

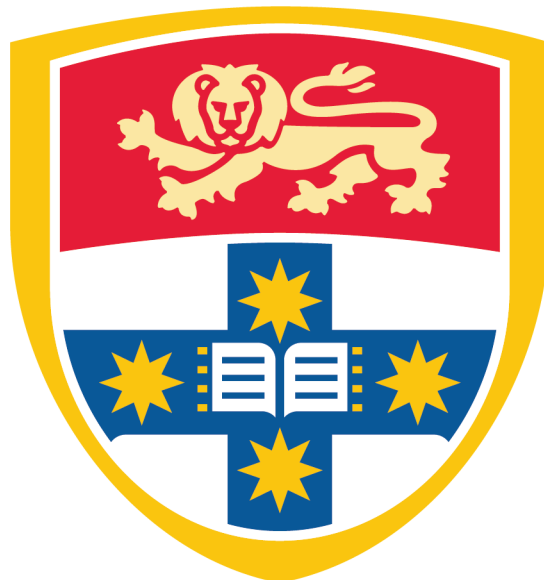
# **Using empirical evidence to predict if and how a DNA variant will disrupt RNA splicing in rare disorders.**

Ruebena Dawes

Primary supervisor: Professor Sandra T. Cooper

Auxiliary supervisors: Associate Professor Kristi Jones and Associate Professor  
Monkol Lek

*A thesis submitted to fulfil requirements for the degree of Doctor of Philosophy*



Kids Neuroscience Centre, Kids Research  
The Children's Hospital at Westmead, Westmead, NSW, Australia

Sydney Medical School, Faculty of Medicine and Health

The University of Sydney, Sydney, NSW, Australia

**2022**

## **Statement of Originality**

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Ruebena Dawes

Mary 2022

## Authorship Attribution Statement

Chapter 2 of this thesis was published as:

**Dawes, R.**, Joshi, H. & Cooper, S.T. Empirical prediction of variant-activated cryptic splice donors using population-based RNA-Seq data. *Nat Commun* **13**, 1655 (2022).

<https://doi.org/10.1038/s41467-022-29271-y>

I performed the data curation and analysis; I wrote the initial draft of the manuscript and took part in review and editing along with Himanshu Joshi and Sandra Cooper. I prepared all main and supplemental figures, and created the accompanying 40K-RNA web portal, with assistance from Himanshu Joshi with database management.

Chapter 3 of this thesis has been accepted as:

**Dawes R.**, Bournazos AM, Bryen SJ et al. SpliceVault predicts the precise nature of variant-associated mis-splicing. *Nat Genet (NG-AN59054R3)*. (2022).

Adam Bournazos and I contributed equally as joint first authors. I performed the data curation and analysis, creating the 300K-RNA database and cross-referencing with experimental data generated by Adam Bournazos. I prepared Figure 1C, Figures 2-4, Extended Data Figures 1-6 and results and methods sections pertaining to these figures, as well as being involved in writing and editing other manuscript sections along with other authors. Himanshu Joshi completed data curation of tissue-specific events required for Figure 3 and S4. I created the accompanying 300K-RNA web portal, with assistance from Himanshu Joshi with database management and tissue-specific capabilities.

Ruebena Dawes

May 2022

## Authorship Attribution Statement

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Professor Sandra T. Cooper

May 2022

### Competing Interests

Professor Sandra Cooper is director of Frontier Genomics Pty Ltd (Australia). Professor Cooper receives no remuneration (salary or consultancy fees) for this role. Frontier Genomics Pty Ltd (Australia) has no existing financial relationships that will benefit from publication of these data.



## **Acknowledgements**

Thank you to my supervisor Prof Sandra Cooper, who has invested so much time, support, and mentorship to my development as a scientist.

Thank you to Himanshu Joshi for providing invaluable and generous help with the technical side of this project, and for always being just a Slack message away.

Thank you to the rest of the team at the Kids Neuroscience Centre for providing such a friendly and supportive environment, and intellectually stimulating involvement in meetings.

I am grateful to have received financial support from the Australian Government Research Training Program scholarship and postgraduate Merit Award, as well as a stipend from Kids Neuroscience Centre.

