J., BULMAN, D. E., WIECZOREK, D., BOYCOTT, K. M. & CONSORTIUM, F. C. 2012. Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly. *Am J Hum Genet,* 90**,** 369-77.

LITWACK, G. 2018. Nucleic Acids and Molecular Genetics. *Human Biochemistry.* 1 ed.: Academic Press.

LIU, H. X., CHEW, S. L., CARTEGNI, L., ZHANG, M. Q. & KRAINER, A. R. 2000. Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol Cell Biol,* 20**,** 1063-71.

LIU, J., ZHANG, Z., BANDO, M., ITOH, T., DEARDORFF, M. A., CLARK, D., KAUR, M., TANDY, S., KONDOH, T., RAPPAPORT, E., SPINNER, N. B., VEGA, H., JACKSON, L. G., SHIRAHIGE, K. & KRANTZ, I. D. 2009. Transcriptional dysregulation in NIPBL and cohesin mutant human cells. *PLoS Biol,* 7**,** e1000119.

LIU, L., LI, Y., LI, S., HU, N., HE, Y., PONG, R., LIN, D., LU, L. & LAW, M. 2012. Comparison of next-generation sequencing systems. *J Biomed Biotechnol,* 2012**,** 251364.

LORD, J., GALLONE, G., SHORT, P. J., MCRAE, J. F., IRONFIELD, H., WYNN, E. H., GERETY, S. S., HE, L., KERR, B., JOHNSON, D. S., MCCANN, E., KINNING, E., FLINTER, F., TEMPLE, I. K., CLAYTON-SMITH, J., MCENTAGART, M., LYNCH, S. A., JOSS, S., DOUZGOU, S., DABIR, T., CLOWES, V., MCCONNELL, V. P. M., LAM, W., WRIGHT, C. F., FITZPATRICK, D. R., FIRTH, H. V., BARRETT, J. C., HURLES, M. E. & STUDY, D. D. D. 2019. Pathogenicity and selective constraint on variation near splice sites. *Genome Res,* 29**,** 159-170.

MANTERE, T., KERSTEN, S. & HOISCHEN, A. 2019. Long-Read Sequencing Emerging in Medical Genetics. *Front Genet,* 10**,** 426.

MARCO-PUCHE, G., LOIS, S., BENÍTEZ, J. & TRIVINO, J. C. 2019. RNA-Seq Perspectives to Improve Clinical Diagnosis. *Front Genet,* 10**,** 1152.

MARQUEZ, Y., HÖPFLER, M., AYATOLLAHI, Z., BARTA, A. & KALYNA, M. 2015. Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity. *Genome Res,* 25**,** 995-1007.

MARSTON, S., COPELAND, O., JACQUES, A., LIVESEY, K., TSANG, V., MCKENNA, W. J., JALILZADEH, S., CARBALLO, S., REDWOOD, C. & WATKINS, H. 2009. Evidence from human myectomy samples that MYBPC3 mutations cause hypertrophic cardiomyopathy through haploinsufficiency. *Circ Res,* 105**,** 219-22.

MARX, V. 2021. Method of the Year: spatially resolved transcriptomics. *Nat Methods,* 18**,** 9-14.

MASLEN, C., BABCOCK, D., RAGHUNATH, M. & STEINMANN, B. 1997. A rare branch-point mutation is associated with missplicing of fibrillin-2 in a large family with congenital contractural arachnodactyly. *Am J Hum Genet,* 60**,** 1389-98.

MASUNAGA, T., NIIZEKI, H., YASUDA, F., YOSHIDA, K., AMAGAI, M. & ISHIKO, A. 2015. Splicing abnormality of integrin β4 gene (ITGB4) due to nucleotide substitutions far from splice site underlies pyloric atresia-junctional epidermolysis bullosa syndrome. *J Dermatol Sci,* 78**,** 61-6.

MATERA, A. G. & WANG, Z. 2014. A day in the life of the spliceosome. *Nat Rev Mol Cell Biol,* 15**,** 108-21.

MATTICK, J. S. 2001. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep,* 2**,** 986-91.

MATTICK, J. S., DINGER, M., SCHONROCK, N. & COWLEY, M. 2018. Whole genome sequencing provides better diagnostic yield and future value than whole exome sequencing. *Med J Aust,* 209**,** 197-199.

MAYER, K., BALLHAUSEN, W., LEISTNER, W. & ROTT, H. 2000. Three novel types of splicing aberrations in the tuberous sclerosis TSC2 gene caused by mutations apart from splice consensus sequences. *Biochim Biophys Acta,* 1502**,** 495-507.

MCCARTHY, D. J., HUMBURG, P., KANAPIN, A., RIVAS, M. A., GAULTON, K., CAZIER, J. B. & DONNELLY, P. 2014. Choice of transcripts and software has a large effect on variant annotation. *Genome Med,* 6**,** 26.

MCDONALD, T. L., ZHOU, W., CASTRO, C. P., MUMM, C., SWITZENBERG, J. A., MILLS, R. E. & BOYLE, A. P. 2021. Cas9 targeted enrichment of mobile elements using nanopore sequencing. *Nat Commun,* 12, 3586.

MCKIE, A. B., MCHALE, J. C., KEEN, T. J., TARTTELIN, E. E., GOLIATH, R., VAN LITH-VERHOEVEN, J. J., GREENBERG, J., RAMESAR, R. S., HOYNG, C. B., CREMERS, F. P., MACKEY, D. A., BHATTACHARYA, S. S., BIRD, A. C., MARKHAM, A. F. & INGLEHEARN, C. F. 2001. Mutations in the pre-mRNA splicing factor gene PRPC8 in autosomal dominant retinitis pigmentosa (RP13). *Hum Mol Genet,* 10, 1555-62.

MCLAREN, W., GIL, L., HUNT, S. E., RIAT, H. S., RITCHIE, G. R., THORMANN, A., FLICEK, P. & CUNNINGHAM, F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol,* 17, 122.

MEHTA, A., BECK, M. & SUNDER-PLASSMANN, G. 2006. Fabry Disease: Perspectives from 5 Years of FOS.

MERCER, T. R., CLARK, M. B., ANDERSEN, S. B., BRUNCK, M. E., HAERTY, W., CRAWFORD, J., TAFT, R. J., NIELSEN, L. K., DINGER, M. E. & MATTICK, J. S. 2015. Genome-wide discovery of human splicing branchpoints. *Genome Res,* 25, 290-303.

MERENDINO, L., GUTH, S., BILBAO, D., MARTÍNEZ, C. & VALCÁRCEL, J. 1999. Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG. *Nature,* 402, 838-41.

MERICO, D., ROIFMAN, M., BRAUNSCHWEIG, U., YUEN, R. K., ALEXANDROVA, R., BATES, A., REID, B., NALPATHAMKALAM, T., WANG, Z., THIRUVAHINDRAPURAM, B., GRAY, P., KAKAKIOS, A., PEAKE, J., HOGARTH, S., MANSON, D., BUNCIC, R., PEREIRA, S. L., HERBRICK, J. A., BLENCOWE, B. J., ROIFMAN, C. M. & SCHERER, S. W. 2015. Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing. *Nat Commun,* 6, 8718.

MERKER, J. D., WENGER, A. M., SNEDDON, T., GROVE, M., ZAPPALA, Z., FRESARD, L., WAGGOTT, D., UTIRAMERUR, S., HOU, Y., SMITH, K. S., MONTGOMERY, S. B., WHEELER, M., BUCHAN, J. G., LAMBERT, C. C., ENG, K. S., HICKEY, L., KORLACH, J., FORD, J. & ASHLEY, E. A. 2018. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med,* 20, 159-163.

MERKER, J. D., WENGER, A. M., SNEDDON, T., GROVE, M., ZAPPALA, Z., FRESARD, L., WAGGOTT, D., UTIRAMERUR, S., HOU, Y., SMITH, K. S., MONTGOMERY, S. B., WHEELER, M., BUCHAN, J. G., LAMBERT, C. C., ENG, K. S., HICKEY, L., KORLACH, J., FORD, J. & ASHLEY, E. A. 2018. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med,* 20, 159-163.

MERTES, C., SCHELLER, I. F., YÉPEZ, V. A., ÇELIK, M. H., LIANG, Y., KREMER, L. S., GUSIC, M., PROKISCH, H. & GAGNEUR, J. 2021. Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat Commun,* 12, 529.

MESTER, J. L., GHOSH, R., PESARAN, T., HUETHER, R., KARAM, R., HRUSKA, K. S., COSTA, H. A., LACHLAN, K., NGEOW, J., BARNHOLTZ-SLOAN, J., SESOCK, K., HERNANDEZ, F., ZHANG, L., MILKO, L., PLON, S. E., HEGDE, M. & ENG, C. 2018. Gene-specific criteria for PTEN variant curation: Recommendations from the ClinGen PTEN Expert Panel. *Hum Mutat,* 39, 1581-1592.

MIDDLETON, R., GAO, D., THOMAS, A., SINGH, B., AU, A., WONG, J. J., BOMANE, A., COSSON, B., EYRAS, E., RASKO, J. E. & RITCHIE, W. 2017. IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol,* 18, 51.

MIURA, Y., MARDY, S., AWAYA, Y., NIHEI, K., ENDO, F., MATSUDA, I. & INDO, Y. 2000. Mutation and polymorphism analysis of the TRKA (NTRK1) gene encoding a high-affinity receptor for nerve growth factor in congenital insensitivity to pain with anhidrosis (CIPA) families. *Hum Genet,* 106, 116-24.

MONTALBAN, G., BONACHE, S., MOLES-FERNANDEZ, A., GISBERT-BEAMUD, A., TENES, A., BACH, V., CARRASCO, E., LOPEZ-FERNANDEZ, A., STJEPANOVIC, N., BALMANA, J., DIEZ, O. & GUTIERREZ-ENRIQUEZ, S. 2019. Screening of BRCA1/2 deep intronic regions by targeted gene sequencing identifies the first germline BRCA1 variant causing pseudoexon activation in a patient with breast/ovarian cancer. *J Med Genet,* 56, 63-74.

MORLAN, J. D., QU, K. & SINICROPI, D. V. 2012. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. *PLoS One,* 7, e42882.

MURDOCK, D. R., DAI, H., BURRAGE, L. C., ROSENFELD, J. A., KETKAR, S., MÜLLER, M. F., YÉPEZ, V. A., GAGNEUR, J., LIU, P., CHEN, S., JAIN, M., ZAPATA, G., BACINO, C. A., CHAO, H. T., MORETTI, P., CRAIGEN, W. J., HANCHARD, N. A., LEE, B. & NETWORK, U.

D. 2021. Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. *J Clin Invest,* 131.

MUSICH, R., CADLE-DAVIDSON, L. & OSIER, M. V. 2021. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. *Front Plant Sci,* 12**,** 657240.

NAGALAKSHMI, U., WANG, Z., WAERN, K., SHOU, C., RAHA, D., GERSTEIN, M. & SNYDER, M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science,* 320**,** 1344-9.

NARANJO, S., VOESENEK, K., DE LA CALLE-MUSTIENES, E., ROBERT-MORENO, A., KOKOTAS, H., GRIGORIADOU, M., ECONOMIDES, J., VAN CAMP, G., HILGERT, N., MORENO, F., ALSINA, B., PETERSEN, M. B., KREMER, H. & GÓMEZ-SKARMETA, J. L. 2010. Multiple enhancers located in a 1-Mb region upstream of POU3F4 promote expression during inner ear development and may be required for hearing. *Hum Genet,* 128**,** 411-9.

NARAYANASAMY, S., MARKINA, V., THOROGOOD, A., BLAZKOVA, A., SHABANI, M., KNOPPERS, B. M., PRAINSACK, B. & KOESTERS, R. 2020. Genomic Sequencing Capacity, Data Retention, and Personal Access to Raw Data in Europe. *Front Genet,* 11**,** 303.

NARO, C., JOLLY, A., DI PERSIO, S., BIELLI, P., SETTERBLAD, N., ALBERDI, A. J., VICINI, E., GEREMIA, R., DE LA GRANGE, P. & SETTE, C. 2017. An Orchestrated Intron Retention Program in Meiosis Controls Timely Usage of Transcripts during Germ Cell Differentiation. *Dev Cell,* 41**,** 82-93.e4.

NAVARRO, J. F., CROTEAU, D. L., JUREK, A., ANDRUSIVOVA, Z., YANG, B., WANG, Y., OGEDEGBE, B., RIAZ, T., STØEN, M., DESLER, C., RASMUSSEN, L. J., TØNJUM, T., GALAS, M. C., LUNDEBERG, J. & BOHR, V. A. 2020. Spatial Transcriptomics Reveals Genes Associated with Dysregulated Mitochondrial Functions and Stress Signaling in Alzheimer Disease. *iScience,* 23**,** 101556.

NAZARI, I., TAYARA, H. & CHONG, K. T. 2018. Branch Point Selection in RNA Splicing Using Deep Learning. *IEEE Access,* 7**,** 1800-7.

NG, P. C. & HENIKOFF, S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res,* 31**,** 3812-4.

NOMURA, S. 2021. Single-cell genomics to understand disease pathogenesis. *J Hum Genet,* 66**,** 75-84.

O'LEARY, N. A., WRIGHT, M. W., BRISTER, J. R., CIUFO, S., HADDAD, D., MCVEIGH, R., RAJPUT, B., ROBBERTSE, B., SMITH-WHITE, B., AKO-ADJEI, D., ASTASHYN, A., BADRETDIN, A., BAO, Y., BLINKOVA, O., BROVER, V., CHETVERNIN, V., CHOI, J., COX, E., ERMOLAEVA, O., FARRELL, C. M., GOLDFARB, T., GUPTA, T., HAFT, D., HATCHER, E., HLAVINA, W., JOARDAR, V. S., KODALI, V. K., LI, W., MAGLOTT, D., MASTERSON, P., MCGARVEY, K. M., MURPHY, M. R., O'NEILL, K., PUJAR, S., RANGWALA, S. H., RAUSCH, D., RIDDICK, L. D., SCHOCH, C., SHKEDA, A., STORZ, S. S., SUN, H., THIBAUD-NISSEN, F., TOLSTOY, I., TULLY, R. E., VATSAN, A. R., WALLIN, C., WEBB, D., WU, W., LANDRUM, M. J., KIMCHI, A., TATUSOVA, T., DICUCCIO, M., KITTS, P., MURPHY, T. D. & PRUITT, K. D. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res,* 44**,** D733-45.

OSBORNE, R. H., HOPPER, J. L., KIRK, J. A., CHENEVIX-TRENCH, G., THORNE, H. J. & SAMBROOK, J. F. 2000. kConFab: a research resource of Australasian breast cancer families. Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer. *Med J Aust,* 172**,** 463-4.

OSTERGAARD, E., DUNO, M., MØLLER, L. B., KALKANOGLU-SIVRI, H. S., DURSUN, A., ALIEFENDIOGLU, D., LETH, H., DAHL, M., CHRISTENSEN, E. & WIBRAND, F. 2013. Novel Mutations in the PC Gene in Patients with Type B Pyruvate Carboxylase Deficiency. *JIMD Rep,* 9**,** 1-5.

PAGANI, F., RAPONI, M. & BARALLE, F. E. 2005. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc Natl Acad Sci U S A,* 102**,** 6368-72.

PAGEL, K. A., ANTAKI, D., LIAN, A., MORT, M., COOPER, D. N., SEBAT, J., IAKOUCHEVA, L. M., MOONEY, S. D. & RADIVOJAC, P. 2019. Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome. *PLoS Comput Biol,* 15**,** e1007112.

PAGGI, J. M. & BEJERANO, G. 2018. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. *RNA,* 24**,** 1647-1658.

PALMER, E. E., SACHDEV, R., MACINTOSH, R., MELO, U. S., MUNDLOS, S., RIGHETTI, S., KANDULA, T., MINOCHE, A. E., PUTTICK, C., GAYEVSKIY, V., HESSON, L., IDRISOGLU, S., SHOUBRIDGE, C., THAI, M. H. N., DAVIS, R. L., DREW, A. P., SAMPAIO, H., ANDREWS, P. I., LAWSON, J., CARDAMONE, M., MOWAT, D., COLLEY, A., KUMMERFELD, S., DINGER, M. E., COWLEY, M. J., ROSCIOLI, T., BYE, A. & KIRK, E. 2021. Diagnostic Yield of Whole Genome Sequencing After Nondiagnostic Exome Sequencing or Gene Panel in Developmental and Epileptic Encephalopathies. *Neurology,* 96**,** e1770-e1782.

PAMUŁA-PIŁAT, J., TĘCZA, K., KALINOWSKA-HEROK, M. & GRZYBOWSKA, E. 2020. Genetic 3'UTR variations and clinical factors significantly contribute to survival prediction and clinical response in breast cancer patients. *Sci Rep,* 10**,** 5736.

PAUPER, M., KUCUK, E., WENGER, A. M., CHAKRABORTY, S., BAYBAYAN, P., KWINT, M., VAN DER SANDEN, B., NELEN, M. R., DERKS, R., BRUNNER, H. G., HOISCHEN, A., VISSERS, L. E. L. M. & GILISSEN, C. 2020. Long-read trio sequencing of individuals with unsolved intellectual disability. *Eur J Hum Genet*.

PERTEA, M., LIN, X. & SALZBERG, S. L. 2001. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res,* 29**,** 1185-90.

PETROVSKI, S., WANG, Q., HEINZEN, E. L., ALLEN, A. S. & GOLDSTEIN, D. B. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet,* 9**,** e1003709.

PINEDA, J. M. B. & BRADLEY, R. K. 2018. Most human introns are recognized via multiple and tissue-specific branchpoints. *Genes Dev,* 32**,** 577-591.

POLLARD, K. S., HUBISZ, M. J., ROSENBLOOM, K. R. & SIEPEL, A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res,* 20**,** 110-21.

POLLASTRO, P. & RAMPONE, S. 2002. HS^{3}D, a dataset of Homo sapiens splice regions, and its extraction procedure from a major public database. *Int. J. Mod. Phys. C,* 13**,** 1105-1117.

POPLIN, R., RUANO-RUBIO, V., DEPRISTO, M. A., FENNELL, T. J., CARNEIRO, M. O., VAN DER AUWERA, G. A., KLING, D. E., GAUTHIER, L. D., LEVY-MOONSHINE, A., ROAZEN, D., SHAKIR, K., THIBAULT, J., CHANDRAN, S., WHELAN, C., LEK, M., GABRIEL, S., DALY, M. J., NEALE, B., MACARTHUR, D. G. & BANKS, E. 2018. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*.

PRAY, L. A. 2008. DNA replication and causes of mutation. 1**,** 214.

PRUITT, K. D., TATUSOVA, T. & MAGLOTT, D. R. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res,* 35**,** D61-5.

PSORAKIS, I., ROBERTS, S. J., REZEK, I. & SHELDON, B. C. 2012. Inferring social network structure in ecological systems from spatio-temporal data streams. *J R Soc Interface,* 9**,** 3055-3056.

QIAO, Y., REN, C., HUANG, S., YUAN, J., LIU, X., FAN, J., LIN, J., WU, S., CHEN, Q., BO, X., LI, X., HUANG, X., LIU, Z. & SHU, W. 2020. High-resolution annotation of the mouse preimplantation embryo transcriptome using long-read sequencing. *Nat Commun,* 11**,** 2653.

RACZY, C., PETROVSKI, R., SAUNDERS, C. T., CHORNY, I., KRUGLYAK, S., MARGULIES, E. H., CHUANG, H. Y., KÄLLBERG, M., KUMAR, S. A., LIAO, A., LITTLE, K. M., STRÖMBERG, M. P. & TANNER, S. W. 2013. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics,* 29**,** 2041-3.

RAJ, B. & BLENCOWE, B. J. 2015. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. *Neuron,* 87**,** 14-27.

RANIERI, D., ROSATO, B., NANNI, M., MAGENTA, A., BELLEUDI, F. & TORRISI, M. R. 2016. Expression of the FGFR2 mesenchymal splicing variant in epithelial cells drives epithelial-mesenchymal transition. *Oncotarget,* 7**,** 5440-60.

RAY, T. A., COCHRAN, K., KOZLOWSKI, C., WANG, J., ALEXANDER, G., CADY, M. A., SPENCER, W. J., RUZYCKI, P. A., CLARK, B. S., LAEREMANS, A., HE, M. X., WANG, X., PARK, E., HAO, Y., IANNACCONE, A., HU, G., FEDRIGO, O., SKIBA, N. P., ARSHAVSKY, V. Y. & KAY, J. N. 2020. Comprehensive identification of mRNA isoforms reveals the diversity of neural cell-surface molecules with roles in retinal development and disease. *Nat Commun,* 11**,** 3328.

REED, R. 1989. The organization of 3' splice-site sequences in mammalian introns. *Genes Dev,* 3**,** 2113-23.

REESE, M. G., EECKMAN, F. H., KULP, D. & HAUSSLER, D. 1997. Improved splice site detection in Genie. *J Comput Biol,* 4**,** 311-23.

REESE, M. G., EECKMAN, F. H., KULP, D. & HAUSSLER, D. 1997. Improved splice site detection in Genie. *J Comput Biol,* 4, 311-23.

RENTZSCH, P., SCHUBACH, M., SHENDURE, J. & KIRCHER, M. 2021. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med,* 13, 31.

RENTZSCH, P., WITTEN, D., COOPER, G. M., SHENDURE, J. & KIRCHER, M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res,* 47, D886-D894.

RENTZSCH, P., WITTEN, D., COOPER, G. M., SHENDURE, J. & KIRCHER, M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res,* 47, D886-D894.

RICHARDS, A. J., MARTIN, S., YATES, J. R., SCOTT, J. D., BAGULEY, D. M., POPE, F. M. & SNEAD, M. P. 2000. COL2A1 exon 2 mutations: relevance to the Stickler and Wagner syndromes. *Br J Ophthalmol,* 84, 364-71.