Thank you also to my family, especially my mum Sharon Dawes for all the help with proofreading and for listening to all my presentations, and the million other things you've done to help me with this degree.

## **Abstract**

### **Background**

The diagnostic rate in Mendelian disorders continues to hover around 50% after genomic testing, meaning that around half of families and clinicians are left with no actionable answer. Variants affecting splicing motifs are particularly challenging to interpret. To conclusively link a splicing variant to disease it's necessary to determine the consequences of altered splicing on the final mRNA transcript and subsequent protein. Consequently, most probable splicing variants are classified as VUS and unactionable.

A range of powerful but opaque algorithms have proliferated for predicting whether a variant alters splicing. Many are based on machine learning and deep learning, with the data and features used to make a specific prediction usually unavailable to be verified and weighted by clinicians. Without detailing the nature and source(s) of evidence used to make each prediction, these algorithms are relegated to the lowest evidence weighting according to globally-accepted, gold standard variant classification rules, established by the ACMG-AMP.

In addition, most algorithms currently make no attempt to predict mis-splicing outcomes which will occur as the result of a variant, meaning that bespoke functional testing is still required to discover the variant impact on pre-mRNA splicing and allow ACMG-AMP guided variant reclassification for a definitive molecular diagnosis.

There is an urgent need for evidence-based, clinically-validated tools for pathology interpretation of splicing variants.

### **Aims**

To bridge the gap between data science and genetic pathology, by developing methods based on empirical evidence to predict if and how a DNA variant will disrupt RNA splicing in rare disease.

To determine empirical features that accurately inform:

## Abstract

- 1) spliceosomal selection of a cryptic-donor, in preference to the ‘authentic-donor’ (positioned at the exon-intron junction), and other nearby decoy-donors (any GT or GC) that are not used by the spliceosome, and
- 2) The mis-splicing events which will occur because of a variant precluding use of the authentic-donor or authentic-acceptor.

## Methods

We use empirical and clinically relevant data to define and evaluate measurable features enriched in (1) cryptic-donors selected by the spliceosome vs decoy-donors (any GT/GC motif) which were not selected by the spliceosome and (2) mis-splicing events (exon skipping or cryptic activation) which occurred because of a splicing variant.

## Results

For 1) we evaluated the use of current algorithms to show that while intrinsic splice-site strength and proximity to the authentic-donor strongly influence spliceosomal selection of a cryptic-donor, these factors alone are not sufficient for accurate prediction.

For 2) we find that natural, stochastic mis-splicing events seen in population-based RNA-Seq are remarkably prescient of the mis-splicing events that will occur predominantly after the inactivation of an authentic splice site.

## Conclusions

We’ve created an accurate, evidence-based method to predict the nature of variant -induced mis-splicing. The ability to confidently predict the outcome of a splicing variant is a major step forward which will greatly aid in genetic diagnosis of families with Mendelian disorders.

## Conferences and other proceedings

**Dawes, R.,** Bournazos, A. Bryen, S.J., Bommireddipalli, S., Joshi, H., Cooper, S.T.  
SpliceVault: predicting the precise nature of variant-associated mis-splicing. Invited Oral  
Presentation: 2022 Personalized Medicine Gordon Research Seminar, June 25-26, Ventura,  
California.

**Dawes, R.,** Bournazos, A. Bryen, S.J., Bommireddipalli, S., Joshi, H., Cooper, S.T.  
SpliceVault: predicting the precise nature of variant-associated mis-splicing. Poster  
presentation: RNA 2022 May 31-June 5, Boulder, Colorado.

**Dawes, R.,** Bournazos, A. Bryen, S.J., Bommireddipalli, S., Joshi, H., Cooper, S.T.  
SpliceVault: predicting the precise nature of variant-associated mis-splicing. Poster  
presentation: Australasian RNA Biology and Biotechnology Conference (A-RNA) 2022,  
May 15-18, Thredbo NSW.

**Dawes, R.,** Joshi, H., Cooper, S.T. SpliceVault: predicting the precise nature of variant-  
associated mis-splicing. Poster presentation: Australasian RNA Biology and Biotechnology  
Conference (A-RNA), 2022 May 15-18, Thredbo NSW.