RICHARDS, A. L., JONES, L., MOSKVINA, V., KIROV, G., GEJMAN, P. V., LEVINSON, D. F., SANDERS, A. R., PURCELL, S., VISSCHER, P. M., CRADDOCK, N., OWEN, M. J., HOLMANS, P., O'DONOVAN, M. C., (MGS), M. G. O. S. C. & (ISC), I. S. C. 2012. Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. *Mol Psychiatry,* 17, 193-201.

RICHARDS, S., AZIZ, N., BALE, S., BICK, D., DAS, S., GASTIER-FOSTER, J., GRODY, W. W., HEGDE, M., LYON, E., SPECTOR, E., VOELKERDING, K. & REHM, H. L. 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. *Genet Med,* 17, 405-24.

RIEDHAMMER, K. M., STOCKLER, S., PLOSKI, R., WENZEL, M., ADIS-DUTSCHMANN, B., AHTING, U., ALHADDAD, B., BLASCHEK, A., HAACK, T. B., KOPAJTICH, R., LEE, J., MURCIA PIENKOWSKI, V., POLLAK, A., SZYMANSKA, K., TARAILO-GRAOVAC, M., VAN DER LEE, R., VAN KARNEBEEK, C. D., MEITINGER, T., KRÄGELOH-MANN, I. & VILL, K. 2021. De novo stop-loss variants in CLDN11 cause hypomyelinating leukodystrophy. *Brain,* 144, 411-419.

RIEPE, T. V., KHAN, M., ROOSING, S., CREMERS, F. P. M. & 'T HOEN, P. A. C. 2021. Benchmarking deep learning splice prediction tools using functional splice assays. *Hum Mutat,* 42, 799-810.

RODERO, M. P. & CROW, Y. J. 2016. Type I interferon-mediated monogenic autoinflammation: The type I interferonopathies, a conceptual overview. *J Exp Med,* 213, 2527-2538.

RODRIGUEZ, J. M., POZO, F., DI DOMENICO, T., VAZQUEZ, J. & TRESS, M. L. 2020. An analysis of tissue-specific alternative splicing at the protein level. *PLoS Comput Biol,* 16, e1008287.

ROLLER, E., IVAKHNO, S., LEE, S., ROYCE, T. & TANNER, S. 2016. Canvas: versatile and scalable detection of copy number variants. *bioRxiv*.

ROSENBERG, A. B., PATWARDHAN, R. P., SHENDURE, J. & SEELIG, G. 2015. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell,* 163, 698-711.

ROSENTHAL, A., JOUET, M. & KENWRICK, S. 1992. Aberrant splicing of neural cell adhesion molecule L1 mRNA in a family with X-linked hydrocephalus. *Nat Genet,* 2, 107-12.

ROTHROCK, C. R., HOUSE, A. E. & LYNCH, K. W. 2005. HnRNP L represses exon splicing via a regulated exonic splicing silencer. *EMBO J,* 24, 2792-802.

ROWLANDS, C. F., BARALLE, D. & ELLINGFORD, J. M. 2019. Machine Learning Approaches for the Prioritization of Genomic Variants Impacting Pre-mRNA Splicing. *Cells,* 8.

ROWLANDS, C. F., THOMAS, H. B., LORD, J., WAI, H., ARNO, G., BEAMAN, G., SERGOUNIOTIS, P., GOMES-SILVA, B., CAMPBELL, C., GOSSAN, N., HARDCASTLE, C., WEBB, K., O'CALLAGHAN, C., HIRST, R., RAMSDEN, S., JONES, E., CLAYTON-SMITH, J., WEBSTER, A., DOUGLAS, A. G. L., O'KEEFE, R. T., NEWMAN, W. G., BARALLE, D., BLACK, G. C. M. & ELLINGFORD, J. 2021. Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. *Sci Rep*, 11:20607.

RUSKIN, B. & GREEN, M. R. 1985. Role of the 3' splice site consensus sequence in mammalian pre-mRNA splicing. *Nature,* 317, 732-4.

SANGERMANO, R., GARANTO, A., KHAN, M., RUNHART, E. H., BAUWENS, M., BAX, N. M., VAN DEN BORN, L. I., KHAN, M. I., CORNELIS, S. S., VERHEIJ, J., POTT, J. R., THIADENS, A., KLAVER, C. C. W., PUECH, B., MEUNIER, I., NAESSENS, S., ARNO, G., FAKIN, A., CARSS, K. J., RAYMOND, F. L., WEBSTER, A. R., DHAENENS, C. M., STOHR, H., GRASSMANN, F., WEBER, B. H. F., HOYNG, C. B., DE BAERE, E., ALBERT, S., COLLIN, R. W. J. & CREMERS, F. P. M. 2019. Deep-intronic ABCA4 variants explain missing heritability in Stargardt disease and allow correction of splice defects by antisense oligonucleotides. *Genet Med,* 21**,** 1751-1760.

SARGIANNIDOU, I., KIM, G. H., KYRIAKOUDI, S., EUN, B. L. & KLEOPA, K. A. 2015. A start codon CMT1X mutation associated with transient encephalomyelitis causes complete loss of Cx32. *Neurogenetics,* 16**,** 193-200.

SATOH, M. & KUROIWA, T. 1991. Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. *Exp Cell Res,* 196**,** 137-40.

SCHNEIDER, C., WILL, C. L., MAKAROVA, O. V., MAKAROV, E. M. & LÜHRMANN, R. 2002. Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions. *Mol Cell Biol,* 22**,** 3219-29.

SCHNEIDER, W. M., CHEVILLOTTE, M. D. & RICE, C. M. 2014. Interferon-stimulated genes: a complex web of host defenses. *Annu Rev Immunol,* 32**,** 513-45.

SCHOCH, K., TAN, Q. K., STONG, N., DEAK, K. L., MCCONKIE-ROSELL, A., MCDONALD, M. T., GOLDSTEIN, D. B., JIANG, Y. H., SHASHI, V. & NETWORK, U. D. 2020. Alternative transcripts in variant interpretation: the potential for missed diagnoses and misdiagnoses. *Genet Med,* 22**,** 1269-1275.

SCOTTI, M. M. & SWANSON, M. S. 2016. RNA mis-splicing in disease. *Nat Rev Genet,* 17**,** 19-32.

SHAO, W., KIM, H. S., CAO, Y., XU, Y. Z. & QUERY, C. C. 2012. A U1-U2 snRNP interaction network during intron definition. *Mol Cell Biol,* 32**,** 470-8.

SHAW, M. A., BRUNETTI-PIERRI, N., KÁDASI, L., KOVÁCOVÁ, V., VAN MALDERGEM, L., DE BRASI, D., SALERNO, M. & GÉCZ, J. 2003. Identification of three novel SEDL mutations, including mutation in the rare, non-canonical splice site of exon 4. *Clin Genet,* 64**,** 235-42.

SHEN, S., PARK, J. W., LU, Z. X., LIN, L., HENRY, M. D., WU, Y. N., ZHOU, Q. & XING, Y. 2014. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A,* 111**,** E5593-601.

SHEN, X. & COREY, D. R. 2018. Chemistry, mechanism and clinical status of antisense oligonucleotides and duplex RNAs. *Nucleic Acids Res,* 46**,** 1584-1600.

SHERRY, S. T., WARD, M. H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res,* 29**,** 308-11.

SHETH, N., ROCA, X., HASTINGS, M. L., ROEDER, T., KRAINER, A. R. & SACHIDANANDAM, R. 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res,* 34**,** 3955-67.

SHOFFNER, J. M., LOTT, M. T., LEZZA, A. M., SEIBEL, P., BALLINGER, S. W. & WALLACE, D. C. 1990. Myoclonic epilepsy and ragged-red fiber disease (MERRF) is associated with a mitochondrial DNA tRNA(Lys) mutation. *Cell,* 61**,** 931-7.

SHORT, P. J., MCRAE, J. F., GALLONE, G., SIFRIM, A., WON, H., GESCHWIND, D. H., WRIGHT, C. F., FIRTH, H. V., FITZPATRICK, D. R., BARRETT, J. C. & HURLES, M. E. 2018. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature,* 555**,** 611-616.

SHU, X., BLACK, G. C., RICE, J. M., HART-HOLDEN, N., JONES, A., O'GRADY, A., RAMSDEN, S. & WRIGHT, A. F. 2007. RPGR mutation analysis and disease: an update. *Hum Mutat,* 28**,** 322-8.

SIEPEL, A., BEJERANO, G., PEDERSEN, J. S., HINRICHS, A. S., HOU, M., ROSENBLOOM, K., CLAWSON, H., SPIETH, J., HILLIER, L. W., RICHARDS, S., WEINSTOCK, G. M., WILSON, R. K., GIBBS, R. A., KENT, W. J., MILLER, W. & HAUSSLER, D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res,* 15**,** 1034-50.

SIGNAL, B., GLOSS, B. S., DINGER, M. E. & MERCER, T. R. 2018. Machine learning annotation of human branchpoints. *Bioinformatics,* 34**,** 920-927.

SINNREICH, M., THERRIEN, C. & KARPATI, G. 2006. Lariat branch point mutation in the dysferlin gene with mild limb-girdle muscular dystrophy. *Neurology,* 66**,** 1114-6.

SITTLER, A., DEVYS, D., WEBER, C. & MANDEL, J. L. 1996. Alternative splicing of exon 14 determines nuclear or cytoplasmic localisation of fmr1 protein isoforms. *Hum Mol Genet,* 5**,** 95-102.

SMITH, C. W., CHU, T. T. & NADAL-GINARD, B. 1993. Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Mol Cell Biol,* 13**,** 4939-52.

SMITH, S. A. & LYNCH, K. W. 2014. Cell-based splicing of minigenes. *Methods Mol Biol,* 1126**,** 243-55.

SOEMEDI, R., CYGAN, K. J., RHINE, C. L., WANG, J., BULACAN, C., YANG, J., BAYRAK-TOYDEMIR, P., MCDONALD, J. & FAIRBROTHER, W. G. 2017. Pathogenic variants that alter protein code often disrupt splicing. *Nat Genet,* 49**,** 848-855.

STAKNIS, D. & REED, R. 1994. SR proteins promote the first specific recognition of Pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex. *Mol Cell Biol,* 14**,** 7670-82.

STENSON, P. D., MORT, M., BALL, E. V., SHAW, K., PHILLIPS, A. D. & COOPER, D. N. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics,* 133**,** 1-9.

STEWART, J. B. & CHINNERY, P. F. 2015. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat Rev Genet,* 16**,** 530-42.

TAANMAN, J. W. 1999. The mitochondrial genome: structure, transcription, translation and replication. *Biochim Biophys Acta,* 1410**,** 103-23.

TAGGART, A. J., DESIMONE, A. M., SHIH, J. S., FILLOUX, M. E. & FAIRBROTHER, W. G. 2012. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat Struct Mol Biol,* 19**,** 719-21.

TAGGART, A. J., LIN, C. L., SHRESTHA, B., HEINTZELMAN, C., KIM, S. & FAIRBROTHER, W. G. 2017. Large-scale analysis of branchpoint usage across species and cell lines. *Genome Res,* 27**,** 639-649.

TAKATA, A., MATSUMOTO, N. & KATO, T. 2017. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat Commun,* 8**,** 14519.

TAVTIGIAN, S. V., GREENBLATT, M. S., HARRISON, S. M., NUSSBAUM, R. L., PRABHU, S. A., BOUCHER, K. M., BIESECKER, L. G. & CLINGEN SEQUENCE VARIANT INTERPRETATION WORKING, G. 2018. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med,* 20**,** 1054-1060.

TAYLOR, J. C., MARTIN, H. C., LISE, S., BROXHOLME, J., CAZIER, J. B., RIMMER, A., KANAPIN, A., LUNTER, G., FIDDY, S., ALLAN, C., ARICESCU, A. R., ATTAR, M., BABBS, C., BECQ, J., BEESON, D., BENTO, C., BIGNELL, P., BLAIR, E., BUCKLE, V. J., BULL, K., CAIS, O., CARIO, H., CHAPEL, H., COPLEY, R. R., CORNALL, R., CRAFT, J., DAHAN, K., DAVENPORT, E. E., DENDROU, C., DEVUYST, O., FENWICK, A. L., FLINT, J., FUGGER, L., GILBERT, R. D., GORIELY, A., GREEN, A., GREGER, I. H., GROCOCK, R., GRUSZCZYK, A. V., HASTINGS, R., HATTON, E., HIGGS, D., HILL, A., HOLMES, C., HOWARD, M., HUGHES, L., HUMBURG, P., JOHNSON, D., KARPE, F., KINGSBURY, Z., KINI, U., KNIGHT, J. C., KROHN, J., LAMBLE, S., LANGMAN, C., LONIE, L., LUCK, J., MCCARTHY, D., MCGOWAN, S. J., MCMULLIN, M. F., MILLER, K. A., MURRAY, L., NEMETH, A. H., NESBIT, M. A., NUTT, D., ORMONDROYD, E., OTURAI, A. B., PAGNAMENTA, A., PATEL, S. Y., PERCY, M., PETOUSI, N., PIAZZA, P., PIRET, S. E., POLANCO-ECHEVERRY, G., POPITSCH, N., POWRIE, F., PUGH, C., QUEK, L., ROBBINS, P. A., ROBSON, K., RUSSO, A., SAHGAL, N., VAN SCHOUWENBURG, P. A., SCHUH, A., SILVERMAN, E., SIMMONS, A., SORENSEN, P. S., SWEENEY, E., TAYLOR, J., THAKKER, R. V., TOMLINSON, I., TREBES, A., TWIGG, S. R., UHLIG, H. H., VYAS, P., VYSE, T., WALL, S. A., WATKINS, H., WHYTE, M. P., WITTY, L., et al. 2015. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet*.

TIPPIN, B., PHAM, P. & GOODMAN, M. F. 2004. Error-prone replication for better or worse. *Trends Microbiol,* 12**,** 288-95.

TOLLNER, T. L., VENNERS, S. A., HOLLOX, E. J., YUDIN, A. I., LIU, X., TANG, G., XING, H., KAYS, R. J., LAU, T., OVERSTREET, J. W., XU, X., BEVINS, C. L. & CHERR, G. N. 2011. A common mutation in the defensin DEFB126 causes impaired sperm function and subfertility. *Sci Transl Med,* 3**,** 92ra65.

TREANGEN, T. J. & SALZBERG, S. L. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet,* 13**,** 36-46.

TURNBULL, C., SCOTT, R. H., THOMAS, E., JONES, L., MURUGAESU, N., PRETTY, F. B., HALAI, D., BAPLE, E., CRAIG, C., HAMBLIN, A., HENDERSON, S., PATCH, C., O'NEILL, A., DEVEREAU, SMITH, K., MARTIN, A. R., SOSINSKY, A., MCDONAGH, E. M., SULTANA, R., MUELLER, M., SMEDLEY, D., TOMS, A., DINH, L., FOWLER, T., BALE, M., HUBBARD, T., RENDON, A., HILL, S., CAULFIELD, M. J. & PROJECT, G. 2018. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ,* 361**,** k1687.

TURNPENNY, P., ELLARD, S. & CLEAVER, R. 2017. Common Disease, Polygenic and Multifactorial Genetics. *Emery's Elements of Medical Genetics.* 15 ed.: Elsevier.

TURUNEN, J. J., NIEMELÄ, E. H., VERMA, B. & FRILANDER, M. J. 2013. The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA,* 4**,** 61-76.

UDDIN, S. A., CESARATO, N., HUMBATOVA, A., SCHMIDT, A., URREHMAN, F., NAEEM, M., TAREEN, A. S., WOLF, S., PANEZAI, M. A., THIELE, H., WALI, A., FÖLSTER-HOLST, R., BASIT, S., AYUB, M. & BETZ, R. C. 2020. Apparent Missense Variant in COL7A1 Causes a Severe Form of Recessive Dystrophic Epidermolysis Bullosa via Effects on Splicing. *Acta Derm Venereol,* 100**,** adv00275.

VALLETTA, J. J., TORNEY, C., KINGS, M., THORNTON, A. & MADDEN, J. 2017. Applications of machine learning in animal behaviour studies. *Animal Behav,* 124**,** 203-220.

VARGA, R. E., SCHÜLE, R., FADEL, H., VALENZUELA, I., SPEZIANI, F., GONZALEZ, M., RUDENSKAIA, G., NÜRNBERG, G., THIELE, H., ALTMÜLLER, J., ALVAREZ, V., GAMEZ, J., GARBERN, J. Y., NÜRNBERG, P., ZUCHNER, S. & BEETZ, C. 2013. Do not trust the pedigree: reduced and sex-dependent penetrance at a novel mutation hotspot in ATL1 blurs autosomal dominant inheritance of spastic paraplegia. *Hum Mutat,* 34**,** 860-3.

VERBAKEL, S. K., FADAIE, Z., KLEVERING, B. J., VAN GENDEREN, M. M., FEENSTRA, I., CREMERS, F. P. M., HOYNG, C. B. & ROOSING, S. 2019. The identification of a RNA splice variant in TULP1 in two siblings with early-onset photoreceptor dystrophy. *Mol Genet Genomic Med,* 7**,** e660.

VERMA, B., AKINYI, M. V., NORPPA, A. J. & FRILANDER, M. J. 2018. Minor spliceosome and disease. *Semin Cell Dev Biol,* 79**,** 103-112.

VERVOORT, R., LENNON, A., BIRD, A. C., TULLOCH, B., AXTON, R., MIANO, M. G., MEINDL, A., MEITINGER, T., CICCODICOLA, A. & WRIGHT, A. F. 2000. Mutational hot spot within a new RPGR exon in X-linked retinitis pigmentosa. *Nat Genet,* 25**,** 462-6.

VIG, A., POULTER, J. A., OTTAVIANI, D., TAVARES, E., TOROPOVA, K., TRACEWSKA, A. M., MOLLICA, A., KANG, J., KEHELWATHUGODA, O., PATON, T., MAYNES, J. T., WHEWAY, G., ARNO, G., KHAN, K. N., MCKIBBIN, M., TOOMES, C., ALI, M., DI SCIPIO, M., LI, S., ELLINGFORD, J., BLACK, G., WEBSTER, A., RYDZANICZ, M., STAWIŃSKI, P., PŁOSKI, R., VINCENT, A., CHEETHAM, M. E., INGLEHEARN, C. F., ROBERTS, A., HEON, E. & CONSORTIUM, G. E. R. 2020. DYNC2H1 hypomorphic or retina-predominant variants cause nonsyndromic retinal degeneration. *Genet Med,* 22**,** 2041-2051.

VITHANA, E. N., ABU-SAFIEH, L., ALLEN, M. J., CAREY, A., PAPAIOANNOU, M., CHAKAROVA, C., AL-MAGHTHEH, M., EBENEZER, N. D., WILLIS, C., MOORE, A. T., BIRD, A. C., HUNT, D. M. & BHATTACHARYA, S. S. 2001. A human homolog of yeast pre-mRNA splicing gene, PRP31, underlies autosomal dominant retinitis pigmentosa on chromosome 19q13.4 (RP11). *Mol Cell,* 8**,** 375-81.

VOLPI, S., PICCO, P., CAORSI, R., CANDOTTI, F. & GATTORNO, M. 2016. Type I interferonopathies in pediatric rheumatology. *Pediatr Rheumatol Online J,* 14**,** 35.

WAI, H. A., LORD, J., LYON, M., GUNNING, A., KELLY, H., CIBIN, P., SEABY, E. G., SPIERS-FITZGERALD, K., LYE, J., ELLARD, S., THOMAS, N. S., BUNYAN, D. J., DOUGLAS, A. G. L., BARALLE, D., WORKING, S. A. D. & GROUP 2020a. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet Med,* 22**,** 1005-1014.

WANG, K., LI, M. & HAKONARSON, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res,* 38**,** e164.

WANG, K., WU, D., ZHANG, H., DAS, A., BASU, M., MALIN, J., CAO, K. & HANNENHALLI, S. 2018. Comprehensive map of age-associated splicing changes across human tissues and their contributions to age-associated diseases. *Sci Rep,* 8**,** 10929.

WANG, S., PENG, J., MA, J. & XU, J. 2016. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci Rep,* 6**,** 18962.

WANG, Z., GERSTEIN, M. & SNYDER, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet,* 10**,** 57-63.

WEISSCHUH, N., BUENA-ATIENZA, E. & WISSINGER, B. 2021. Splicing mutations in inherited retinal diseases. *Prog Retin Eye Res,* 80**,** 100874.

WESTRA, H. J. & FRANKE, L. 2014. From genome to function by studying eQTLs. *Biochim Biophys Acta,* 1842**,** 1896-1902.

WHIFFIN, N., KARCZEWSKI, K. J., ZHANG, X., CHOTHANI, S., SMITH, M. J., EVANS, D. G., ROBERTS, A. M., QUAIFE, N. M., SCHAFER, S., RACKHAM, O., ALFÖLDI, J., O'DONNELL-LURIA, A. H., FRANCIOLI, L. C., COOK, S. A., BARTON, P. J. R., MACARTHUR, D. G., WARE, J. S., TEAM, G. A. D. P. & CONSORTIUM, G. A. D. 2020. Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *Nat Commun,* 11**,** 2523.

WILL, C. L. & LÜHRMANN, R. 2011. Spliceosome structure and function. *Cold Spring Harb Perspect Biol,* 3.

WIMMER, K., SCHAMSCHULA, E., WERNSTEDT, A., TRAUNFELLNER, P., AMBERGER, A., ZSCHOCKE, J., KROISEL, P., CHEN, Y., CALLENS, T. & MESSIAEN, L. 2020. AG-exclusion zone revisited: Lessons to learn from 91 intronic NF1 3' splice site mutations outside the canonical AG-dinucleotides. *Hum Mutat,* 41**,** 1145-1156.

WONG, J. J., RITCHIE, W., EBNER, O. A., SELBACH, M., WONG, J. W., HUANG, Y., GAO, D., PINELLO, N., GONZALEZ, M., BAIDYA, K., THOENG, A., KHOO, T. L., BAILEY, C. G., HOLST, J. & RASKO, J. E. 2013. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell,* 154**,** 583-95.

WOOD, K. A., ROWLANDS, C. F., QURESHI, W. M. S., THOMAS, H. B., BUCZEK, W. A., BRIGGS, T. A., HUBBARD, S. J., HENTGES, K. E., NEWMAN, W. G. & O'KEEFE, R. T. 2019. Disease modeling of core pre-mRNA splicing factor haploinsufficiency. *Hum Mol Genet,* 28**,** 3704-3723.

WOOD, K. A., ROWLANDS, C. F., THOMAS, H. B., WOODS, S., O'FLAHERTY, J., DOUZGOU, S., KIMBER, S. J., NEWMAN, W. G. & O'KEEFE, R. T. 2020. Modelling the developmental spliceosomal craniofacial disorder Burn-McKeown syndrome using induced pluripotent stem cells. *PLoS One,* 15**,** e0233582.

WU, S., ROMFO, C. M., NILSEN, T. W. & GREEN, M. R. 1999. Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature,* 402**,** 832-5.

XIONG, H. Y., ALIPANAHI, B., LEE, L. J., BRETSCHNEIDER, H., MERICO, D., YUEN, R. K., HUA, Y., GUEROUSSOV, S., NAJAFABADI, H. S., HUGHES, T. R., MORRIS, Q., BARASH, Y., KRAINER, A. R., JOJIC, N., SCHERER, S. W., BLENCOWE, B. J. & FREY, B. J. 2015. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science,* 347**,** 1254806.

YANG, H. & WANG, K. 2015. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc,* 10**,** 1556-66.

YANG, Y., MUZNY, D. M., REID, J. G., BAINBRIDGE, M. N., WILLIS, A., WARD, P. A., BRAXTON, A., BEUTEN, J., XIA, F., NIU, Z., HARDISON, M., PERSON, R., BEKHEIRNIA, M. R., LEDUC, M. S., KIRBY, A., PHAM, P., SCULL, J., WANG, M., DING, Y., PLON, S. E., LUPSKI, J. R., BEAUDET, A. L., GIBBS, R. A. & ENG, C. M. 2013. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med,* 369**,** 1502-11.

YANG, Y., MUZNY, D. M., XIA, F., NIU, Z., PERSON, R., DING, Y., WARD, P., BRAXTON, A., WANG, M., BUHAY, C., VEERARAGHAVAN, N., HAWES, A., CHIANG, T., LEDUC, M., BEUTEN, J., ZHANG, J., HE, W., SCULL, J., WILLIS, A., LANDSVERK, M., CRAIGEN, W. J., BEKHEIRNIA, M. R., STRAY-PEDERSEN, A., LIU, P., WEN, S., ALCARAZ, W., CUI, H., WALKIEWICZ, M., REID, J., BAINBRIDGE, M., PATEL, A., BOERWINKLE, E., BEAUDET, A. L., LUPSKI, J. R., PLON, S. E., GIBBS, R. A. & ENG, C. M. 2014. Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. *JAMA*.

YEH, F. L., CHANG, S. L., AHMED, G. R., LIU, H. I., TUNG, L., YEH, C. S., LANIER, L. S., MAEDER, C., LIN, C. M., TSAI, S. C., HSIAO, W. Y., CHANG, W. H. & CHANG, T. H. 2021. Activation of Prp28 ATPase by phosphorylated Npl3 at a critical step of spliceosome remodeling. *Nat Commun,* 12**,** 3082.

YEO, G. & BURGE, C. B. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol,* 11**,** 377-94.

YOSHIDA, K., NAKAMURA, A., YAZAKI, M., IKEDA, S. & TAKEDA, S. 1998. Insertional mutation by transposable element, L1, in the DMD gene results in X-linked dilated cardiomyopathy. *Hum Mol Genet,* 7**,** 1129-32.

YUEN, R. K., THIRUVAHINDRAPURAM, B., MERICO, D., WALKER, S., TAMMIMIES, K., HOANG, N., CHRYSLER, C., NALPATHAMKALAM, T., PELLECCHIA, G., LIU, Y., GAZZELLONE, M. J., D'ABATE, L., DENEAULT, E., HOWE, J. L., LIU, R. S., THOMPSON, A., ZARREI, M., UDDIN, M., MARSHALL, C. R., RING, R. H., ZWAIGENBAUM, L., RAY, P. N., WEKSBERG, R., CARTER, M. T., FERNANDEZ, B. A., ROBERTS, W., SZATMARI, P. & SCHERER, S. W.

2015. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat Med,* 21**,** 185-91.

ZENG, Z. & BROMBERG, Y. 2019. Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives. *Front Genet,* 10**,** 914.

ZHANG, Q., FAN, X., WANG, Y., SUN, M. A., SHAO, J. & GUO, D. 2017. BPP: a sequence-based algorithm for branch point prediction. *Bioinformatics,* 33**,** 3166-3172.

ZHANG, X., WAKELING, M., WARE, J. & WHIFFIN, N. 2020. Annotating high-impact 5'untranslated region variants with the UTRannotator. *Bioinformatics*.

ZHANG, Z., RIGO, N., DYBKOV, O., FOURMANN, J. B., WILL, C. L., KUMAR, V., URLAUB, H., STARK, H. & LÜHRMANN, R. 2021. Structural insights into how Prp5 proofreads the pre-mRNA branch site. *Nature,* 596**,** 296-300.

ZHOU, X. P., WAITE, K. A., PILARSKI, R., HAMPEL, H., FERNANDEZ, M. J., BOS, C., DASOUKI, M., FELDMAN, G. L., GREENBERG, L. A., IVANOVICH, J., MATLOFF, E., PATTERSON, A., PIERPONT, M. E., RUSSO, D., NASSIF, N. T. & ENG, C. 2003. Germline PTEN promoter mutations and deletions in Cowden/Bannayan-Riley-Ruvalcaba syndrome result in aberrant PTEN protein and dysregulation of the phosphoinositol-3-kinase/Akt pathway. *Am J Hum Genet,* 73**,** 404-11.

ZHU, C., WU, J., SUN, H., BRIGANTI, F., MEDER, B., WEI, W. & STEINMETZ, L. M. 2021. Single-molecule, full-length transcript isoform sequencing reveals disease-associated RNA isoforms in cardiomyocytes. *Nat Commun,* 12**,** 4203.

ZHU, J., MAYEDA, A. & KRAINER, A. R. 2001. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol Cell,* 8**,** 1351-61.

ZORIO, D. A. & BLUMENTHAL, T. 1999. Both subunits of U2AF recognize the 3' splice site in Caenorhabditis elegans. *Nature,* 402**,** 835-8.

Appendix 1 – Supplementary Tables

Supplementary Table S1. *List of functionally assayed variants for comparison of splice prediction tools.* 249 functionally assayed variants are listed in order of chromosomal location (positions are given according to the GRCh37 genome build). The impact of each variant is given in the “Outcome” column, with TP representing true negatives - i.e. variants for which no significant impact on splicing was observed - and TP representing true positives, variants that appear to significantly disrupt splicing in the employed assay.

HGVSc	Gene symbol	Chromosome	Position	Ref	Alt	Splice region	Assay type	Outcome
NM_014874.3:c.838C>T	*MFN2*	chr1	12061479	C	T	exonic	RT-PCR	TN
NM_001146289.1:c.1224-80G>A	*P3H1 (LEPRE1)*	chr1	43220741	C	T	deep intronic	RT-PCR + RNA-seq	TP
NM_000350.2:c.5584+6T>C	*ABCA4*	chr1	94476812	A	G	5′ extended	Minigene	TP
NM_005850.4:c.417C>T	*SF3B4*	chr1	149898557	G	A	exonic	RT-PCR + RNA-seq	TP
NM_019032.5:c.2559G>A	*ADAMTSL4*	chr1	150531125	G	A	exonic	RT-PCR	TP
NM_206933.2:c.14343+36C>G	*USH2A*	chr1	215823898	G	C	5′ extended	Minigene	TN
NM_001011.3:c.507+3A>G	*RPS7*	chr2	3627853	A	G	5′ intronic	RT-PCR	TN
NM_000179.2:c.806C>G	*MSH6*	chr2	48025928	C	G	exonic	RT-PCR	TN
NM_000179.2:c.3416G>A	*MSH6*	chr2	48030802	G	A	exonic	RT-PCR	TN
NM_000179.2:c.3439-16C>T	*MSH6*	chr2	48032033	C	T	3′ intronic	RT-PCR	TN
NM_006343.2:c.2486+6T>A	*MERTK*	chr2	112779977	T	A	5′ extended	Minigene	TP
NM_001040142.1:c.2919+3A>G	*SCN2A*	chr2	166201424	A	G	5′ intronic	Minigene	TP
NM_000090.3:c.1815+5G>A	*COL3A1*	chr2	189861949	G	A	5′ extended	RT-PCR	TN
NM_000090.3:c.3133G>A	*COL3A1*	chr2	189871110	G	A	exonic	RT-PCR	TN
NM_018297.3:c.930C>T	*NGLY1*	chr3	25778898	G	A	exonic	RT-PCR	TP
NM_000249.3:c.80G>A	*MLH1*	chr3	37035118	G	A	exonic	RT-PCR	TN
NM_000249.3:c.122A>G	*MLH1*	chr3	37038115	A	G	exonic	RT-PCR	TP
NM_000249.3:c.935A>C	*MLH1*	chr3	37061851	A	C	exonic	RT-PCR	TN
NM_000249.3:c.1989+6T>G	*MLH1*	chr3	37090106	T	G	5′ extended	RT-PCR	TN
NM_052985.3:c.3039+4A>G	*IFT122*	chr3	129226609	A	G	5′ extended	Minigene	TP
NM_000283.3:c.2130-15G>A	*PDE6B*	chr4	658655	G	A	3′ intronic	Minigene	TP
NM_002890.2:c.2011+6T>G	*RASA1*	chr5	86670739	T	G	5′ extended	RT-PCR	TP
NM_002397.4:c.835-9T>G	*MEF2C*	chr5	88025173	A	C	3′ intronic	RT-PCR	TP
NM_001354896.1:c.295C>T	*APC*	chr5	112102960	C	T	exonic	RT-PCR	TN
NM_000038.5:c.1549-8A>G	*APC*	chr5	112163618	A	G	3′ intronic	RT-PCR	TP
NM_001999.3:c.4594+3A>G	*FBN2*	chr5	127654568	T	C	5′ intronic	RT-PCR	TP
NM_080680.2:c.2682G>A	*COL11A2*	chr6	33141279	C	T	exonic	RT-PCR	TN

NM_001142800.1:c.6571+4558A>G	*EYS*	chr6	64787191	T	C	deep intronic	Minigene	TN
NM_006208.2:c.241G>T	*ENPP1*	chr6	132168916	G	T	exonic	RT-PCR	TP
NM_001277115.1:c.6547-963G>A	*DNAH11*	chr7	21746354	G	A	deep intronic	Minigene	TP
NM_000492.3:c.3874-4522A>G	*CFTR*	chr7	117288374	A	G	deep intronic	Minigene	TP
NM_001174067.1:c.1029G>A	*FGFR1*	chr8	38282027	C	T	exonic	RT-PCR	TP
NM_017890.4:c.7226C>T	*VPS13B*	chr8	100779102	C	T	exonic	RT-PCR	TN
NM_000264.4:c.2704-11C>T	*PTCH1*	chr9	98222076	G	A	3′ intronic	RT-PCR	TN
NM_001306210.1:c.574G>C	*TGFBR1*	chr9	101895009	G	C	exonic	RT-PCR	TP
NM_000368.4:c.211-7T>G	*TSC1*	chr9	135801133	A	C	3′ intronic	RT-PCR	TP
NM_000093.3:c.1195G>A	*COL5A1*	chr9	137623372	G	A	exonic	RT-PCR	TN
NM_000314.4:c.253G>C	*PTEN*	chr10	89690846	G	C	exonic	RT-PCR	TP
NM_000314.4:c.373A>G	*PTEN*	chr10	89692889	A	G	exonic	RT-PCR	TN
NM_000314.4:c.553C>G	*PTEN*	chr10	89711935	C	G	exonic	RT-PCR	TN
NM_000314.4:c.593T>C	*PTEN*	chr10	89711975	T	C	exonic	RT-PCR	TN
NM_000314.4:c.830C>T	*PTEN*	chr10	89720679	C	T	exonic	RT-PCR	TN
NM_006204.3:c.1072-11T>C	*PDE6C*	chr10	95389004	T	C	3′ intronic	Minigene	TN
NM_000256.3:c.3815-10T>G	*MYBPC3*	chr11	47353442	A	C	3′ intronic	RT-PCR	TN
NM_000256.3:c.1624+4A>T	*MYBPC3*	chr11	47364125	T	A	5′ extended	RT-PCR	TP
NM_000256.3:c.1457+5G>C	*MYBPC3*	chr11	47364376	C	G	5′ extended	RT-PCR	TP
NM_000256.3:c.1224-21A>G	*MYBPC3*	chr11	47364834	T	C	3′ intronic	RT-PCR	TP
NM_130799.2:c.1050-3C>G	*MEN1*	chr11	64573245	G	C	3′ intronic	RT-PCR	TP
NM_024649.4:c.592-21A>T	*BBS1*	chr11	66287067	A	T	3′ intronic	Minigene	TN
NM_006946.2:c.4150C>A	*SPTBN2*	chr11	66463876	G	T	exonic	RT-PCR	TN
NM_007103.3:c.1080G>A	*NDUFV1*	chr11	67379040	G	A	exonic	RT-PCR	TP
NM_002335.2:c.1413-7T>A	*LRP5*	chr11	68157342	T	A	3′ intronic	Minigene	TN
NM_016401.3:c.539+3A>G	*HIKESHI (C11orf73)*	chr11	86055766	A	G	5′ extended	RT-PCR	TP
NM_003002.2:c.314+5G>A	*SDHD*	chr11	111959740	G	A	5′ extended	RT-PCR	TP
NM_000059.3:c.79A>G	*BRCA2*	chr13	32893225	A	G	exonic	RT-PCR	TN

NM_000059.3:c.167A>C	*BRCA2*	chr13	32893313	A	C	exonic	RT-PCR	TN
NM_000059.3:c.223G>C	*BRCA2*	chr13	32893369	G	C	exonic	RT-PCR	TN
NM_000059.3:c.280C>T	*BRCA2*	chr13	32893426	C	T	exonic	RT-PCR	TN
NM_000059.3:c.441A>G	*BRCA2*	chr13	32900253	A	G	exonic	RT-PCR	TN
NM_000059.3:c.506A>G	*BRCA2*	chr13	32900409	A	G	exonic	RT-PCR + RNA-seq	TN
NM_000059.3:c.520C>T	*BRCA2*	chr13	32900639	C	T	exonic	RT-PCR	TP
NM_000059.3:c.598A>G	*BRCA2*	chr13	32900717	A	G	exonic	RT-PCR	TN
NM_000059.3:c.632-3C>G	*BRCA2*	chr13	32903577	C	G	3′ intronic	RT-PCR	TP
NM_000059.3:c.772C>A	*BRCA2*	chr13	32905146	C	A	exonic	RT-PCR	TN
NM_000059.3:c.1127T>G	*BRCA2*	chr13	32906742	T	G	exonic	RT-PCR + RNA-seq	TN
NM_000059.3:c.1291A>C	*BRCA2*	chr13	32906906	A	C	exonic	RT-PCR	TN
NM_000059.3:c.1480G>A	*BRCA2*	chr13	32907095	G	A	exonic	RT-PCR + RNA-seq	TN
NM_000059.3:c.1804G>A	*BRCA2*	chr13	32907419	G	A	exonic	RT-PCR	TN
NM_000059.3:c.1936A>C	*BRCA2*	chr13	32910428	A	C	exonic	RT-PCR	TN
NM_000059.3:c.2803G>C	*BRCA2*	chr13	32911295	G	C	exonic	RT-PCR	TN
NM_000059.3:c.2810A>C	*BRCA2*	chr13	32911302	A	C	exonic	RT-PCR	TN
NM_000059.3:c.3032C>G	*BRCA2*	chr13	32911524	C	G	exonic	RT-PCR	TN
NM_000059.3:c.3073A>G	*BRCA2*	chr13	32911565	A	G	exonic	RT-PCR	TN
NM_000059.3:c.6938-4C>T	*BRCA2*	chr13	32920960	C	T	3′ intronic	RT-PCR	TN
NM_000059.3:c.7021C>T	*BRCA2*	chr13	32929011	C	T	exonic	RT-PCR	TN
NM_000059.3:c.7610A>G	*BRCA2*	chr13	32930739	A	G	exonic	RT-PCR	TN
NM_000059.3:c.7822C>G	*BRCA2*	chr13	32936676	C	G	exonic	RT-PCR	TN
NM_000059.3:c.8192A>G	*BRCA2*	chr13	32937531	A	G	exonic	RT-PCR	TN
NM_000059.3:c.8258T>C	*BRCA2*	chr13	32937597	T	C	exonic	RT-PCR	TN
NM_000059.3:c.8378G>T	*BRCA2*	chr13	32944585	G	T	exonic	RT-PCR	TP
NM_000059.3:c.8486A>G	*BRCA2*	chr13	32944693	A	G	exonic	RT-PCR	TP
NM_000059.3:c.8963G>A	*BRCA2*	chr13	32953896	G	A	exonic	RT-PCR	TN
NM_000059.3:c.9104A>C	*BRCA2*	chr13	32954037	A	C	exonic	RT-PCR	TN
NM_000059.3:c.9104A>C	*BRCA2*	chr13	32954037	A	C	exonic	RT-PCR	TN
NM_000059.3:c.9242T>C	*BRCA2*	chr13	32954268	T	C	exonic	RT-PCR	TN

NM_000059.3:c.9367A>G	*BRCA2*	chr13	32968936	A	G	exonic	RT-PCR	TN
NM_000059.3:c.9456G>T	*BRCA2*	chr13	32969025	G	T	exonic	RT-PCR	TN
NM_000059.3:c.9502-13C>G	*BRCA2*	chr13	32971022	C	G	3′ intronic	RT-PCR + RNA-seq	TN
NM_000059.3:c.9586A>G	*BRCA2*	chr13	32971119	A	G	exonic	RT-PCR	TN
NM_000059.3:c.9604C>T	*BRCA2*	chr13	32971137	C	T	exonic	RT-PCR	TN
NM_000059.3:c.10045A>G	*BRCA2*	chr13	32972695	A	G	exonic	RT-PCR	TN
NM_000059.3:c.10249T>C	*BRCA2*	chr13	32972899	T	C	exonic	RT-PCR + RNA-seq	TN
NM_020366.3:c.491-386C>T	*RPGRIP1*	chr14	21770261	C	T	deep intronic	Minigene	TN
NM_030621.4:c.1509G>A	*DICER1*	chr14	95583959	C	T	exonic	RT-PCR	TN
NM_002420.5:c.899+29G>A	*TRPM1*	chr15	31355292	C	T	5′ extended	Minigene	TP
NM_000138.4:c.8149G>A	*FBN1*	chr15	48704843	C	T	exonic	RT-PCR	TP
NM_000138.4:c.8037C>G	*FBN1*	chr15	48707747	G	C	exonic	RT-PCR	TN
NM_000138.4:c.7916A>G	*FBN1*	chr15	48707868	T	C	exonic	RT-PCR	TN
NM_000138.4:c.7754T>C	*FBN1*	chr15	48712949	A	G	exonic	RT-PCR	TN
NM_000138.4:c.7664G>T	*FBN1*	chr15	48713790	C	A	exonic	RT-PCR	TN
NM_000138.4:c.7633C>T	*FBN1*	chr15	48713821	G	A	exonic	RT-PCR	TN
NM_000138.4:c.7606G>A	*FBN1*	chr15	48713848	C	T	exonic	RT-PCR	TN
NM_000138.4:c.7582T>C	*FBN1*	chr15	48713872	A	G	exonic	RT-PCR	TP
NM_000138.4:c.7379A>G	*FBN1*	chr15	48717640	T	C	exonic	RT-PCR	TN
NM_000138.4:c.7204+7C>G	*FBN1*	chr15	48719757	G	C	5′ extended	RT-PCR	TN
NM_000138.4:c.7204G>C	*FBN1*	chr15	48719764	C	G	exonic	RT-PCR	TN
NM_000138.4:c.7203A>G	*FBN1*	chr15	48719765	T	C	exonic	RT-PCR	TP
NM_000138.4:c.7003C>T	*FBN1*	chr15	48719965	G	A	exonic	RT-PCR	TP
NM_000138.4:c.6815A>G	*FBN1*	chr15	48722924	T	C	exonic	RT-PCR	TN
NM_000138.4:c.6740-3C>G	*FBN1*	chr15	48723002	G	C	3′ intronic	RT-PCR	TN
NM_000138.4:c.6694T>C	*FBN1*	chr15	48725108	A	G	exonic	RT-PCR	TP
NM_000138.4:c.6453C>T	*FBN1*	chr15	48729201	G	A	exonic	RT-PCR	TN
NM_000138.4:c.6313+3A>T	*FBN1*	chr15	48729962	T	A	5′ intronic	RT-PCR	TP
NM_000138.4:c.6251G>C	*FBN1*	chr15	48730027	C	G	exonic	RT-PCR	TP