**Dawes, R.,** Joshi, H., Cooper, S.T. 90% of cryptic-donors activated in genetic disorders are  
present in variant-free RNA-Seq samples. Poster presentation: American Society of Human  
Genetics Annual Meeting, 2021 October 18-21, virtual.

**Dawes, R.,** Joshi, H., Cooper, S.T. Predicting cryptic splice site selection in genetic  
disorders. Poster presentation: RNA 2021, May 25-June 4, virtual.

**Dawes, R.,** Joshi, H., Bryen S., Bournazos, A., Cooper, S.T. Features that determine donor  
cryptic splice site selection in genetic disorders. Poster presentation: Precision Medicine: A  
Revolution in Patient Care Conference 2021 May 20-21, Sydney, NSW.

**Dawes, R.,** Cooper S.T. Features that determine 5' cryptic splice site selection in genetic disorders. Oral presentation: ABACBS. 2020, November 24-26, virtual.

**Dawes, R.,** Lek, M., Cooper, S.T. Gene discovery informatics toolkit defines candidate genes for unexplained infertility and prenatal or infantile mortality. Poster presentation: Genome Informatics, 2019 November 6-9, Cold Spring Harbor, New York.

**Dawes, R.,** Lek, M., Cooper, S.T. Gene discovery informatics toolkit defines candidate genes for unexplained infertility and prenatal or infantile mortality. Poster presentation and rapid talk: American Society of Human Genetics Annual Meeting, 2019 October 15-19, Houston, Texas.

**Dawes, R.** and Cooper, S.T.C. Predicting the use of cryptic splice sites as a result of genetic variants. Oral presentation: The University of Sydney Children's Hospital Westmead Clinical School 2019 HDR Student Conference, 2019 August 16, Sydney, NSW.

**Dawes, R.,** Lek, M., Cooper, S.T. Gene discovery informatics toolkit defines candidate genes for unexplained infertility and prenatal or infantile mortality. ASMR NSW Annual Scientific Meeting, 2019 May 31, Surry Hills, Sydney.

## Abbreviations

ACMG	American College of Medical Genetics and Genomics
AGEZ	AG Exclusion Zone
AMP	Association of Molecular Pathologists
API	Application Programming Interface
CHX	cyclohexamide
DMSO	dimethyl sulfoxide
DNA	Deoxyribonucleic acid
EBV	Epstein-Barr virus
ENCODE	Encyclopedia of DNA Elements
HAL	Hexamer Additive Linear
HGMD	Human Gene Mutation Database
HSF	Human Splicing Finder
IQR	Interquartile Range
kb	kilobase
LaBranchoR	Long short-term memory network Branchpoint Retriever
LCL	Lymphoblastoid cell line
LOESS	Locally Weighted Scatterplot Smoothing
MES	MaxEntScan
MPRA	Massively Parallel Reporter Assay
MPS	Massively Parallel Sequencing
NCBI	Lymphoblastoid cell line
NMD	Nonsense-Mediated Decay
nt	nucleotide
OMIM	Online Mendelian Inheritance in Man
PGD	Preimplantation Genetic Diagnosis
PPV	Positive-Predictive Value
PSI	Percent Spliced In
PTC	Premature Termination Codon
PWM	Position Weight Matrix
REF	Reference

## Abbreviations

RNA	Ribonucleic acid
RT-PCR	Reverse Transcription Polymerase Chain Reaction
S-CAP	Splicing Clinically Applicable Pathogenicity Prediction
SAI	SpliceAI
SNV	Single Nucleotide Variant
SPANR	Splicing based Analysis of Variants
SPiCE	Splicing Prediction in Consensus Elements
SpliceACORD	Australasian Consortium for RNA Diagnostics Super Quick Information-content Random-forest Learning of Splice
SQUIRLS	Variants
SRA	Sequence Read Archive
SRE	Splicing Regulatory Element
SSF-like	SpliceSiteFinder-like
TPM	Transcripts Per Million
TrAP	Transcript-inferred Pathogenicity
VAR	Variant
VUS	Variant of Uncertain Significance
WGS	Whole Genome Sequencing