NM_000138.4:c.6164-3C>T	*FBN1*	chr15	48730117	G	A	3′ intronic	RT-PCR	TN
NM_000138.4:c.6158G>A	*FBN1*	chr15	48733923	C	T	exonic	RT-PCR	TN
NM_000138.4:c.6031T>G	*FBN1*	chr15	48736744	A	C	exonic	RT-PCR	TN
NM_000138.4:c.5926G>A	*FBN1*	chr15	48736849	C	T	exonic	RT-PCR	TN
NM_000138.4:c.5834G>A	*FBN1*	chr15	48737656	C	T	exonic	RT-PCR	TN
NM_000138.4:c.5789A>G	*FBN1*	chr15	48737701	T	C	exonic	RT-PCR	TN
NM_000138.4:c.5788+5G>A	*FBN1*	chr15	48738898	C	T	5′ extended	RT-PCR	TN
NM_000138.4:c.5707G>A	*FBN1*	chr15	48738984	C	T	exonic	RT-PCR	TP
NM_000138.4:c.5678A>G	*FBN1*	chr15	48739013	T	C	exonic	RT-PCR	TN
NM_000138.4:c.5627G>A	*FBN1*	chr15	48741009	C	T	exonic	RT-PCR	TN
NM_000138.4:c.5497T>C	*FBN1*	chr15	48744807	A	G	exonic	RT-PCR	TN
NM_000138.4:c.5377T>C	*FBN1*	chr15	48748879	A	G	exonic	RT-PCR	TN
NM_000138.4:c.5372G>T	*FBN1*	chr15	48748884	C	A	exonic	RT-PCR	TN
NM_000138.4:c.5296G>A	*FBN1*	chr15	48752443	C	T	exonic	RT-PCR	TN
NM_000138.4:c.5021G>A	*FBN1*	chr15	48756140	C	T	exonic	RT-PCR	TP
NM_000138.4:c.4780G>A	*FBN1*	chr15	48758023	C	T	exonic	RT-PCR	TN
NM_000138.4:c.4747+5G>A	*FBN1*	chr15	48760130	C	T	5′ extended	RT-PCR	TN
NM_000138.4:c.4582G>T	*FBN1*	chr15	48760609	C	A	exonic	RT-PCR	TP
NM_000138.4:c.4343A>G	*FBN1*	chr15	48762947	T	C	exonic	RT-PCR	TP
NM_000138.4:c.4096G>A	*FBN1*	chr15	48766566	C	T	exonic	RT-PCR	TN
NM_000138.4:c.4031G>A	*FBN1*	chr15	48766781	C	T	exonic	RT-PCR	TN
NM_000138.4:c.4027G>A	*FBN1*	chr15	48766785	C	T	exonic	RT-PCR	TN
NM_000138.4:c.3974A>T	*FBN1*	chr15	48766838	T	A	exonic	RT-PCR	TN
NM_000138.4:c.3964G>C	*FBN1*	chr15	48773852	C	G	exonic	RT-PCR	TN
NM_000138.4:c.3963A>G	*FBN1*	chr15	48773853	T	C	exonic	RT-PCR	TP
NM_000138.4:c.3772C>T	*FBN1*	chr15	48776081	G	A	exonic	RT-PCR	TN
NM_000138.4:c.3712G>A	*FBN1*	chr15	48777571	C	T	exonic	RT-PCR	TN
NM_000138.4:c.3533A>G	*FBN1*	chr15	48779328	T	C	exonic	RT-PCR	TN
NM_000138.4:c.3509G>A	*FBN1*	chr15	48779352	C	T	exonic	RT-PCR	TN
NM_000138.4:c.3463+3A>C	*FBN1*	chr15	48779506	T	G	5′ intronic	RT-PCR	TN

NM_000138.4:c.3344A>G	*FBN1*	chr15	48779628	T	C	exonic	RT-PCR	TN
NM_000138.4:c.3332G>A	*FBN1*	chr15	48780315	C	T	exonic	RT-PCR	TN
NM_000138.4:c.3268C>T	*FBN1*	chr15	48780379	G	A	exonic	RT-PCR	TN
NM_000138.4:c.3209-13T>A	*FBN1*	chr15	48780451	A	T	3′ intronic	RT-PCR	TN
NM_000138.4:c.3197G>C	*FBN1*	chr15	48780576	C	G	exonic	RT-PCR	TP
NM_000138.4:c.3124G>C	*FBN1*	chr15	48780649	C	G	exonic	RT-PCR	TN
NM_000138.4:c.2953G>A	*FBN1*	chr15	48782177	C	T	exonic	RT-PCR	TN
NM_000138.4:c.2952C>A	*FBN1*	chr15	48782178	G	T	exonic	RT-PCR	TN
NM_000138.4:c.2927G>A	*FBN1*	chr15	48782203	C	T	exonic	RT-PCR	TN
NM_000138.4:c.2645C>T	*FBN1*	chr15	48787352	G	A	exonic	RT-PCR	TN
NM_000138.4:c.2638G>A	*FBN1*	chr15	48787359	C	T	exonic	RT-PCR	TN
NM_000138.4:c.2369G>A	*FBN1*	chr15	48788347	C	T	exonic	RT-PCR	TN
NM_000138.4:c.2293G>A	*FBN1*	chr15	48789463	C	T	exonic	RT-PCR	TN
NM_000138.4:c.1916G>A	*FBN1*	chr15	48797266	C	T	exonic	RT-PCR	TP
NM_000138.4:c.1909T>C	*FBN1*	chr15	48797273	A	G	exonic	RT-PCR	TN
NM_000138.4:c.1883G>A	*FBN1*	chr15	48797299	C	T	exonic	RT-PCR	TN
NM_000138.4:c.1846G>A	*FBN1*	chr15	48797336	C	T	exonic	RT-PCR	TN
NM_000138.4:c.1633C>T	*FBN1*	chr15	48802322	G	A	exonic	RT-PCR	TN
NM_000138.4:c.1595A>G	*FBN1*	chr15	48802360	T	C	exonic	RT-PCR	TN
NM_000138.4:c.1588G>A	*FBN1*	chr15	48805746	C	T	exonic	RT-PCR	TN
NM_000138.4:c.1510T>C	*FBN1*	chr15	48805824	A	G	exonic	RT-PCR	TN
NM_000138.4:c.1426T>A	*FBN1*	chr15	48807626	A	T	exonic	RT-PCR	TN
NM_000138.4:c.1169C>T	*FBN1*	chr15	48808538	G	A	exonic	RT-PCR	TN
NM_000138.4:c.736G>A	*FBN1*	chr15	48829808	C	T	exonic	RT-PCR	TN
NM_000138.4:c.640G>A	*FBN1*	chr15	48829904	C	T	exonic	RT-PCR	TP
NM_000138.4:c.538+4A>G	*FBN1*	chr15	48888476	T	C	5′ intronic	RT-PCR	TN
NM_000138.4:c.433T>C	*FBN1*	chr15	48892345	A	G	exonic	RT-PCR	TN
NM_000138.4:c.364C>T	*FBN1*	chr15	48892414	G	A	exonic	RT-PCR	TN
NM_000138.4:c.247+9A>G	*FBN1*	chr15	48905198	T	C	5′ extended	RT-PCR	TN
NM_000138.4:c.184C>T	*FBN1*	chr15	48905270	G	A	exonic	RT-PCR	TP

NM_004855.4:c.847-10A>G	*PIGB*	chr15	55632800	A	G	3′ intronic	RT-PCR	TN
NM_005902.3:c.802C>T	*SMAD3*	chr15	67473722	C	T	exonic	RT-PCR + RNA-seq	TP
NM_001271.3:c.4138-6T>G	*CHD2*	chr15	93545401	T	G	3′ intronic	RT-PCR	TN
NM_000548.3:c.4492A>C	*TSC2*	chr16	2134715	A	C	exonic	RT-PCR	TP
NM_001009944.2:c.1723-23T>C	*PKD1*	chr16	2166142	A	G	3′ intronic	RT-PCR	TP
NM_024675.3:c.3201+5G>C	*PALB2*	chr16	23625320	C	G	5′ extended	RT-PCR	TN
NM_001605.2:c.2286G>A	*AARS*	chr16	70289631	C	T	exonic	RT-PCR	TP
NM_001142864.2:c.6963C>T	*PIEZO1*	chr16	88782855	G	A	exonic	RT-PCR	TP
NM_001142864.2:c.6651C>A	*PIEZO1*	chr16	88783440	G	T	exonic	RT-PCR	TN
NM_001256182.1:c.5511G>A	*ANKRD11*	chr16	89347439	C	T	exonic	RT-PCR	TN
NM_000430.3:c.900+3A>G	*PAFAH1B1*	chr17	2577585	A	G	5′ intronic	RT-PCR	TN
NM_000430.3:c.1002+6T>A	*PAFAH1B1*	chr17	2579906	T	A	5′ extended	RT-PCR	TP
NM_000546.5:c.783-60G>A	*TP53*	chr17	7577215	C	T	deep intronic	RT-PCR	TP
NM_000546.5:c.623A>G	*TP53*	chr17	7578226	T	C	exonic	RT-PCR + RNA-seq	TN
NM_000180.3:c.3043+5G>A	*GUCY2D*	chr17	7919164	G	A	5′ extended	Minigene	TN
NM_005208.4:c.213C>T	*CRYBA1*	chr17	27577316	C	T	exonic	Minigene	TP
NM_001042492.2:c.1062+3A>G	*NF1*	chr17	29527616	A	G	5′ intronic	RT-PCR	TP
NM_001042492.2:c.7895A>G	*NF1*	chr17	29684312	A	G	Exonic	RT-PCR + RNA-seq	TP
NM_007294.3:c.5453A>G	*BRCA1*	chr17	41199674	T	C	exonic	RT-PCR	TP
NM_007294.3:c.5431C>A	*BRCA1*	chr17	41199696	G	T	exonic	RT-PCR	TP
NM_007294.3:c.5425G>T	*BRCA1*	chr17	41199702	C	A	exonic	RT-PCR	TN
NM_007294.3:c.5407G>T	*BRCA1*	chr17	41199720	C	A	exonic	RT-PCR	TN
NM_007294.3:c.5332+13G>T	*BRCA1*	chr17	41203067	C	A	5′ extended	RT-PCR	TN
NM_007294.3:c.5252G>A	*BRCA1*	chr17	41209094	C	T	exonic	RT-PCR	TN
NM_007294.3:c.5207T>C	*BRCA1*	chr17	41209139	A	G	exonic	RT-PCR	TN
NM_007294.3:c.5198A>G	*BRCA1*	chr17	41209148	T	C	exonic	RT-PCR	TN
NM_007294.3:c.5157G>T	*BRCA1*	chr17	41215386	C	A	exonic	RT-PCR	TN
NM_007294.3:c.5153-26A>G	*BRCA1*	chr17	41215416	T	C	3′ intronic	RT-PCR	TN
NM_007294.3:c.5152+6T>C	*BRCA1*	chr17	41215885	A	G	5′ extended	RT-PCR	TP
NM_007294.3:c.5152+5G>C	*BRCA1*	chr17	41215886	C	G	5′ extended	RT-PCR	TP

NM_007294.3:c.5117G>A	*BRCA1*	chr17	41215926	C	T	exonic	RT-PCR	TP
NM_007294.3:c.5096G>A	*BRCA1*	chr17	41215947	C	T	exonic	RT-PCR	TN
NM_007294.3:c.5075-6C>A	*BRCA1*	chr17	41215974	G	T	3′ intronic	RT-PCR	TN
NM_007294.3:c.5074+7T>C	*BRCA1*	chr17	41219618	G	G	5′ extended	Minigene	TN
NM_007294.3:c.5024C>T	*BRCA1*	chr17	41219675	G	A	exonic	RT-PCR	TN
NM_007294.3:c.5024C>T	*BRCA1*	chr17	41219675	G	A	exonic	RT-PCR	TN
NM_007294.3:c.4987-11T>C	*BRCA1*	chr17	41219723	A	G	3′ intronic	RT-PCR + RNA-seq	TN
NM_007294.3:c.4868C>G	*BRCA1*	chr17	41223063	G	C	exonic	RT-PCR	TN
NM_007294.3:c.4676-8C>G	*BRCA1*	chr17	41223263	G	C	3′ intronic	RT-PCR + RNA-seq	TP
NM_007294.3:c.4484+15T>C	*BRCA1*	chr17	41228490	A	G	5′ extended	RT-PCR	TN
NM_007294.3:c.4393A>G	*BRCA1*	chr17	41228596	T	C	exonic	RT-PCR	TN
NM_007294.3:c.4357+6T>C	*BRCA1*	chr17	41234415	A	G	5′ extended	RT-PCR	TN
NM_007294.3:c.4343G>A	*BRCA1*	chr17	41234435	C	T	exonic	RT-PCR	TP
NM_007294.3:c.3845A>T	*BRCA1*	chr17	41243703	T	A	exonic	RT-PCR	TN
NM_007294.3:c.3047A>G	*BRCA1*	chr17	41244501	T	C	exonic	RT-PCR	TN
NM_007294.3:c.1731A>G	*BRCA1*	chr17	41245817	T	C	exonic	RT-PCR + RNA-seq	TN
NM_007294.3:c.612G>C	*BRCA1*	chr17	41247921	C	G	exonic	RT-PCR	TN
NM_007294.3:c.509G>A	*BRCA1*	chr17	41251830	C	T	exonic	RT-PCR	TN
NM_007294.3:c.286G>C	*BRCA1*	chr17	41256900	C	G	exonic	RT-PCR	TN
NM_007294.3:c.286G>A	*BRCA1*	chr17	41256900	C	T	exonic	RT-PCR	TN
NM_007294.3:c.213-5T>G	*BRCA1*	chr17	41256978	A	C	3′ intronic	RT-PCR	TN
NM_007294.3:c.213-14C>G	*BRCA1*	chr17	41256987	G	C	3′ intronic	RT-PCR	TP
NM_007294.3:c.212G>T	*BRCA1*	chr17	41258473	C	A	exonic	RT-PCR	TP
NM_007294.3:c.189A>T	*BRCA1*	chr17	41258496	T	A	exonic	RT-PCR	TP
NM_007294.3:c.134+3A>T	*BRCA1*	chr17	41267740	T	A	5′ intronic	RT-PCR	TN
NM_007294.3:c.81-14C>T	*BRCA1*	chr17	41267810	G	A	3′ intronic	RT-PCR	TP
NM_007294.3:c.81-65G>C	*BRCA1*	chr17	41267861	C	G	deep intronic	RT-PCR	TN
NM_007294.3:c.36A>G	*BRCA1*	chr17	41276078	T	C	exonic	RT-PCR	TN
NM_007294.3:c.19C>T	*BRCA1*	chr17	41276095	G	A	exonic	RT-PCR	TN
NM_015443.3:c.2725-5T>G	*KANSL1*	chr17	44110563	A	C	3′ intronic	RT-PCR	TN

NM_015443.3:c.1848G>A	*KANSL1*	chr17	44143903	C	T	exonic	RT-PCR	TP
NM_005993.4:c.1922G>C	*TBCD*	chr17	80863929	G	C	exonic	RT-PCR	TP
NM_007254.3:c.578+4A>G	*PNKP*	chr19	50367577	T	C	5′ intronic	RT-PCR	TP
NM_001308632.1:c.1620C>T	*POLD1*	chr19	50910365	C	T	exonic	RT-PCR	TP
NM_015629.3:c.528-38C>T	*PRPF31*	chr19	54627090	C	T	3′ intronic	Minigene	TN
NM_018848.3:c.803T>G	*MKKS*	chr20	10393360	A	C	3′ exonic	RT-PCR	TN
NM_018848.3:c.749G>A	*MKKS*	chr20	10393414	C	T	exonic	RT-PCR	TN
NM_015600.4:c.867+5G>A	*ABHD12*	chr20	25288597	C	T	5′ extended	Minigene	TN
NM_001098.2:c.526-642C>T	*ACO2*	chr22	41910739	C	T	deep intronic	Minigene	TP
NM_001034853.1:c.1754-3C>G	*RPGR*	chrX	38146501	G	C	3′ intronic	Minigene	TP
NM_001034853.1:c.247G>T	*RPGR*	chrX	38182106	C	A	exonic	Minigene	TP
NM_001399.4:c.957C>G	*EDA*	chrX	69255240	C	G	exonic	RT-PCR	TN
NM_001286074.1:c.488C>A	*TAF1*	chrX	70595092	C	A	exonic	RT-PCR	TN
NM_004208.3:c.697-27T>G	*AIFM1*	chrX	129274619	A	C	3′ intronic	RT-PCR	TP
NM_001110556.1:c.2747A>T	*FLNA*	chrX	153590426	T	A	exonic	RT-PCR	TN
NM_001363.3:c.915+10G>A	*DKC1*	chrX	153997595	G	A	5′ intronic	RT-PCR + RNA-seq	TP

Supplementary Table S2. *Pairwise comparisons of the AUC for each in-silico prioritization tool.* Displayed are the p-values for each comparison, calculated using the pROC R package (bootstrap test, 2,000 iterations). Significant differences in the AUC after Bonferroni correction are indicated by bold text, showing that SpliceAI and a consensus approach perform better than all other approaches. All approaches perform better than CADD alone.

	SpliceAI	SPIDEX	MMSplice	MaxEntScan	KipoiSplice	TraP	S-CAP	CADD	Consensus
SpliceAI	x								
SPIDEX	**4.61E-09**	x							
MMSplice	**5.35E-06**	0.05125	x						
MaxEntScan	**9.23E-08**	0.08028	0.9835	x					
KipoiSplice	**5.37E-05**	0.009853	0.69	0.6723	x				
TraP	**9.79E-05**	**0.005707**	0.5183	0.4253	0.7236	x			
S-CAP	**2.01E-10**	0.5741	0.09294	0.1365	0.06732	0.03671	x		
CADD	**2.20E-16**	**8.08E-05**	**1.69E-12**	**6.73E-12**	**2.13E-11**	**1.01E-10**	**3.39E-13**	x	
Consensus	0.2358	**1.80E-09**	**2.74E-05**	**8.37E-08**	**5.43E-08**	**1.09E-05**	**4.47E-10**	**2.20E-16**	x

Supplementary Table S3. *Optimal thresholds and values calculated through ROC-AUC for 250 variants of uncertain significance which had received functional analysis through blood-based RNA analysis or synthetic minigene assays.* We determined 80/250 variants to impact splicing (true positives) and 170/250 variants to not impact splicing (true negatives). ROC curves were created using the pROC ggplot2 package in R, and 95% confidence intervals were calculated using 2,000 stratified bootstrap replicates. Optimal thresholds were selected based on the maximum Youden's J statistic, as calculated using pROC.

Prioritization Approach	Optimal Threshold	Specificity	Sensitivity	AUC (95% CI)
SpliceAI	0.145	0.9	0.9113924	0.9536 (0.9238-0.9771)
SPIDEX	1.84715	0.8882353	0.5443038	0.7305 (0.6570-0.8018)
MMSplice	0.2892759	0.8235294	0.721519	0.8081 (0.7459-0.8659)
MaxEntScan	1.715	0.7411765	0.8607595	0.8103 (0.7576-0.8653)
KipoiSplice	0.509906	0.9176471	0.721519	0.8299 (0.7638-0.8942)
TraP	0.4675	0.8176471	0.7974684	0.8428 (0.7858-0.8938)
S-CAP	0.00245211	0.8117647	0.6329114	0.7471 (0.6823-0.8075)
CADD	25.2	0.4470588	0.7088608	0.4716 (0.4004-0.5486)
Consensus	3.5	0.8647059	0.8860759	0.9375 (0.9106-0.9651)
Weighted metric (SpliceAI & Consensus)	0.6275	0.9117647	0.9240506	0.9635 (0.9413-0.9818)

Supplementary Table S4. *Thresholds utilized for each of the* in silico *splicing algorithms to identify variants expected to impact splicing.*

Tool	**Threshold**
SpliceAI	0.2
MaxEntScan	Increase or decrease of 1
SPIDEX	Increase or decrease of 5
TraP	
Non-coding	0.289
Coding	0.416
MMSplice	Increase or decrease of 2
KipoiSplice	0.95
CADD	
Exonic	7.39
5extended	0.005
3intronic	0.006
5intronic	0.006
Anything else	0.006
S-CAP	
Exonic	0.009
5extended	0.005
3intronic	0.006
5intronic	0.006
5core	0.033
3core	0.034

Supplementary Table S5. *Variability in accuracy of splice prediction tools.* Using the standard error of the mean (SEM), 95% confidence intervals (CIs) were calculated across the 2,000 bootstraps used to calculate model accuracy, as depicted in **Figure 9c**. We observed that, for all tools investigated, variability in accuracy was low across the bootstraps.

Approach	Mean accuracy	SEM	95% CI
CADD	0.3877	0.001340	0.3864-0.3890
Consensus	0.8679	0.0009484	0.8670-0.8689
KipoiSplice	0.7475	0.001224	0.7462-0.7487
MaxEntScan	0.7247	0.001279	0.7234-0.7260
MMSplice	0.7560	0.001193	0.7548-0.7572
S-CAP	0.7512	0.001201	0.7500-0.7524
SPIDEX	0.6963	0.001262	0.6951-0.6976
SpliceAI	0.9082	0.0008068	0.9074-0.9090
TraP	0.7807	0.001172	0.7796-0.7820

Supplementary Table S6. *Summary of rare unique variants prioritized by each in silico splicing prediction tool/strategy.*

Tool	Number of prioritized variants
CADD	16,110
SPIDEX	684
SpliceAI	674
MaxEntScan	3,115
MMSplice	259
KipoiSplice	224
TraP	2,024
S-CAP	6033
Consensus (4/8)	632

Supplementary Table S7. *Number of rare unique variants prioritized using different levels of consensus between in silico splicing tools.*

Consensus (*n* of 8 tools)	Number of prioritized variants	Number of 5' / 3' core variants
1	8,660	1
2	6,373	1
3	1,439	1
4	279	4 (1%)
5	102	3 (3%)
6	76	21 (28%)
7	82	52 (63%)
8	93	83 (89%)

Supplementary Table S8. *Summary of SpliceAI variant scores by their overlap with other in silico splicing tools also prioritizing the variant.* LCI = lower 95% confidence interval; UCI = upper 95% confidence interval.