## Table of Contents

<i>Statement of Originality</i> -----	2
<i>Authorship Attribution Statement</i> -----	3
<i>Acknowledgements</i> -----	5
<i>Abstract</i> -----	6
<i>Conferences and other proceedings</i> -----	8
<i>Abbreviations</i> -----	10
<i>List of Figures</i> -----	14
<i>Introduction</i> -----	15
1.1    Variation in genes can lead to Mendelian disorders-----	15
1.2    The role of variants affecting RNA splicing in rare disease.-----	17
1.2.1    RNA splicing in protein-coding genes-----	17
1.2.2    Alternative splicing and mis-splicing-----	19
1.2.3    RNA splicing variants in Mendelian disorders -----	22
1.3    Clinical interpretation of splicing variants-----	23
1.3.1    RNA functional studies for variant interpretation-----	25
1.4    In silico tools used for the interpretation of splicing variants -----	28
1.4.1    Motif-based Algorithms -----	28
1.4.2    Contemporary Machine Learning Algorithms-----	32
1.5    Integrating <i>in silico</i> splicing tools into clinical practice -----	36
1.6    Thesis aims and outline -----	38
<i>Empirical prediction of variant-activated cryptic splice donors using population-based RNA-Seq data</i> -----	40
2.1    Overview -----	40
<i>SpliceVault predicts the precise nature of variant-associated mis-splicing.</i> -----	61
3.1    Overview -----	61
<i>Discussion</i> -----	92



## Table of Contents

4.1	Developing a method to predict the mis-splicing outcomes of splicing variants -----	92
4.1.1	Current splicing algorithms are not geared towards aiding pathology interpretation.-----	92
4.1.2	Past splicing behaviour is a potent predictor of future behaviour-----	93
4.1.3	Implications for variant interpretation and RNA functional studies. -----	94
4.2	Future Directions- improving and expanding SpliceVault -----	96
4.3	Conclusions -----	99
<i>References</i> -----		<i>100</i>

## List of Figures

Figure 1-1 Whole Genome Sequencing (WGS) for rare disease diagnosis .....	16
Figure 1-2 RNA splicing in protein coding genes .....	18
Figure 1-3 Alternative Splicing and Mis-splicing .....	22
Figure 1-4 Variants at splicing motifs in Clinvar 2021 .....	23
Figure 1-5 Schematic representation of RT-PCR and RNA-Seq for the detection of two mRNA isoforms.....	26
Figure 1-6 Timeline of Splicing Algorithm development.....	30
Figure 1-7 Citations by year of Splicing Algorithms published prior to 2021.....	37
Figure 3-1 Re-presented Figure S5 from Bournazos et al. 2022 .....	62

## Introduction

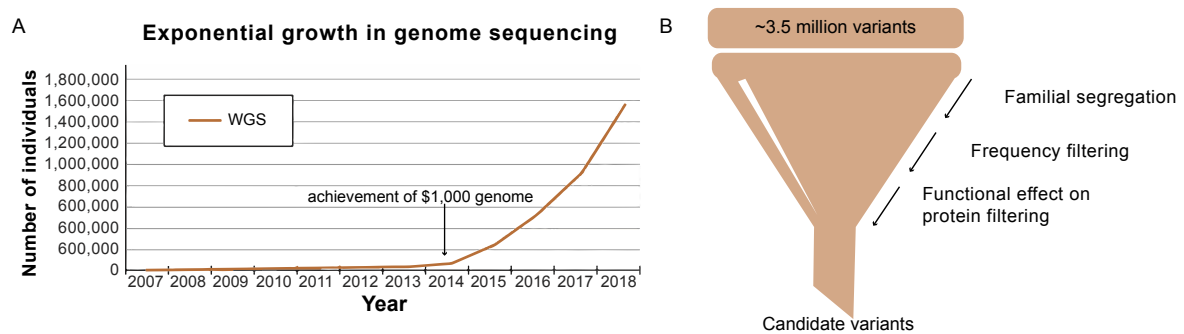
### 1.1 Variation in genes can lead to Mendelian disorders

Mendelian disorders derive their name from their inheritance patterns within families - likened to the inheritance of traits in the pea plants studied by Gregor Mendel<sup>1</sup>. These disorders each have a genetic cause and fall under the umbrella of 'rare diseases', which are defined by their rarity but a large subset of which have a genetic aetiology. As of December 2021, 4,188 genes have been linked to one or more Mendelian disorders<sup>2</sup>. Although individually rare, cumulatively Mendelian disorders affect around 1 in 50 individuals of European descent<sup>3</sup>, and an estimated 300 million people worldwide at any point in time<sup>4</sup>. Genetic rare disorders disproportionately affect the nervous systems of children, are chronic and progressive, and often lead to severe life-long intellectual and physical disabilities<sup>3</sup>. Mendelian disorders are a leading cause of critical illness and mortality in infancy<sup>5</sup>, and in one study Mendelian disorders were found to underlie 71% of paediatric hospital admissions<sup>6</sup>.

Due to the severity and distinctive presentation of many Mendelian disorders, their diagnosis usually begins with the observance of a patient's phenotype. Initial clinical investigations may allow a provisional clinical diagnosis and suggest likely gene candidates, however a precise genetic diagnosis is still crucial to allow accurate prognosis, treatment and family planning<sup>3</sup>. Additionally, diagnosis of Mendelian disease based on clinical features alone is challenging due to phenotypic overlap between the ~10,000 currently known disorders<sup>7,8</sup>.

The diagnosis of genetic disease has been revolutionised by recent technological advances and the drop in cost of Massively Parallel Sequencing (MPS)<sup>9</sup>, with genetic investigation of a Mendelian phenotype now frequently beginning with sequencing of the patient's whole genome ('whole genome sequencing', WGS) (Figure 1.1A)<sup>7,10-12</sup>. WGS sequences the whole genome of a patient and compares it to a reference human genome, to allow identification of all the nucleotides that differ ('variants'). The ~3.5 million variants identified per patient<sup>13</sup>

undergo filtering to prioritise those most likely to cause a pathogenic defect in the encoded gene product (Figure 1.1B).



**Figure 1 Whole Genome Sequencing (WGS) for rare disease diagnosis.** **A)** Exponential growth in individuals undergoing WGS since 2014, when the cost of sequencing dropped to \$1000. adapted from 7. **B)** The ~3.5 million variants<sup>13</sup> identified in an individual who has undergone WGS must be filtered to identify those most likely to cause their phenotype. Familial segregation = the variant identified with the phenotype within the individual's family; frequency filtering = the variant is rare in the healthy population; functional effect on protein filtering = the variant can confidently be predicted to cause Loss-of-Function of the gene or cause disease by some other deleterious mechanism.

If deleterious variants in a high-likelihood candidate gene are found, these can be pursued as the likely causative variant(s). The identification of the causative gene defect can then establish a definitive and accurate genetic diagnosis, which is critical for patient care<sup>14,15</sup>. Genetic diagnosis provides information to clinicians, affected individuals, and their families on prognosis and the best course of treatment<sup>16</sup>. Due to the inherited nature of Mendelian disorders, it also allows family planning, through understanding of the risk of recurrence and enabling disease prevention via prenatal counselling and Preimplantation Genetic Diagnosis (PGD)<sup>17</sup>.

Oftentimes however no causal variant is found, and individuals must embark on a 'Diagnostic Odyssey' to find the genetic explanation for their disorder, spanning on average ~5 years, involving more than 7 physicians and specialists, and causing significant burden and distress to affected patients and their families. In the last 10 years, a bottleneck has formed as variant interpretation lagged behind rapid advances in genomic sequencing technologies<sup>18</sup>, leading to the classification of many patient variants as 'Variants of Uncertain Significance' or VUS. A VUS cannot be used in clinical decision-making and leaves patients and clinical teams with no genetic answer<sup>19</sup>. By 2019, 48% of all variants in ClinVar, the

public archive of human genetic variants, were classified VUS<sup>20,21</sup>. This interpretative bottleneck has critical implications for patient care, with ~50% of patients being left without a genetic diagnosis even after WGS<sup>22–24</sup>.

In cohorts of VUS, ~20-30% of variants have been found to affect the molecular process called splicing<sup>20,25</sup>, which in longform is actually ‘precursor messenger RNA (pre-mRNA) splicing’. Due to challenges interpreting exonic or intronic variants that may or may not disrupt pre-mRNA splicing underpinning as genetic disorder, splicing variants are overrepresented among variants currently classified as VUS as they are frequently overlooked by diagnostic pipelines<sup>26</sup>.