Consensus	Median	*n*	LCI	UCI
SpliceAI alone	0.36	2	0.060	0.661
1 other	0.267	48	0.230	0.304
2 others	0.290	101	0.251	0.330
3 others	0.424	103	0.375	0.474
4 others	0.541	63	0.477	0.604
5 others	0.783	50	0.719	0.846
6 others	0.961	73	0.914	1
7 others	0.974	88	0.945	1

Supplementary Table S9. *Summary of SpliceAI variant scores by the impacted region of the genome, as defined in Jagadeesh et al. (2019)*. LCI, lower 95% confidence interval; UCI, upper 95% confidence interval.

Region	Median	*n*	LCI	UCI
3′ core	0.982	63	0.941	1
3′ intronic	0.432	78	0.369	0.495
5′ core	0.973	84	0.947	1
5′ extended	0.621	84	0.571	0.671
5′ intronic	0.487	22	0.391	0.584
exonic	0.303	197	0.271	0.336

Supplementary Table 10. *175 genes from the MCGM retinal dystrophy panel and their associated transcripts for routine genetic testing.*

Gene symbol	Ensembl Gene ID	RefSeq transcript ID(s)	Ensembl transcript ID(s)
ABCA4	ENSG00000198691	NM_000350	ENST00000370225
ABHD12	ENSG00000100997	NM_001042472	ENST00000339157
ACBD5	ENSG00000107897	NM_145698	ENST00000396271
ADAM9	ENSG00000168615	NM_003816	ENST00000487273
ADAMTS18	ENSG00000140873	NM_199355	ENST00000282849
AHI1	ENSG00000135541	NM_001134832	ENST00000327035
		NM_017651	ENST00000457866
AIPL1	ENSG00000129221	NM_014336	ENST00000381129
ARL2BP	ENSG00000102931	NM_012106	ENST00000219204
ARL6	ENSG00000113966	NM_032146	ENST00000335979
BBIP1	ENSG00000214413	NM_001195306	ENST00000448814
BBS1	ENSG00000174483	NM_024649	ENST00000318312
BBS10	ENSG00000179941	NM_024685	ENST00000393262
BBS12	ENSG00000181004	NM_001178007	ENST00000542236
BBS2	ENSG00000125124	NM_031885	ENST00000245157
BBS4	ENSG00000140463	NM_033028	ENST00000268057
BBS5	ENSG00000163093	NM_152384	ENST00000295240
BBS7	ENSG00000138686	NM_176824	ENST00000264499
BBS9	ENSG00000122507	NM_198428	ENST00000242067
BEST1	ENSG00000167995	NM_004183	ENST00000378043
C1QTNF5	ENSG00000223953	NM_015645	ENST00000445041
C21orf2	ENSG00000160226	NM_004928	ENST00000339818
C2orf71	ENSG00000179270	NM_001029883	ENST00000331664
C8orf37	ENSG00000156172	NM_177965	ENST00000286688
CA4	ENSG00000167434	NM_000717	ENST00000300900
CABP4	ENSG00000175544	NM_145200	ENST00000325656
CACNA1F	ENSG00000102001	NM_005183	ENST00000376265
CACNA2D4	ENSG00000151062	NM_172364	ENST00000382722
CAPN5	ENSG00000149260	NM_004055	ENST00000278559
CC2D2A	ENSG00000048342	NM_001080522	ENST00000503292
CDH23	ENSG00000107736	NM_022124	ENST00000224721
CDH3	ENSG00000062038	NM_001793	ENST00000264012
CDHR1	ENSG00000148600	NM_001171971	ENST00000332904
		NM_033100	ENST00000372117
CEP164	ENSG00000110274	NM_014956	ENST00000278935
CEP290	ENSG00000198707	NM_025114	ENST00000552810
CERKL	ENSG00000188452	NM_001030311	ENST00000410087
CHM	ENSG00000188419	NM_000390	ENST00000357749
CIB2	ENSG00000136425	NM_006383	ENST00000258930
CLN3	ENSG00000188603	NM_000086	ENST00000359984
CLRN1	ENSG00000163646	NM_001195794	ENST00000328863

		NM_052995	ENST00000295911
CNGA1	ENSG00000198515	NM_001142564	ENST00000544810
CNGA3	ENSG00000144191	NM_001298	ENST00000393504
CNGB1	ENSG00000070729	NM_001297	ENST00000251102
CNGB3	ENSG00000170289	NM_019098	ENST00000320005
CNNM4	ENSG00000158158	NM_020184	ENST00000377075
CRB1	ENSG00000134376	NM_201253	ENST00000367400
CRX	ENSG00000105392	NM_000554	ENST00000221996
CSPP1	ENSG00000104218	NM_024790	ENST00000262210
CYP4V2	ENSG00000145476	NM_207352	ENST00000378802
DFNB31	ENSG00000095397	NM_015404	ENST00000362057
DHDDS	ENSG00000117682	NM_024887	ENST00000360009
DTHD1	ENSG00000197057	NM_001136536	ENST00000357504
		NM_001170700	ENST00000456874
EFEMP1	ENSG00000115380	NM_001039348	ENST00000394555
ELOVL4	ENSG00000118402	NM_022726	ENST00000369816
EMC1	ENSG00000127463	NM_015047	ENST00000477853
EYS	ENSG00000188107	NM_001142800	ENST00000503581
FAM161A	ENSG00000170264	NM_001201543	ENST00000404929
FLVCR1	ENSG00000162769	NM_014053	ENST00000366971
FSCN2	ENSG00000186765	NM_001077182	ENST00000417245
FZD4	ENSG00000174804	NM_012193	ENST00000531380
GNAT1	ENSG00000114349	NM_000172	ENST00000433068
GNAT2	ENSG00000134183	NM_005272	ENST00000351050
GNPTG	ENSG00000090581	NM_032520	ENST00000204679
GPR125	ENSG00000197177	NM_145290	ENST00000334304
GPR179	ENSG00000277399	NM_001004334	ENST00000342292
GPR98	ENSG00000164199	NM_032119	ENST00000405460
GRM6	ENSG00000113262	NM_000843	ENST00000231188
GUCA1A	ENSG00000048545	NM_000409	ENST00000053469
GUCA1B	ENSG00000112599	NM_002098	ENST00000230361
GUCY2D	ENSG00000132518	NM_000180	ENST00000254854
HARS	ENSG00000170445	NM_002109	ENST00000504156
HMX1	ENSG00000215612	NM_018942	ENST00000400677
IDH3B	ENSG00000101365	NM_006899	ENST00000380843
		NM_174855	ENST00000380851
IFT140	ENSG00000187535	NM_014714	ENST00000426508
IMPDH1	ENSG00000106348	NM_000883	ENST00000338791
IMPG1	ENSG00000112706	NM_001563	ENST00000369950
IMPG2	ENSG00000081148	NM_016247	ENST00000193391
INPP5E	ENSG00000148384	NM_019892	ENST00000371712
INVS	ENSG00000119509	NM_014425	ENST00000262457
IQCB1	ENSG00000173226	NM_001023570	ENST00000310864
ITM2B	ENSG00000136156	NM_021999	ENST00000378565

KCNJ13	ENSG00000115474	NM_002242	ENST00000233826
KCNV2	ENSG00000168263	NM_133497	ENST00000382082
KIAA1549	ENSG00000122778	NM_001164665	ENST00000440172
		NM_020910	ENST00000440172
KIF11	ENSG00000138160	NM_004523	ENST00000260731
KLHL7	ENSG00000122550	NM_001031710	ENST00000339077
LCA5	ENSG00000135338	NM_181714	ENST00000392959
LRAT	ENSG00000121207	NM_004744	ENST00000336356
LRP5	ENSG00000162337	NM_002335	ENST00000294304
LZTFL1	ENSG00000163818	NM_020347	ENST00000296135
MERTK	ENSG00000153208	NM_006343	ENST00000421804
MFRP	ENSG00000235718	NM_031433	ENST00000445041
MKKS	ENSG00000125863	NM_018848	ENST00000399054
MKS1	ENSG00000011143	NM_001165927	ENST00000537529
		NM_017777	ENST00000393119
MVK	ENSG00000110921	NM_000431	ENST00000228510
MYO7A	ENSG00000137474	NM_000260	ENST00000409709
NDP	ENSG00000124479	NM_000266	ENST00000378062
NEK2	ENSG00000117650	NM_002497	ENST00000366999
NMNAT1	ENSG00000173614	NM_022787	ENST00000377205
NPHP1	ENSG00000144061	NM_000272	ENST00000393272
NPHP3	ENSG00000113971	NM_153240	ENST00000337331
NPHP4	ENSG00000131697	NM_015102	ENST00000378156
NR2E3	ENSG00000278570	NM_014249	ENST00000617575
NRL	ENSG00000129535	NM_006177	ENST00000397002
NYX	ENSG00000188937	NM_022567	ENST00000342595
OAT	ENSG00000065154	NM_000274	ENST00000368845
OFD1	ENSG00000046651	NM_003611	ENST00000340096
OTX2	ENSG00000165588	NM_021728	ENST00000339475
PANK2	ENSG00000125779	NM_153638	ENST00000316562
PCDH15	ENSG00000150275	NM_001142763	ENST00000361849
		NM_001142769	ENST00000395445
		NM_001142770	ENST00000395438
		NM_001142771	ENST00000373965
PCYT1A	ENSG00000161217	NM_005017	ENST00000292823
PDE6A	ENSG00000132915	NM_000440	ENST00000255266
PDE6B	ENSG00000133256	NM_000283	ENST00000255622
PDE6C	ENSG00000095464	NM_006204	ENST00000371447
PDE6G	ENSG00000185527	NM_002602	ENST00000331056
PEX1	ENSG00000127980	NM_000466	ENST00000248633
PEX2	ENSG00000164751	NM_000318	ENST00000357039
PEX7	ENSG00000112357	NM_000288	ENST00000318471
PHYH	ENSG00000107537	NM_006214	ENST00000263038
PITPNM3	ENSG00000091622	NM_031220	ENST00000262483

PLA2G5	ENSG00000127472	NM_000929	ENST00000375108
PRCD	ENSG00000214140	NM_001077620	-
PROM1	ENSG00000007062	NM_006017	ENST00000447510
PRPF3	ENSG00000117360	NM_004698	ENST00000324862
PRPF31	ENSG00000105618	NM_015629	ENST00000321030
PRPF4	ENSG00000136875	NM_004697	ENST00000374198
PRPF6	ENSG00000101161	NM_012469	ENST00000266079
PRPF8	ENSG00000174231	NM_006445	ENST00000304992
PRPH2	ENSG00000112619	NM_000322	ENST00000230381
RAB28	ENSG00000157869	NM_001017979	ENST00000330852
RAX2	ENSG00000173976	NM_032753	ENST00000555978
RBP3	ENSG00000265203	NM_002900	ENST00000224600
RBP4	ENSG00000138207	NM_006744	ENST00000371464
RD3	ENSG00000198570	NM_183059	ENST00000367002
RDH12	ENSG00000139988	NM_152443	ENST00000551171
RDH5	ENSG00000135437	NM_001199771	ENST00000257895
RGR	ENSG00000148604	NM_002921	ENST00000359452
RGS9	ENSG00000108370	NM_001165933	ENST00000443584
		NM_003835	ENST00000262406
RHO	ENSG00000163914	NM_000539	ENST00000296271
RIMS1	ENSG00000079841	NM_001168407	ENST00000401910
		NM_001168410	ENST00000517827
		NM_014989	ENST00000521978
RLBP1	ENSG00000140522	NM_000326	ENST00000268125
ROM1	ENSG00000149489	NM_000327	ENST00000278833
RP1	ENSG00000104237	NM_006269	ENST00000220676
RP1L1	ENSG00000183638	NM_178857	ENST00000382483
RP2	ENSG00000102218	NM_006915	ENST00000218340
RP9	ENSG00000164610	NM_203288	ENST00000297157
RPE65	ENSG00000116745	NM_000329	ENST00000262340
RPGR	ENSG00000156313	NM_001034853	ENST00000378505
RPGRIP1	ENSG00000092200	NM_020366	ENST00000400017
RPGRIP1L	ENSG00000103494	NM_015272	ENST00000379925
RS1	ENSG00000102104	NM_000330	ENST00000379984
SAG	ENSG00000130561	NM_000541	ENST00000409110
SDCCAG8	ENSG00000054282	NM_006642	ENST00000366541
SEMA4A	ENSG00000196189	NM_022367	ENST00000368285
SLC24A1	ENSG00000074621	NM_001254740	ENST00000339868
		NM_004727	ENST00000261892
SNRNP200	ENSG00000144028	NM_014014	ENST00000323853
SPATA7	ENSG00000042317	NM_018418	ENST00000393545
TEAD1	ENSG00000187079	NM_021961	ENST00000361905
TIMP3	ENSG00000100234	NM_000362	ENST00000266085
TMEM237	ENSG00000155755	NM_001044385	ENST00000409883

TOPORS	ENSG00000197579	NM_005802	ENST00000360538
TRIM32	ENSG00000119401	NM_012210	ENST00000450136
TRPM1	ENSG00000134160	NM_002420	ENST00000397795
TSPAN12	ENSG00000106025	NM_012338	ENST00000222747
TTC8	ENSG00000165533	NM_144596	ENST00000380656
TUB	ENSG00000166402	NM_177972	ENST00000299506
TULP1	ENSG00000112041	NM_003322	ENST00000229771
UNC119	ENSG00000109103	NM_005148	ENST00000335765
		NM_054035	ENST00000301032
USH1C	ENSG00000006611	NM_005709	ENST00000318024
		NM_153676	ENST00000005226
USH1G	ENSG00000182040	NM_173477	ENST00000319642
USH2A	ENSG00000042781	NM_206933	ENST00000307340
VCAN	ENSG00000038427	NM_004385	ENST00000265077
VPS13B	ENSG00000132549	NM_017890	ENST00000358544
WDPCP	ENSG00000143951	NM_015910	ENST00000272321
WDR19	ENSG00000157796	NM_025132	ENST00000399820
ZNF423	ENSG00000102935	NM_015069	ENST00000561648
ZNF513	ENSG00000163795	NM_144631	ENST00000323703

Supplementary Table S11. *List of putative branchpoint-impacting SNVs identified through retinal dystrophy gene panel testing*. BP position is "A" for variants affecting the branchpoint residue itself, while "-2" represents variants affecting the conserved residue (most often a thymidine) two bases upstream. A patient's phenotype was deemed to be "solved" if a variant(s) was returned as likely pathogenic or pathogenic and wholly accounted for patient phenotype, including being in the correct zygosity. A given patient may harbour multiple SNVs. Colouring of rows indicates the corresponding section of the pie chart in **Figure 12b**.

Patient ID	Gene	HGVSg	BP position	Supporting datasets	Solved?	Consistent with phenotype/inheritance?	Existing carrier finding?
1	*ADAM9*	chr8:38948762T>C	-2	All	Y	N	N
2	*AHI1*	chr6:135774598T>A	A	BPP	Y	N	N
3	*BBS1*	chr11:66278454A>G	A	Mercer; SVM-BPfinder	N	N	N
4	*BBS1*	chr11:66287067A>T	A	SVM-BPfinder; BPP	N	Y	Y
5	*BBS1*	chr11:66288725C>T	A	BPP	N	N	N
6	*BBS1*	chr11:66290900C>T	-2	Mercer; BPP	N	Y	N
7	*BBS2*	chr16:56531016T>C	A	BPP	Y	N	N
8	*BBS2*	chr16:56533849T>C	A	BPP	Y	N	N
9	*BBS2*	chr16:56543976T>G	A	All	N	Y	N
10	*BBS7*	chr4:122749687A>G	-2	SVM-BPfinder	Y	N	N
11	*BEST1*	chr11:61724829A>T	A	SVM-BPfinder; BPP	N	N	N
12	*CACNA2D4*	chr12:1919533A>T	-2	SVM-BPfinder; BPP	N	N	N
13	*CC2D2A*	chr4:15589418A>T	A	BPP	Y	N	N
14	*CC2D2A*	chr4:15602831T>C	-2	SVM-BPfinder; BPP	Y	N	N
15	*CDH23*	chr10:73491718T>C	-2	BPP	Y	N	N
16	*CDH23*	chr10:73548657T>G	-2	SVM-BPfinder; BPP	Y	N	N
17	*CEP164*	chr11:117244435T>A	-2	SVM-BPfinder; BPP	Y	N	N
18	*CEP164*	chr11:117244435T>A	-2	SVM-BPfinder; BPP	N	N	N
19	*CEP164*	chr11:117244435T>A	-2	SVM-BPfinder; BPP	N	N	N
20	*CEP164*	chr11:117244435T>A	-2	SVM-BPfinder; BPP	N	N	N

21	*CEP164*	chr11:117244435T>A	-2	SVM-BPfinder; BPP	Y	N	N
22	*CEP164*	chr11:117244435T>A	-2	SVM-BPfinder; BPP	Y	N	N
23	*CEP164*	chr11:117244435T>A	-2	SVM-BPfinder; BPP	Y	N	N
24	*CEP164*	chr11:117244435T>A	-2	SVM-BPfinder; BPP	Y	N	N
25	*CEP164*	chr11:117282757C>T	-2	BPP	N	N	N
26	*CEP164*	chr11:117282757C>T	-2	BPP	N	N	N
27	*CEP164*	chr11:117282757C>T	-2	BPP	Y	N	N
28	*CEP164*	chr11:117282757C>T	-2	BPP	Y	N	N
29	*CEP164*	chr11:117282757C>T	-2	BPP	Y	N	N
30	*CEP164*	chr11:117282757C>T	-2	BPP	Y	N	N
31	*CEP164*	chr11:117282757C>T	-2	BPP	Y	N	N
32	*CEP290*	chr12:88457911A>C	-2	SVM-BPfinder	N	N	N
33	*CERKL*	chr2:182412603T>A	A	BPP	Y	N	N
34	*CHM*	chrX:85236842T>G	A	BPP	N	N	N
35	*CLN3*	chr16:28493535G>C	-2	Mercer; BPP	Y	N	N
36	*CLN3*	chr16:28493535G>C	-2	Mercer; BPP	Y	N	N
37	*CLN3*	chr16:28493535G>C	-2	Mercer; BPP	Y	N	N
38	*CSPP1*	chr8:68070663A>T	A	BPP	Y	N	N
39	*EFEMP1*	chr2:56103893A>T	-2	SVM-BPfinder; BPP	N	N	N
40	*FSCN2*	chr17:79503144T>C	-2	SVM-BPfinder; BPP	N	N	N
41	*FSCN2*	chr17:79503144T>C	-2	SVM-BPfinder; BPP	Y	N	N
42	*FSCN2*	chr17:79503144T>C	-2	SVM-BPfinder; BPP	Y	N	N
43	*GNPTG*	chr16:1412428A>G	A	All	N	N	N
44	*GPR179*	chr17:36490765A>G	-2	SVM-BPfinder; BPP	N	N	N
45	*GPR179*	chr17:36491164T>C	A	SVM-BPfinder	Y	N	N
46	*GPR179*	chr17:36493112T>A	A	SVM-BPfinder; BPP	N	N	N
47	*GPR179*	chr17:36493112T>A	A	SVM-BPfinder; BPP	Y	N	N
48	*GPR179*	chr17:36493112T>A	A	SVM-BPfinder; BPP	Y	N	N

49	*GPR179*	chr17:36493112T>A	A	SVM-BPfinder; BPP	Y	N	N
50	*GPR98*	chr5:89938443A>T	A	SVM-BPfinder	N	N	N
51	*GPR98*	chr5:89938443A>T	A	SVM-BPfinder	Y	N	N
52	*GPR98*	chr5:90021338A>G	A	BPP	N	N	N
53	*GRM6*	chr5:178419123T>C	A	SVM-BPfinder	Y	N	N
54	*GUCY2D*	chr17:7919736T>C	-2	SVM-BPfinder; BPP	No diagnostic report available		
55	*GUCY2D*	chr17:7919736T>C	-2	SVM-BPfinder; BPP	No diagnostic report available		
56	*HARS*	chr5:140053942T>C	A	All	Y	N	N
57	*HARS*	chr5:140057027A>G	-2	SVM-BPfinder; BPP	N	N	N
58	*IFT140*	chr16:1574938T>C	A	BPP	Y	N	N
59	*IFT140*	chr16:1576105G>A	-2	Mercer; BPP	Y	N	N
60	*IFT140*	chr16:1621572T>A	A	SVM-BPfinder; BPP	N	N	N
61	*IMPG2*	chr3:100963651T>G	A	SVM-BPfinder; BPP	N	Y	N
62	*IMPG2*	chr3:100963651T>G	A	SVM-BPfinder; BPP	Y	N	N
63	*IMPG2*	chr3:100963651T>G	A	SVM-BPfinder; BPP	Y	N	N
64	*IMPG2*	chr3:100964972T>C	A	SVM-BPfinder; BPP	N	Y	N
65	*KIAA1549*	chr7:138554540T>C	A	BPP	Y	N	N
66	*KIAA1549*	chr7:138593884T>A	A	SVM-BPfinder	N	N	N
67	*LRP5*	chr11:68170919C>G	-2	Mercer; BPP	Y	N	N
68	*LRP5*	chr11:68205880A>G	A	All	N	N	N
69	*MERTK*	chr2:112702519A>T	A	SVM-BPfinder	Y	N	N
70	*MYO7A*	chr11:76900357A>C	A	SVM-BPfinder; BPP	N	N	N
71	*NPHP1*	chr2:110919286A>G	-2	SVM-BPfinder	N	N	N
72	*NPHP1*	chr2:110919286A>G	-2	SVM-BPfinder	No diagnostic report available		
73	*NPHP1*	chr2:110919286A>G	-2	SVM-BPfinder	Y	N	N
74	*NPHP1*	chr2:110919286A>G	-2	SVM-BPfinder	Y	N	N
75	*NPHP3*	chr3:132418941A>G	-2	SVM-BPfinder	Y	N	N
76	*PCDH15*	chr10:55973850A>T	-2	SVM-BPfinder	Y	N	N

77	*PDE6B*	chr4:648589C>T	-2	BPP	Y	N	N
78	*PDE6B*	chr4:659014C>T	-2	BPP	Y	N	N
79	*PDE6C*	chr10:95380375A>G	A	SVM-BPfinder	N	Y	N
80	*PEX1*	chr7:92116899T>C	A	SVM-BPfinder	Y	N	N
81	*PEX1*	chr7:92138745T>C	A	SVM-BPfinder; BPP	N	N	N
82	*PEX1*	chr7:92138745T>C	A	SVM-BPfinder; BPP	N	N	N
83	*PEX1*	chr7:92138745T>C	A	SVM-BPfinder; BPP	Y	N	N
84	*PEX1*	chr7:92151596G>A	-2	BPP	Y	N	N
85	*PEX7*	chr6:137187743A>G	A	SVM-BPfinder; BPP	Y	N	N
44	*PITPNM3*	chr17:6387120G>A	-2	Mercer; BPP	N	N	N
86	*PITPNM3*	chr17:6387120G>A	-2	Mercer; BPP	N	N	N
87	*PITPNM3*	chr17:6387120G>A	-2	Mercer; BPP	N	Y	N
88	*PITPNM3*	chr17:6387120G>A	-2	Mercer; BPP	Y	N	N
89	*PITPNM3*	chr17:6387120G>A	-2	Mercer; BPP	Y	N	N
90	*PITPNM3*	chr17:6387120G>A	-2	Mercer; BPP	Y	N	N
91	*PROM1*	chr4:16017902A>G	-2	SVM-BPfinder; BPP	Y	N	N
92	*PRPF31*	chr19:54627837T>G	A	Mercer	Y	N	N
86	*PRPF31*	chr19:54631645C>T	-2	Mercer; BPP	N	N	N
93	*PRPF31*	chr19:54631645C>T	-2	Mercer; BPP	Y	N	N
94	*PRPF31*	chr19:54631645C>T	-2	Mercer; BPP	Y	N	N
95	*PRPF8*	chr17:1558859T>G	A	Mercer	Y	N	N
96	*PRPH2*	chr6:42666266T>A	A	SVM-BPfinder; BPP	N	N	N
97	*PRPH2*	chr6:42666266T>A	A	SVM-BPfinder; BPP	N	N	N
98	*PRPH2*	chr6:42666266T>A	A	SVM-BPfinder; BPP	N	N	N
99	*PRPH2*	chr6:42666266T>A	A	SVM-BPfinder; BPP	N	Y	N
100	*PRPH2*	chr6:42666266T>A	A	SVM-BPfinder; BPP	Y	N	N
101	*PRPH2*	chr6:42666266T>A	A	SVM-BPfinder; BPP	Y	N	N
102	*PRPH2*	chr6:42666266T>A	A	SVM-BPfinder; BPP	Y	N	N

103	*RGR*	chr10:86007321A>G	A	BPP	N	N	N
104	*RGR*	chr10:86007321A>G	A	BPP	N	N	N
105	*RGR*	chr10:86007321A>G	A	BPP	N	N	N
106	*RGR*	chr10:86007321A>G	A	BPP	Y	N	N
107	*RPGRIP1*	chr14:21762816A>T	A	SVM-BPfinder	N	N	N
108	*RPGRIP1*	chr14:21792750G>A	-2	BPP	N	Y	N
61	*RPGRIP1L*	chr16:53670462A>G	-2	SVM-BPfinder; BPP	N	N	N
109	*RPGRIP1L*	chr16:53670462A>G	-2	SVM-BPfinder; BPP	N	N	N
110	*RPGRIP1L*	chr16:53670462A>G	-2	SVM-BPfinder; BPP	N	N	N
111	*RPGRIP1L*	chr16:53672354T>C	A	SVM-BPfinder; BPP	N	N	N
46	*RPGRIP1L*	chr16:53672354T>C	A	SVM-BPfinder; BPP	N	N	N
112	*RPGRIP1L*	chr16:53672354T>C	A	SVM-BPfinder; BPP	Y	N	N
113	*RPGRIP1L*	chr16:53672354T>C	A	SVM-BPfinder; BPP	Y	N	N
92	*RPGRIP1L*	chr16:53672354T>C	A	SVM-BPfinder; BPP	Y	N	N
114	*SAG*	chr2:234229263A>G	A	SVM-BPfinder	N	Y	N
82	*SEMA4A*	chr1:156128160A>C	A	SVM-BPfinder; BPP	N	Y	N
115	*TULP1*	chr6:35466259G>A	-2	BPP	N	Y	N
65	*USH1C*	chr11:17518381A>G	-2	SVM-BPfinder; BPP	Y	N	N
116	*USH2A*	chr1:216243678A>T	-2	SVM-BPfinder; BPP	Y	N	N
117	*USH2A*	chr1:216390937T>G	A	BPP	N	N	N
118	*USH2A*	chr1:216390937T>G	A	BPP	N	N	N
119	*WDR19*	chr4:39205244A>G	A	SVM-BPfinder	N	N	N

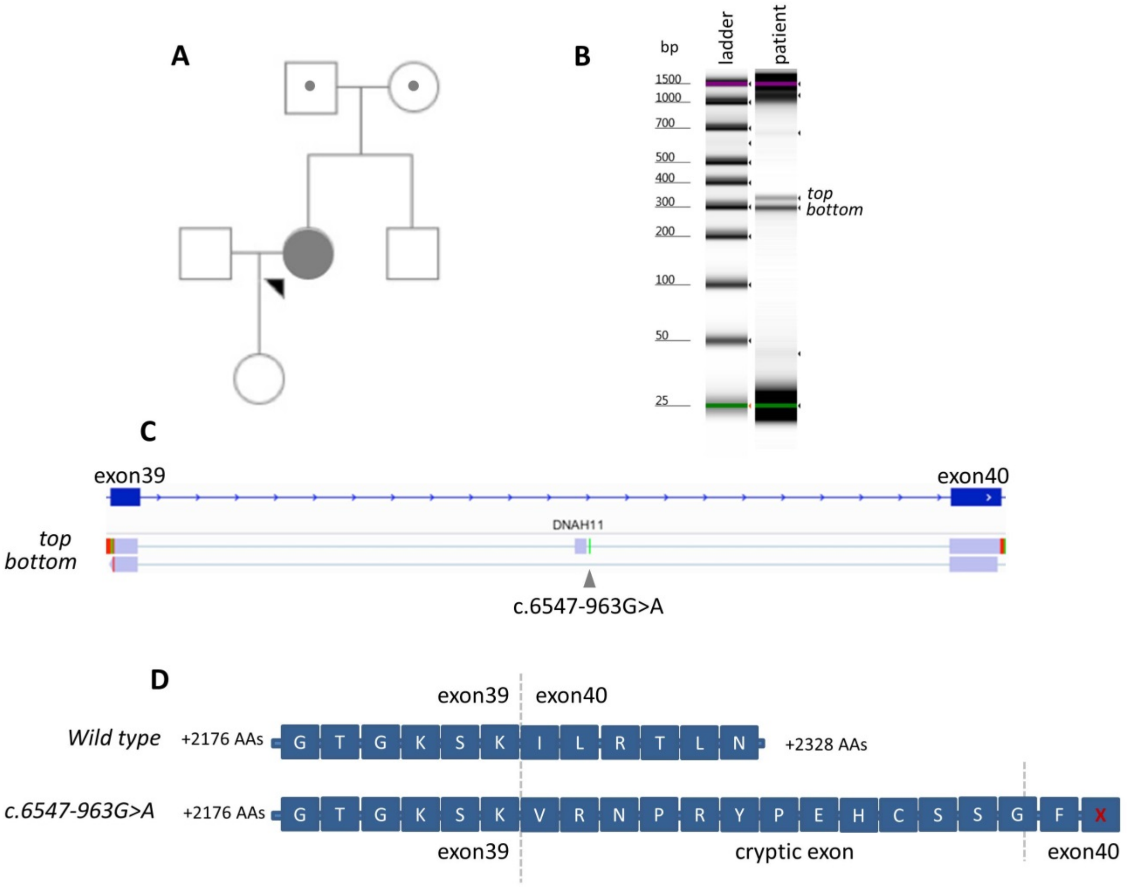
Variant (HGVSg)	Gene	Source of RNA	Phenotype	TPM	MRSD (M reads)
chr2:152,355,017G>T	NEB	Skeletal muscle	Nemaline myopathy	857.9	9.83
chr2:152,389,953A>C					
chr2:152,544,805C>T					
chrX:31,790,694-31,798,498invdel	DMD		Duchenne muscular dystrophy	24.84	79.4
chrX:32,274,692G>A			Myalgia, myoglobinuria		
chr2:179,446,219ATACT>A	TTN		Fetal akinesia	349.5	47.63
chr2:179,642,185G>A			Multi/minicore congenital myopathy		
chr21:47,409,881C>T	COL6A1		Collagen VI-related dystrophy	56.02	16.25
chr21:47,409,881C>T					
chr19:38,958,362C>T	RYR1		Congenital fiber-type disproportion	425.5	3.45
chr1:46,655,129C>A	POMGNT1		α-Dystroglycanopathy	29.26	6.01
chr17:41,199,655C>G	BRCA1	LCL	Inherited breast cancer susceptibility	19.985	217.19
chr17:41,246,879T>C					
chr17:41,246,879T>C					
chr17:41,246,879T>C					
chr17:41,258,551C>A					
chr13:32,945,238G>A	BRCA2			10.16	Unfeasible
chr13:32,969,074A>T					
chr19:33,892,776C>T	PEPD	Whole blood	Prolidase deficiency	18.89	28.31
chr20:35,526,363C>G	SAMHD1		Aicardi-Goutières syndrome	48.53	24.68
chr23:153,997,595G>A	MED13L		MRFACD	5.89	262.34

Supplementary Table S12. *Summary of pathogenic splicing events analyzed in this study.* All co-ordinates are given in relation to the GRCh37 genome build. TPM, transcripts per million; MRSD, minimum required sequencing depth.

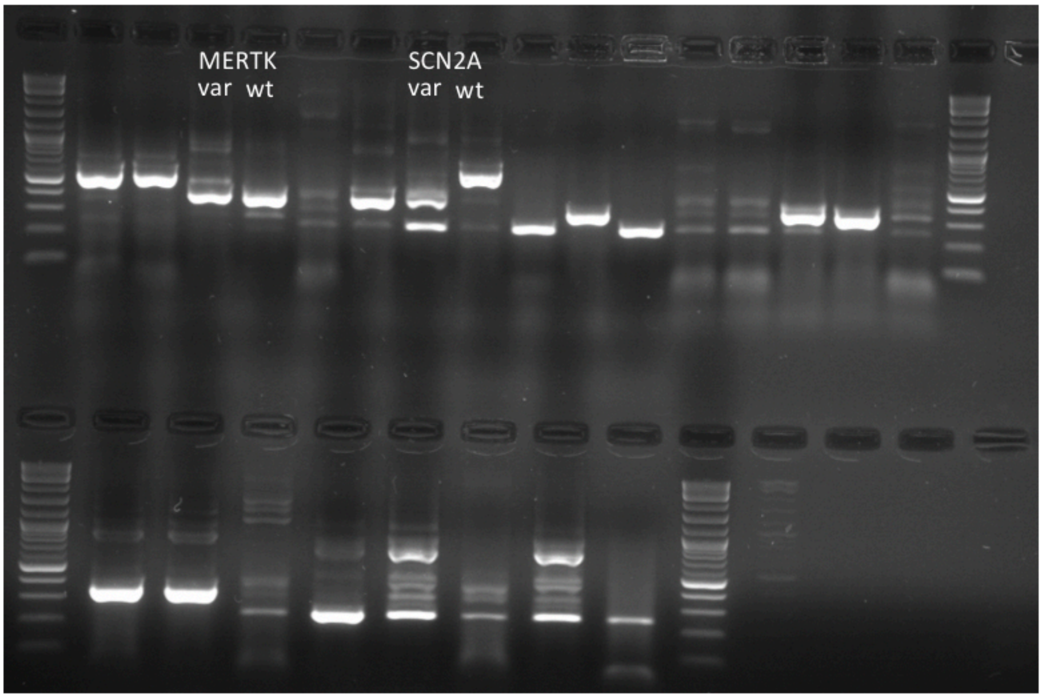
Tissue	No. samples	Source	Sequencing type	Usage
Blood	151	GTEx	75-bp paired end poly-A enrichment, Illumina	Generation of MRSD model, bootstrapping analysis of event counts
LCL	91			
Muscle	184			
Blood	1	Inhouse	150-bp paired end globin depletion, Illumina	Collation of known pathogenic mis-splicing events
	12		75-bp paired end poly-A enrichment, Illumina	Collation of known pathogenic mis-splicing events & MRSD model validation
LCL	20		150-bp paired end poly-A enrichment, Illumina	Collation of known pathogenic mis-splicing events
	4	Inhouse	75-bp paired end poly-A enrichment, Illumina	MRSD model validation
Muscle	52	Previously published data (3)	75-bp paired end poly-A enrichment, Illumina	Collation of known pathogenic mis-splicing events, downsampling of pathogenic events & MRSD model validation

Supplementary Table S13. *Summary of RNA-seq datasets utilized in this the generation and testing of the MRSD scoring framework.* RNA-seq datasets derived using different methodologies were used for various aspects of this section of work. All data used to *generate* the MRSD model was based on data from the GTEx consortium across all three analyzed tissues.

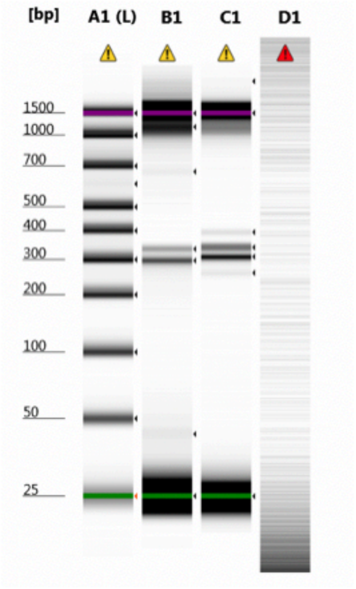
Appendix 2 – Supplementary Figures



Supplementary Figure S1. *DNAH11 c.6547-963G>A.* **(A)** Family pedigree showing the proband and her unaffected father and mother who carry heterozygous alleles of *DNAH11* c.8610C>G and c.6547-963G>A, respectively. **(B)** Gel electrophoresis results for the proband, visualized using an Agilent 2200 Tapestation (original unaltered images are presented in **Supplementary Figure S3**). RNA was reverse transcribed after extraction from whole blood and then amplified using primers specific to exons 39 and 40 of the *DNAH11* gene (NM_001277115.1). The caption shows two distinct cDNA amplicons in the proband sample separated by ~40 base pairs. **(C)** Integrated Genomic Viewer snapshot of the alignment of sequencing products to the human reference genome (GRCh37) showing the introduction of a 38 base pair cryptic exon (chr7:21,746,318-21,746,355) as a result of c.6547-963G>A. The *top* and *bottom* bands were sequenced after being cut from an agarose gel electrophoresis. **(D)** Impact of the cryptic exon on the translated protein. The cryptic exon shifts the reading frame and is expected to introduce a premature stop codon in exon 40, resulting in premature termination of protein synthesis, p.Ile2183Lysfs*15. Amino acids (AAs) are provided with single letter notations, with *X* indicating a stop codon. Vertical intersects indicate transition of the cDNA to the adjacent exon.

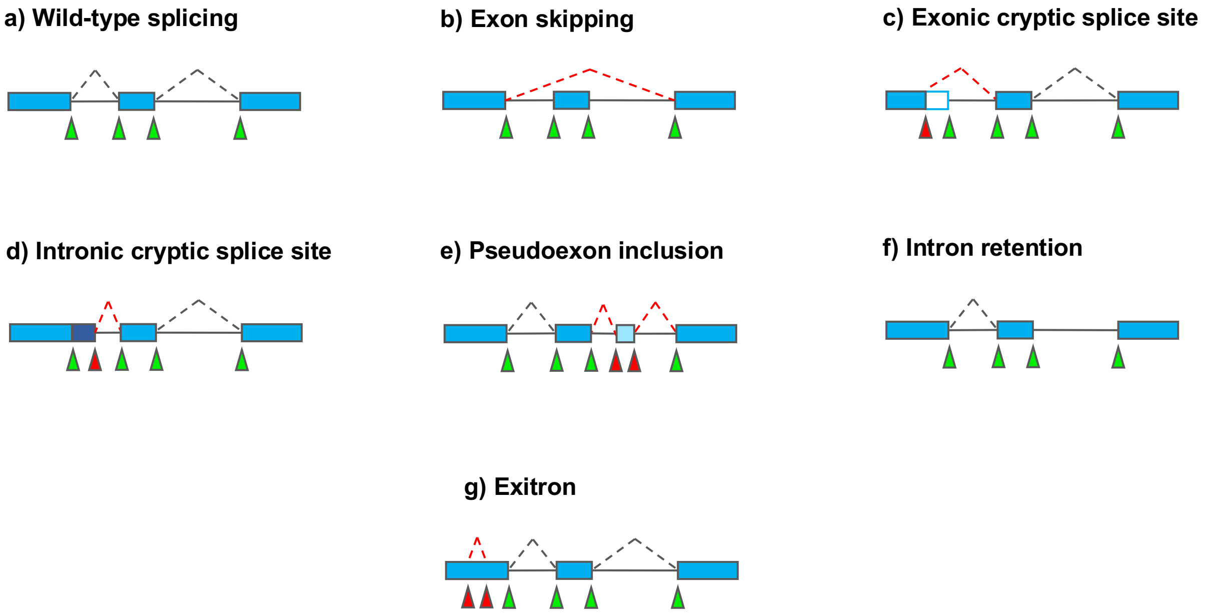


Supplementary Figure S2. *Uncropped gel electrophoresis photographs for data used in Figure 8.* The Invitrogen 1 Kb Plus Ladder was used for prediction of fragment size. The lanes used in **Figure 8** are indicated. Only the lanes indicated are relevant for MERTK c.2486+6T>A and SCN2A c.2919+3A>G.

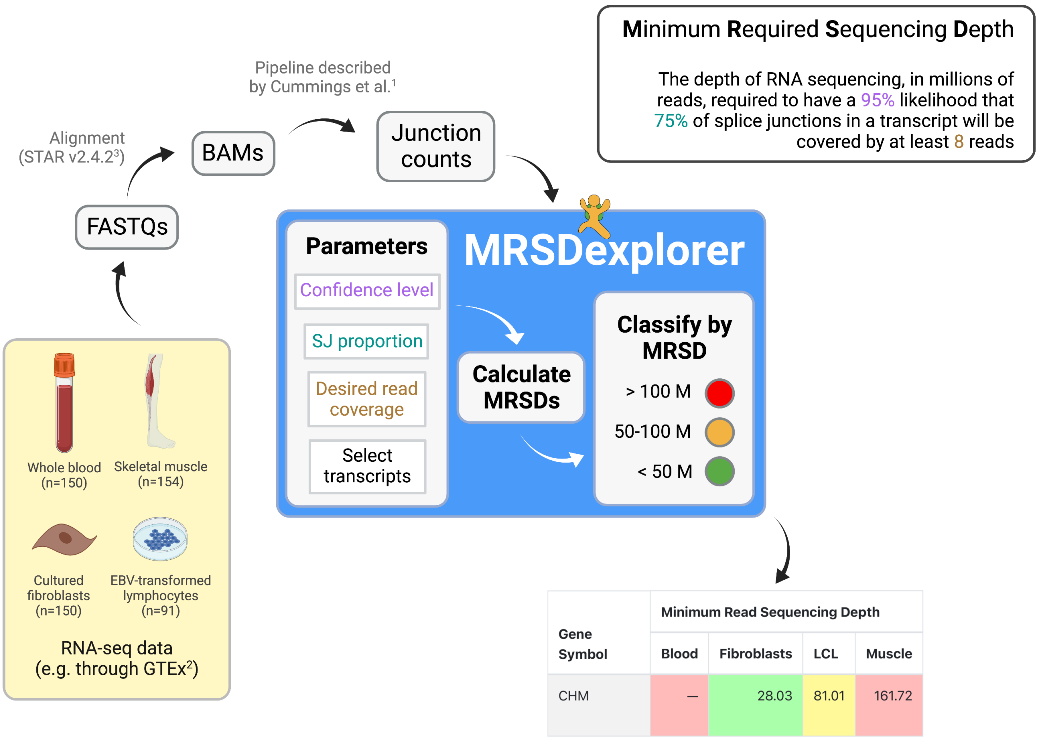


A1 – ladder
B1 - patient (used in Supp Figure 1)
C1 – patient (repeat)
D1 – empty

Supplementary Figure S3. *Uncropped images from the Agilent Tapestation showing cDNA bands amplified from patient sample.* The images shown are the default images from the Agilent 2200 Tapestation system, with band concentrations scaled to each individual sample.