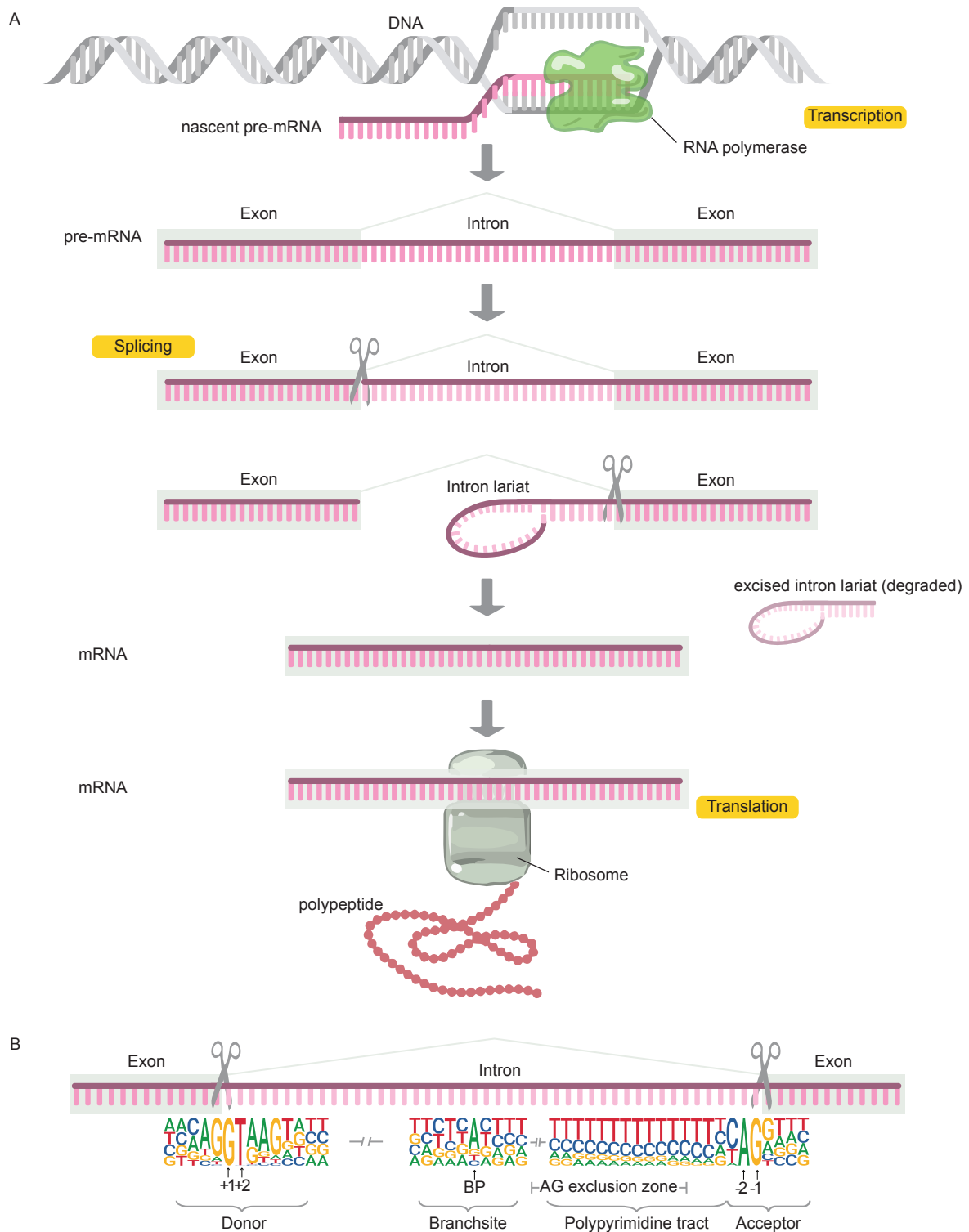
### **1.2 The role of variants affecting RNA splicing in rare disease.**

#### **1.2.1 RNA splicing in protein-coding genes**

Human genes are split into intercalated regions called exons and introns, defined based on their role in the mRNA transcript. After being transcribed from DNA into pre-mRNA (Figure 1.2A, Transcription), introns must be removed from the pre-mRNA transcript in a biochemical reaction called splicing<sup>27</sup> (Figure 1.2A, Splicing). The remaining exons are then ligated together to form the mature mRNA transcript. mRNA is then translated into protein according to the amino acid code, wherein three nucleotide units (called ‘codons’) in the mRNA each encode a specific amino acid, and consecutive amino acids are linked together to create a polypeptide chain that forms a mature protein (Figure 1.2A, Translation).