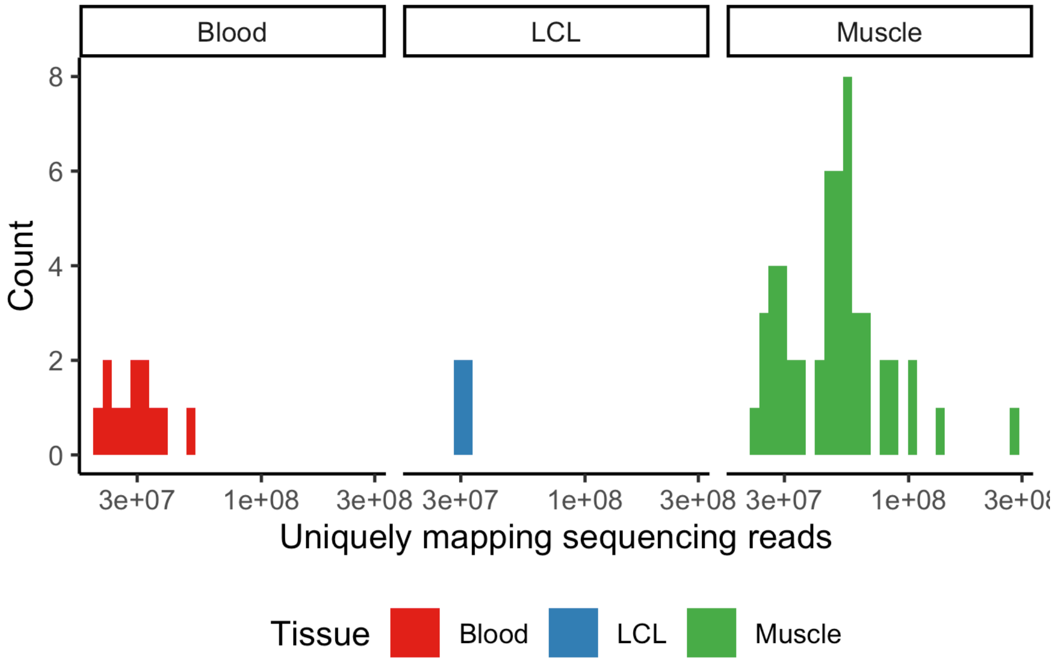


Supplementary Figure S4. *Categories of potentially pathogenic splicing events and their representation in analytical pipeline output.* Disruption of (**a**) wild-type splicing may lead to (**b**) skipping of one or more exons, the creation of novel splice sites in (**c**) exonic or (**d**) intronic regions that may outcompete the canonical sites, or result in (**e**) the generation of an intronic pseudoexon. (**f**) Splicing may be abrogated completely, leading to total retention of the intron. (**g**) Within longer exons, creation of a novel splice site may lead to a so-called "exitron", whereby a central portion of the exon is absent from the final transcript. Green triangles indicate canonical splice sites; red triangles indicate non-canonical sites.

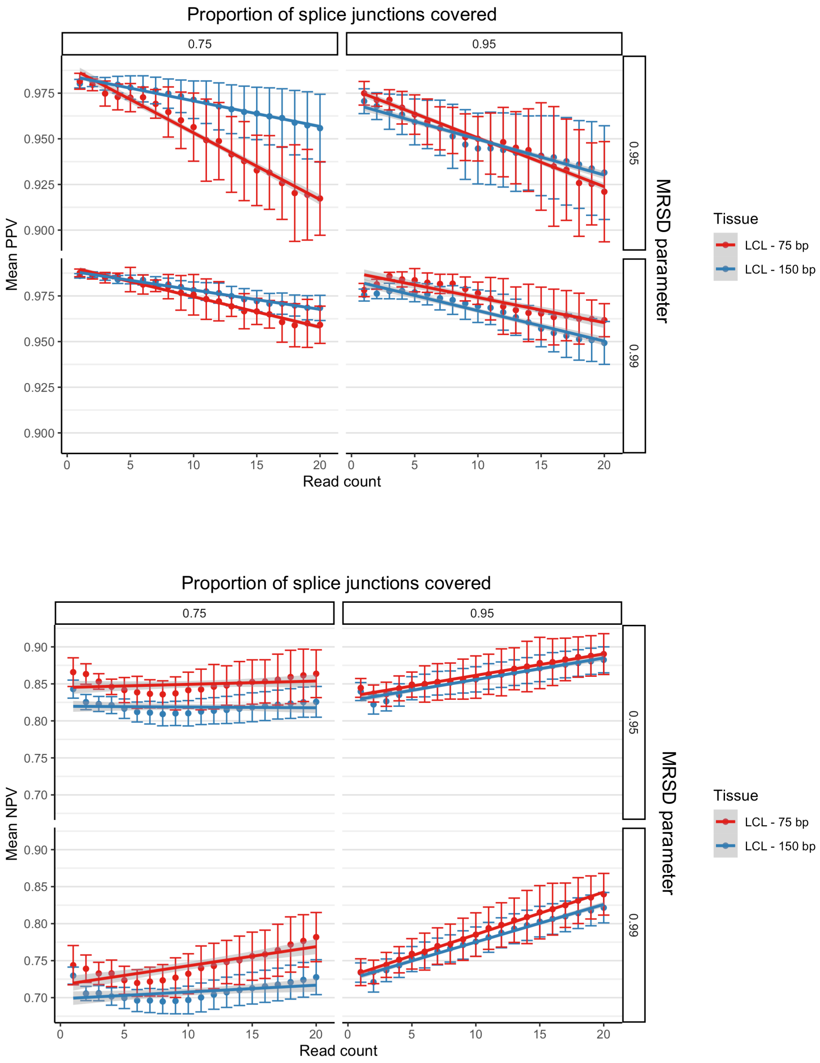

Minimum Required Sequencing Depth
The depth of RNA sequencing, in millions of reads, required to have a 95% likelihood that 75% of splice junctions in a transcript will be covered by at least 8 reads
Pipeline described by Cummings et al.[1]
Alignment (STAR v2.4.2[3])
BAMs
Junction counts
FASTQs
MRSDexplorer
Parameters
Confidence level
SJ proportion
Desired read coverage
Select transcripts
Calculate MRSDs
Classify by MRSD
> 100 M
50-100 M
< 50 M
Whole blood (n=150)
Skeletal muscle (n=154)
Cultured fibroblasts (n=150)
EBV-transformed lymphocytes (n=91)
RNA-seq data (e.g. through GTEx[2])

Gene Symbol	Minimum Read Sequencing Depth			
	Blood	Fibroblasts	LCL	Muscle
CHM	—	28.03	81.01	161.72

Supplementary Figure S5. *Workflow for MRSD score generation.* Users can create their own MRSD scores using the code provided online at https://github.com/mcgm-mrsd/mrsd-explorer. Starting with a set of RNA-seq samples, reads are aligned and the split reads counted using an established pipeline. Then, using our bespoke Python scripts, users can generate their own predictive scores (using parameters of their choice) and classify transcripts according to the level of sequencing required to obtain the specified coverage. Alternatively, users are free to investigate pre-computed scores for all GENCODE v19 genes across four tissues (whole blood, skeletal muscle, cultured fibroblasts and lymphoblastoid cell lines, or LCLs) at our web portal: http://mcgm-mrsd.github.io/

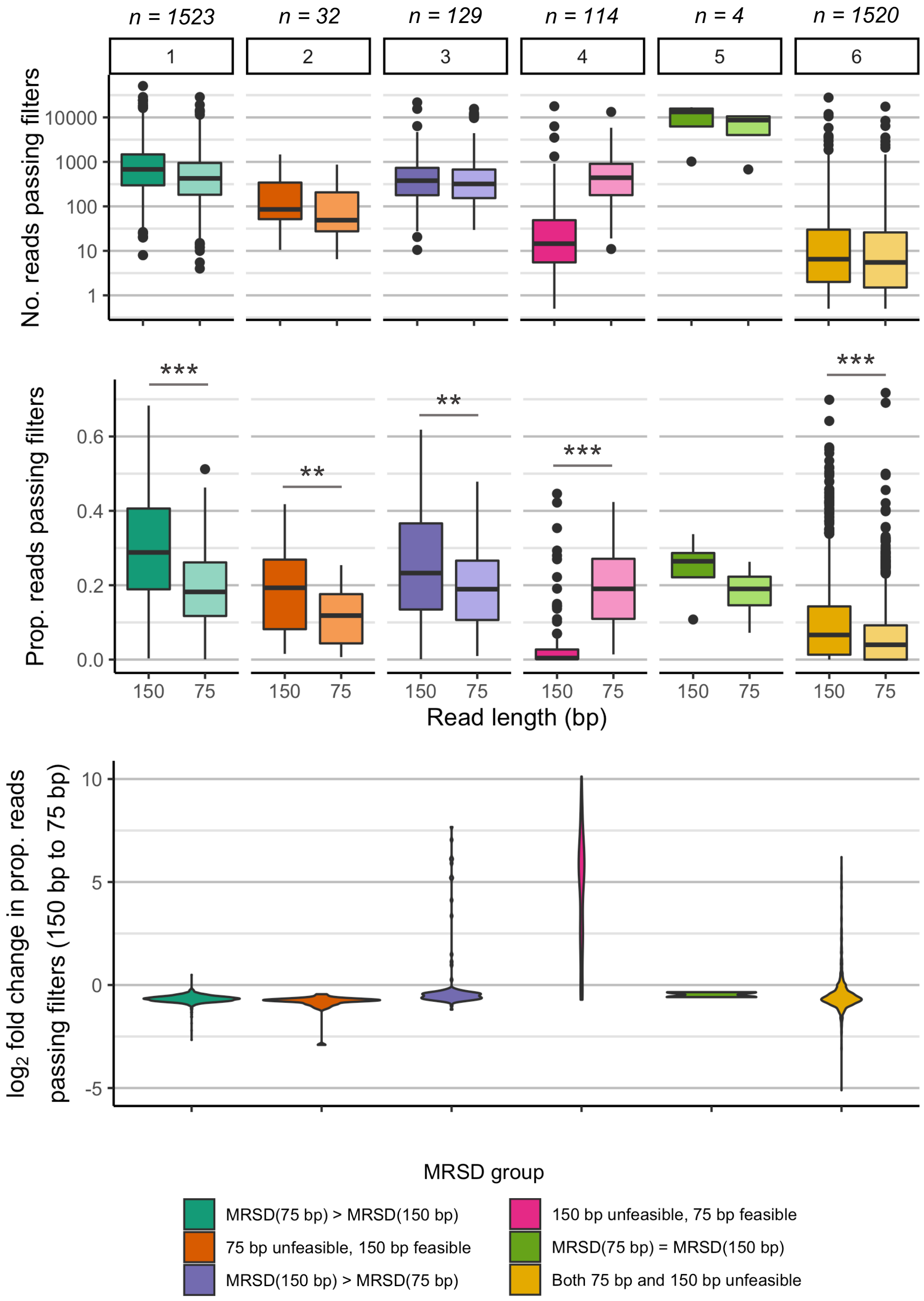


Supplementary Figure S6. *Sequencing depths of RNA-seq samples used for evaluation of MRSD model accuracy.* Whole blood (*n* = 12), LCL (*n* = 4) and skeletal muscle (*n* = 52) RNA-seq samples were derived from in-house or previously published data (Cummings et al., 2017) for validation of the MRSD model efficacy. Sequencing depths across the three tissues ranged from 20.6-281.5 M uniquely mapping reads.

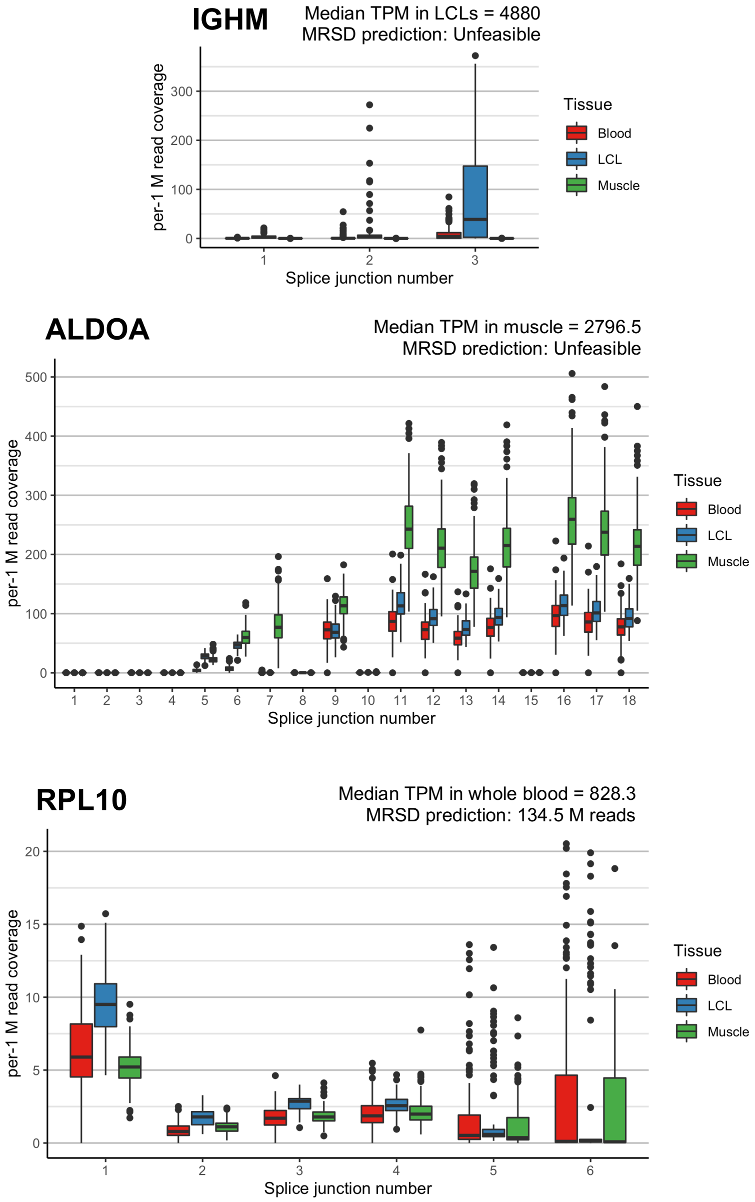


Supplementary Figure S7. *Effect of varying sequencing read length on MRSD model performance.* Despite being derived from 75 bp paired end RNA-seq data, MRSD scores show similar performance when applied to 75 or 150 bp paired end read-based RNA-seq, both in terms of (**top**) PPV and (**bottom**) NPV. When specifying 75% splice junction coverage, MRSD PPV is generally higher when the model is applied to 150 bp read-based data. This likely reflects the fact that junctions predicted to be sufficiently covered by 75 bp reads will be more likely to be sufficiently covered by reads of greater length, and so positive predictions are more likely to hold true when applied to longer-read data. We also observe that NPV for 150 bp read datasets is lower than that for 75 bp across all 4 parameter combinations; conversely to PPV, this is possibly because transcripts not sufficiently covered by 75 bp reads are more likely to be sufficiently covered by 150 bp reads, thus making negative predictions less likely to hold true in longer-read data. In most cases, differences in model performance between 75 and 150 bp is low, suggesting MRSD may, in some cases, provide a suitable approximation of transcript coverage in RNA-seq datasets with read lengths different to those used to construct the model.

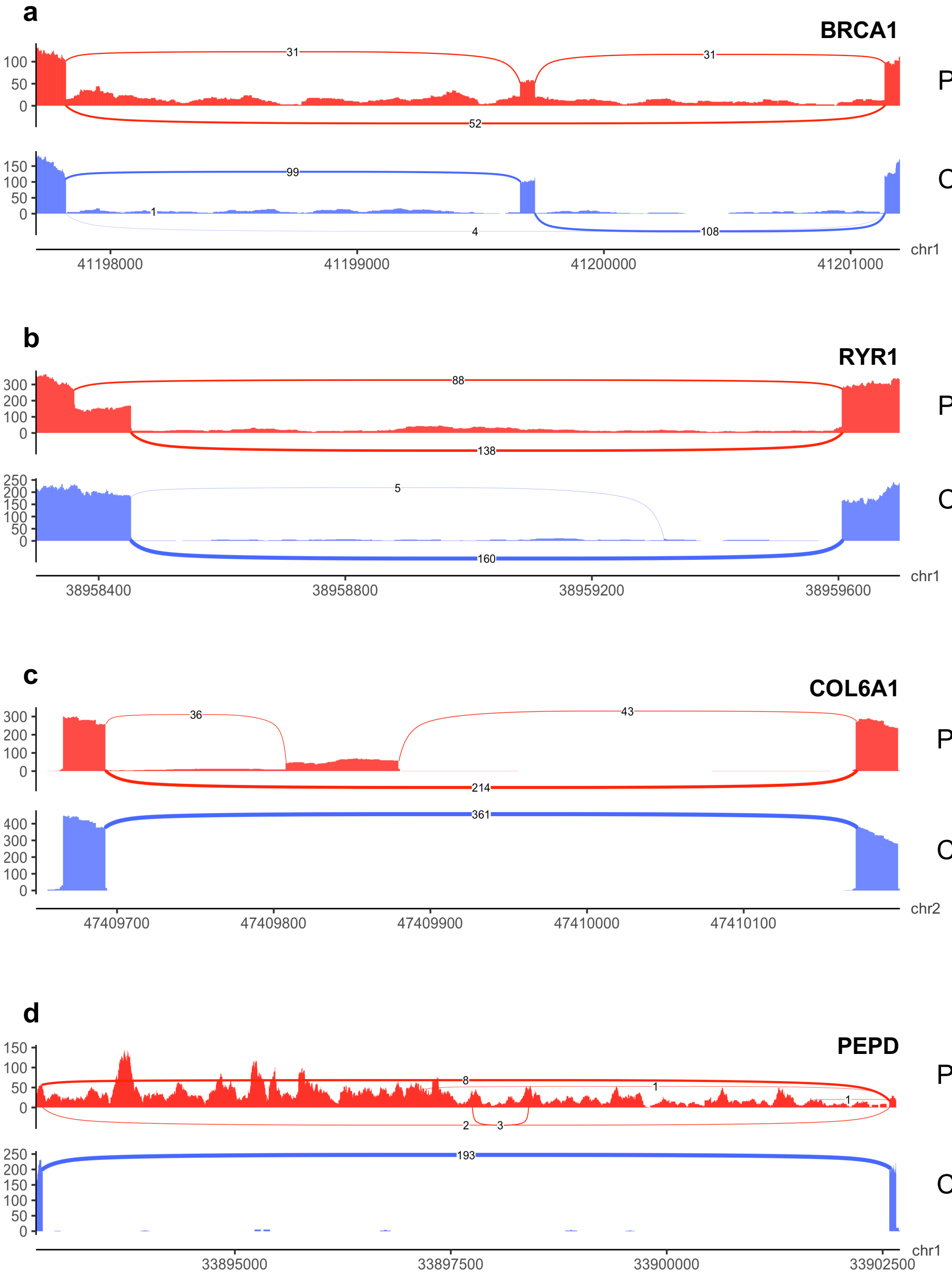
Supplementary Figure S8



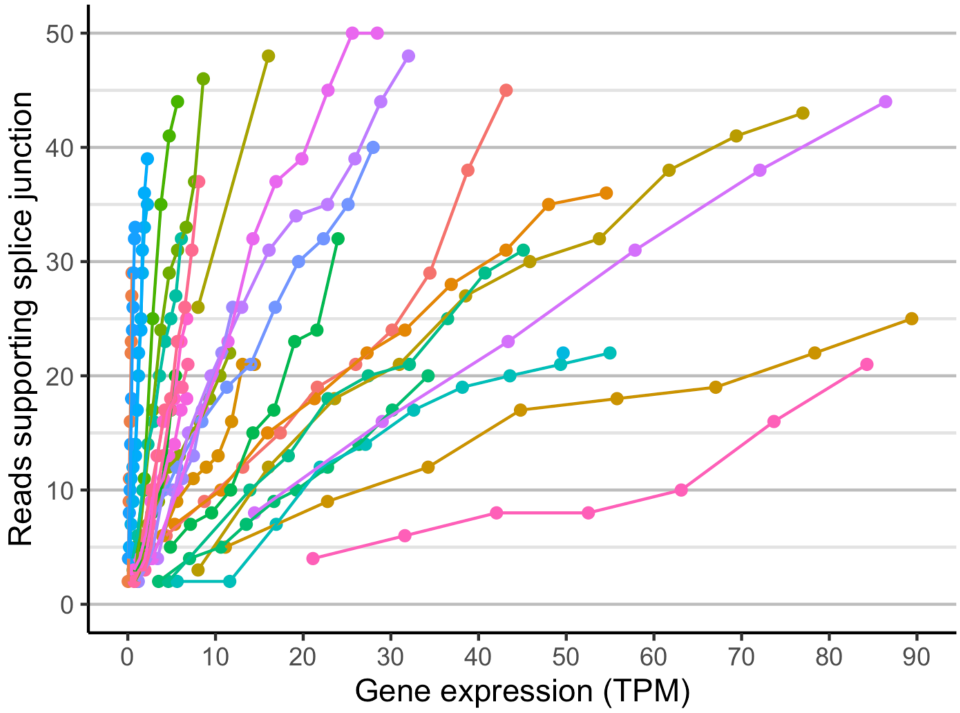
Supplementary Figure S8. *MRSD scores are generally lower when derived from RNA-seq runs of longer read length.* MRSD predictions generated from 20 LCL-based 150 bp RNA sequencing runs were compared against those generated following trimming of the same reads to a maximum of 75 bp. For 45.8% (1520/3322) of disease-associated genes, coverage was too poor to generate an MRSD score regardless of read length (group 6), while MRSDs could be generated but remained the same regardless of read length for just 4/3322 (0.12%) genes (group 5). Intuitively, of the 54.1% (1798/3322) of genes for which at least one dataset allowed MRSD generation, a higher MRSD was observed in the 75 bp dataset for 86.5% (1555/1798, groups 1 and 2). However, for the remaining 13.5% of genes (243/1798, groups 3 and 4), a lower MRSD score was generated using the 75 bp dataset than the 150 bp dataset. For many of these genes, it was determined that a shortening of the reads actually improved their quality to the extent that they were more likely to pass the enforced quality filters – namely, that a mapping event must be the primary alignment, that the read must map successfully (i.e. must have a mapping quality of 60) and that the read must be a split read. We observed that in group 4, comprising genes for which MRSD generation is unfeasible using the 150 bp dataset but feasible using the 75 bp dataset, there was a median 36.8-fold increase in the number of reads passing these read filters following trimming (**bottom**). Further work is needed to investigate alternative causes of this counter-intuitive pattern, and to determine whether the discarding of the longer reads represents an artefactual drawback to the read filtering process, or an effective way to filter reads for quality that is missed using shorter reads.



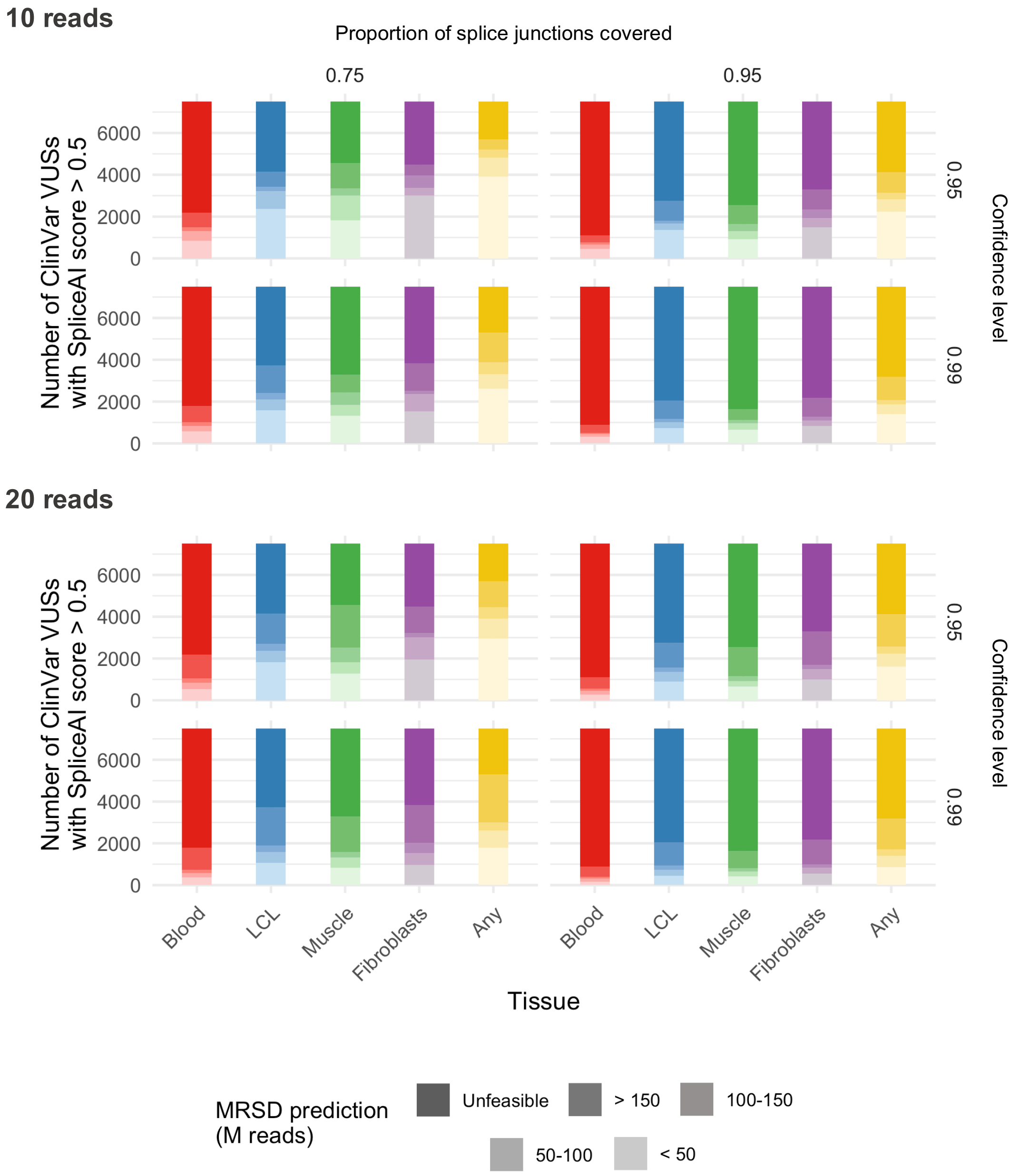
Supplementary Figure S9. *Evidence for 3' sequence bias confounding the use of TPM as a guiding RNA-seq metric.* Analyzing the number of reads (per 1 M uniquely mapping input reads) mapping to individual splice junctions within three genes with substantial TPM-MRSD discrepancy demonstrates that highly expressed genes may exhibit biased coverage of splice junctions. For IGHM (**top**) and ALDOA (**middle**) in LCLs and muscle, respectively, a sufficient proportion of junctions towards the 3' end of the transcript have no read support in a sufficient number of patients, resulting in an MRSD prediction of "unfeasible", despite high coverage of other junctions within the same transcript. Coverage of the final two splice junctions in RPL10 (**bottom**) in LCL-based RNA-seq data is low but not non-zero in many patients, giving a feasible but high MRSD prediction. In some cases, this bias may result from artefacts of library preparation, or may possible reflect genuine isoform shifts in the given tissue. Higher splice junction numbers represent junctions closer to the 3' end of transcripts.



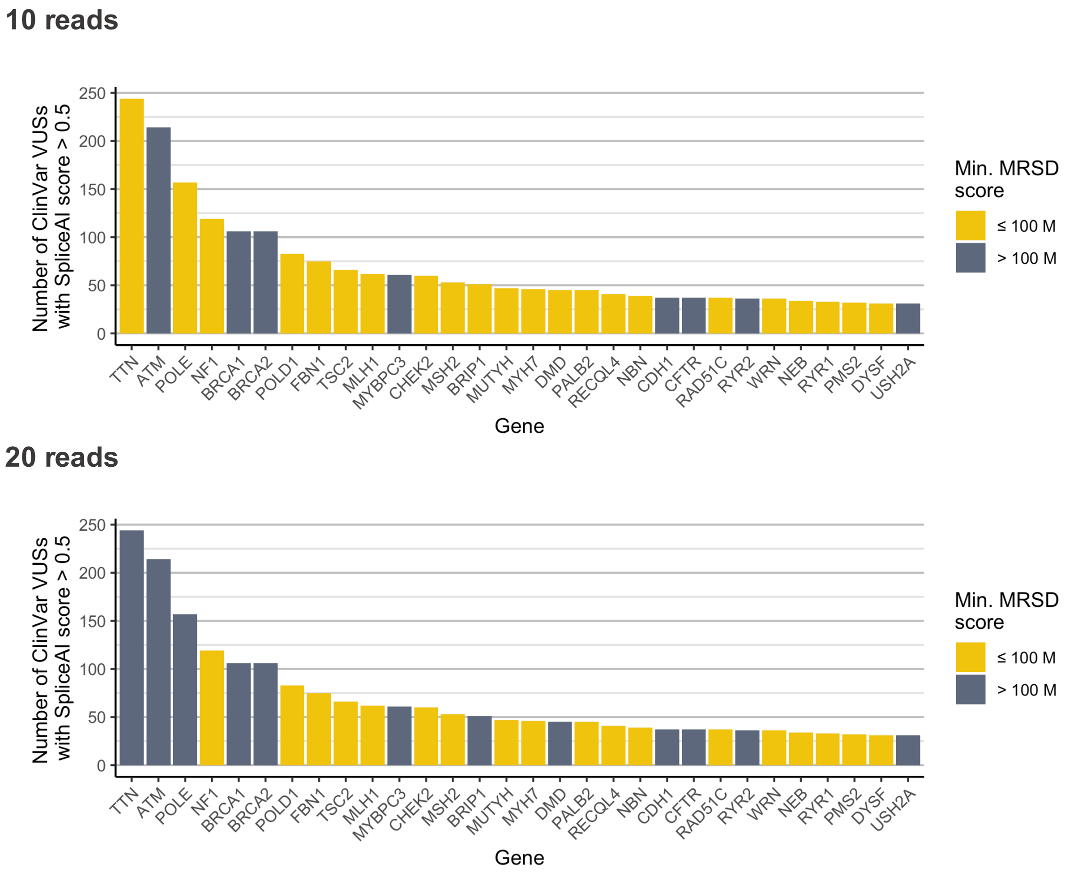
Supplementary Figure S10. *Exemplar events identified during pathogenic splice event analysis.* Selected Sashimi plots for (**a**) exon skipping, (**b**) exonic splice gain, (**c**) pseudoexonization and (**d**) intron retention events identified as the cause of disease in our patient datasets. The presence of aberrant splice junctions with outlying event metrics allowed flagging of these as potentially pathogenic. For (**d**), the intron retention event was identified from the 2 reads supporting usage of an extremely weak alternative splice acceptor four bases downstream of the abrogated canonical acceptor; however, in the absence of any aberrant splicing events, intron retention events are more difficult to identify from RNA-seq data using current bioinformatics pipelines.



Supplementary Figure S11. *Relative gene expression level does not reflect the raw read coverage of transcript splice junctions.* When simulating decreased gene expression by downsampling reads in genes containing novel splicing events identified in upstream analysis, it emerged that expression of a gene (in transcripts per million, TPM) does not directly correlate with the number of reads supporting splice junctions in that gene. Among the events supported by 8 reads, for example, gene expression ranged from 0.17-52 TPM. This may be accounted for by variation in the proportion of transcripts containing the event, variation in the coverage across the length of a transcript (as shown in **Supplementary Figure S9**), or variation in the depth to which a sample has been sequenced. Thus, when specifying a metric threshold above which we expect splice aberration to be observable, relative expression level may not appropriately represent expected read support. Axes are limited for ease of visualization.



Supplementary Figure S12. *Increasing specified read coverage reduces the number of ClinVar variants that can be analyzed.* Similarly to **Figure 19a** (main text), we generated MRSD scores for genes harboring predicted splice-impacting ClinVar variants (SpliceAI score ≥ 0.5; Jaganathan et al., 2019) using more stringent read coverage parameters (10 and 20 reads). We observed only a small reduction in the number of ClinVar variants in low-MRSD genes when specifying 10 reads (24.9-64.0% dependent on parameters). Specifying 20 read coverage, however, drastically reduces the percentage of ClinVar variants in low-MRSD genes to 18.7-52.0%.



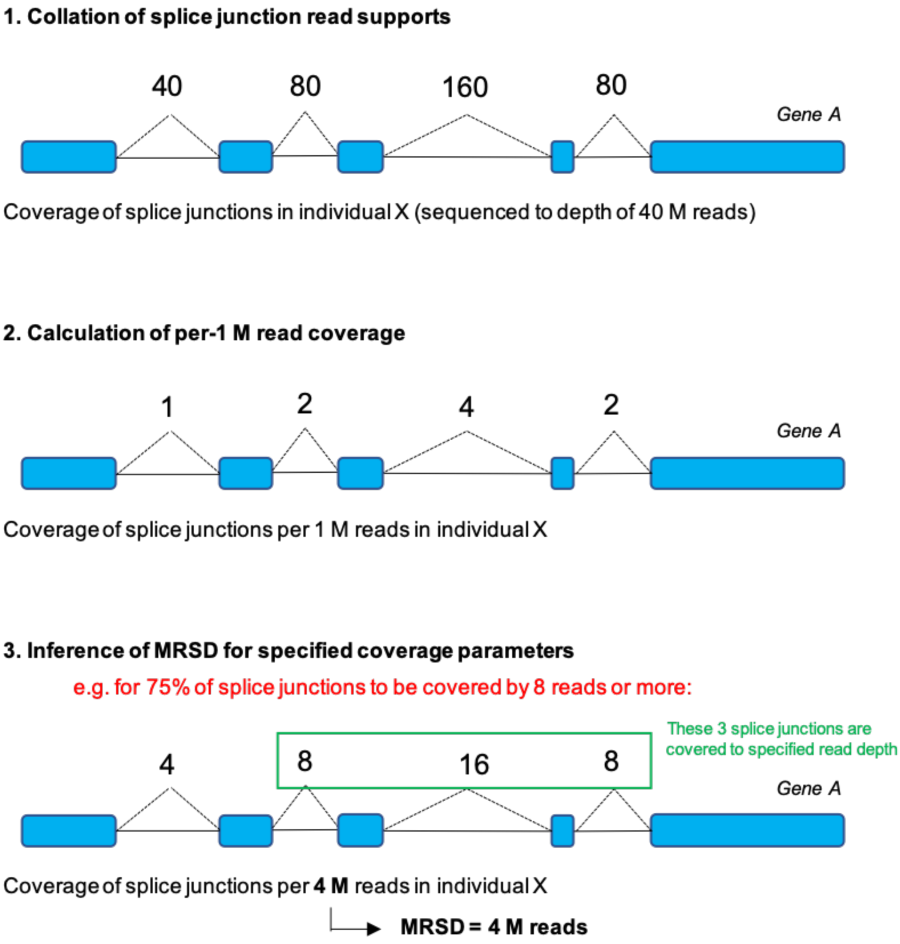
Supplementary Figure S13. *Increasing specified read count removes highly VUS-prone genes from the scope of analysis.* Similarly to **Figure 19b**, we looked among the 30 genes harboring the most predicted splice-impacting ClinVar variants and considered how many were low-MRSD in at least one of the four investigated tissues when specifying increasing levels of read coverage. Only one extra gene, *ATM*, becomes ostensibly high-MRSD when specifying a 10-read coverage parameter when compared with the 8-read coverage data (**Figure 19b**). However, by specifying a 20-read level of coverage, a further four genes are removed from the scope of analysis, leaving 18/30 (60%) still considered low-MRSD.

Appendix 3 – Supplementary Methods

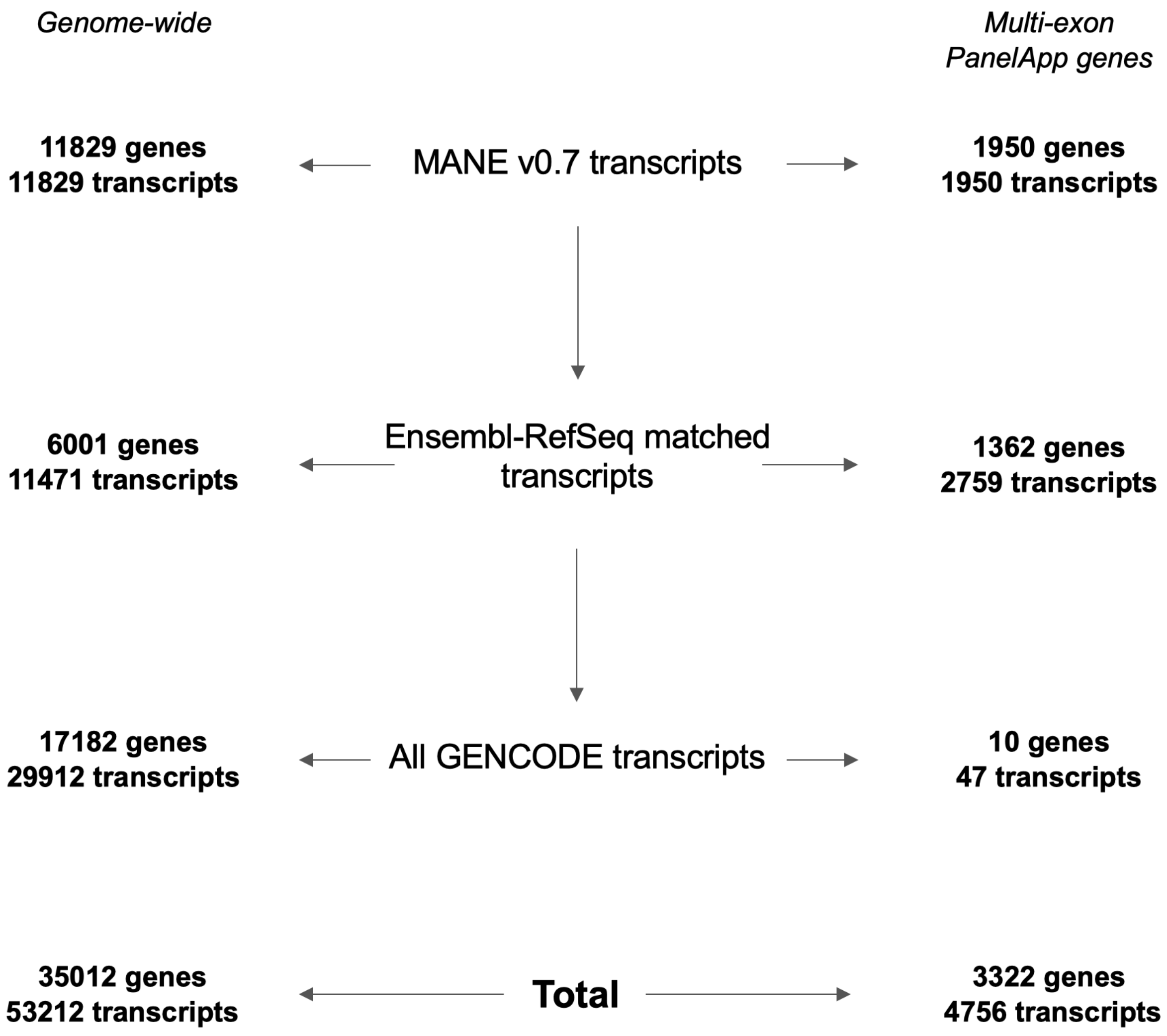
Supplementary Methods 1. *Illustration of MRSD calculation methodology.* MRSD scores utilize the level of read coverage supporting the existence of splice junctions in control RNA-seq datasets to predict the depth of sequencing required to achieve a specified level of splice junction coverage in a transcript of interest. For a given transcript in a given individual:

1. Read coverage values are collated across all splice junctions in the transcript model (with a single transcript assigned to each gene if investigating at the gene level, see **Supplementary Methods 2**)
2. Each of these values is divided by the sequencing depth – by default defined as the number of uniquely mapping sequencing reads (in millions of reads) to produce a per-1 M read coverage value for each junction
3. The desired level of read coverage is divided by the per-1 M read coverage value of the splice junction with the *X*'th percentile lowest read coverage, which gives the depth of sequencing that would be required for X% of junctions to be covered with the desired number of reads or higher. This figure is the sample-specific MRSD.

The sample-specific MRSDs are collated across all control RNA-seq samples, and a global MRSD is then derived by taking the m-th percentile highest prediction from among these; m is termed the MRSD parameter, and represents the proportion of control RNA-seq samples for which sequencing at the returned MRSD would have sufficiently covered that gene. By extension, it is also an approximate measure of the likelihood that a subsequent RNA-seq run at the returned depth will yield the specified coverage.


1. Collation of splice junction read supports
40
80
160
80
Gene A
Coverage of splice junctions in individual X (sequenced to depth of 40 M reads)
2. Calculation of per-1 M read coverage
1
2
4
2
Gene A
Coverage of splice junctions per 1 M reads in individual X
3. Inference of MRSD for specified coverage parameters
e.g. for 75% of splice junctions to be covered by 8 reads or more:
These 3 splice junctions are covered to specified read depth
4
8
16
8
Gene A
Coverage of splice junctions per **4 M** reads in individual X
MRSD = 4 M reads

Supplementary Methods 2. *Tiering methodology for selection of transcripts for MRSD generation.* To calculate MRSD values for all protein-coding genes, a single transcript model was established for each gene. Firstly, transcripts present in the MANE v0.7 curated transcript set were selected for genes where these existed, provided the co-ordinates of all splice junctions in that transcript (given in relation to the GRCh38 reference genome) mapped back to known junctions in build GRCh37. For genes where these conditions were not met, transcript models were formed from the union of all junctions present in all RefSeq transcripts listed for that gene on Ensembl BioMart. Finally, for any genes lacking a corresponding RefSeq transcript(s), a transcript model was derived consisting of the union of all junctions present in all transcripts assigned to that gene in the GENCODE v19 annotation.


Genome-wide
Multi-exon PanelApp genes
11829 genes 11829 transcripts
MANE v0.7 transcripts
1950 genes 1950 transcripts
6001 genes 11471 transcripts
Ensembl-RefSeq matched transcripts
1362 genes 2759 transcripts
17182 genes 29912 transcripts
All GENCODE transcripts
10 genes 47 transcripts
35012 genes 53212 transcripts
Total
3322 genes 4756 transcripts

Supplementary Methods 3. *Tissue-specific criteria for filtering of high-quality GTEx control RNA-seq datasets.* Filtering of GTEx controls was conducted to select the highest quality samples based on the below tissue-specific parameters. Parameters were selected and adjusted on a tissue-by-tissue basis to exclude metric outliers and samples that may confound analysis of pathogenic splicing events (e.g. excluding cancer patients from LCL control cohorts, in which inherited breast cancer was studied). The corresponding column names in the GTEx v8 sample attribute (pht002743.v8) and subject phenotype (pht002742.v8) files are italicized.

Skeletal muscle (as listed in [1])

- RNA integrity number/RIN (*SMRIN*): between 6-9
- Sample ischemic time (*SMTSISCH*): <720 (i.e. <12 hours)
- Hardy scale (*DTHHRDY*): 0, 1 or 2, corresponding to sudden deaths
- Age (*AGE*): <50
    - Unless BMI <30

Whole blood

- Samples included in GTEx analysis freeze, corresponding to higher quality samples (*SMAFRZE*): not flagged EXCLUDE due to technical issues
- RIN (*SMRIN*): between 6-9
- Sample ischemic time (*SMTSISCH*): <0
- Hardy scale (*DTHHRDY*): 0, 1 or 2

EBV-transformed lymphocytes (LCLs)

- *SMAFRZE*: not flagged EXCLUDE due to technical issues
- RIN (*SMRIN*): > 9
- *MHCANCER5*, *MHCANCERC* and *MHCANCERNM* all 0 to eliminate all non-metastatic cancers and all cancers in the past 5 years or current
- *DTHHRDY*: 0, 1 or 2
- No reported history (*MHGENCMT*) of:
    - Breast cancer
    - Ovarian cancer
    - Pancreatic cancer
    - Prostate cancer
    - Colorectal cancer
    - No patients filtered out through this criterion

Cultured fibroblasts

- As for EBV-transformed lymphocytes, except with the addition of the following:
    - RIN (*SMRIN*) > 9.7
    - Uniquely mapping reads (*MPPDUN*): > 60 M