The splicing reaction is completed by a large, super-complex made up of protein and RNA components called the ‘spliceosome’, which must be directed to the intron borders by signals embedded in the mRNA transcript. These signals are short stretches of sequence called ‘splice sites’ and ensure that the splicing reaction occurs at the precise nucleotide border between exon and intron. Splice sites are often represented as consensus sequences which represent how commonly each nucleotide is seen at each position in the motif (Figure 1.2B).

## Introduction



**Figure 2 RNA splicing in protein coding genes. A)** The central dogma of molecular biology, adapted from Clancy & Brown 2008<sup>28</sup> to include splicing. After being transcribed from DNA into pre-mRNA, introns must be removed from the pre-mRNA transcript in a biochemical reaction called splicing (represented here as a pair of scissors). The remaining exons then form the mature mRNA transcript which can be translated into a polypeptide according to the amino acid code, and then folded and processed to become the final protein. **B)** Consensus splice motifs at the exon-intron junctions.

In molecular biology parlance, the splice site at the beginning of the intron is called the ‘donor’ and the splice site at the end of the intron is called the ‘acceptor’. The donor and acceptor sites across the human genome each have an essential dinucleotide: for the donor, GT (or in rare cases, GC), and for the acceptor AG. Neighbouring these dinucleotides is a more variable but still important stretch of sequence – the splicing ‘motif’. The acceptor site additionally requires a ‘polypyrimidine tract’ (a run of variable length containing an enrichment of Ts and Cs) and a ‘branchsite’ (Figure 1.2B).

In the splicing reaction, the transcript is spliced first at the donor, and the donor then loops over to base pair with the branchsite to form a lariat structure<sup>27</sup> (Figure 1.1A, splicing). The sequence after the branchsite is then scanned in the 3’ direction for the acceptor, to cleave specifically at that location, precisely excise the intron, then ligate the exons together. Due to this ‘AG-scanning’ mechanism to identify the acceptor, there is an ‘AG Exclusion Zone’ (AGEZ) where decoy AGs appear very infrequently between branchsite and real acceptor, as they are highly likely to cause splicing at the incorrect nucleotide<sup>29–31</sup> (Figure 1.2B).

In addition to the donor, acceptor and branchsite motifs, the recognition of the exon-intron junction can either be enhanced or suppressed by further Splicing Regulatory Elements (SREs). These may be exonic or intronic stretches of sequences that are complementary to RNA binding proteins (RBPs) or form an RNA structure<sup>32</sup>. There have been significant efforts made to characterise SREs, either through *in vitro* measurements of RBP binding to random RNA sequences<sup>33–38</sup>, mutational scanning of minigene transcripts for changes affecting splicing<sup>39–49</sup>, or *in silico* analyses of enriched or conserved sequences surrounding splice-sites<sup>50–56</sup>. These studies have identified many degenerate motifs which are plentiful in the human genome, however it remains difficult to interpret the cumulative effect and interactions between the many possible SREs in the vicinity of any one splice-site on *in vivo* splicing decisions.

### 1.2.2 Alternative splicing and mis-splicing

The same stretch of pre-mRNA sequence may be spliced at different splice sites in different transcripts, a process called ‘alternative splicing’<sup>57–60</sup>. The ‘building blocks’ (exons) can be

assembled into a final product (mature mRNA transcript) in different ways. For example, a whole exon could be excluded from a transcript (Figure 1.3A, Exon Skipping), or an intron could be included (Figure 1.3A, Intron Retention). There could also be multiple donor or acceptor sites flanking the same exon which may be used at different proportions or in different tissues (Figure 1.3A, alternative donor / alternative acceptor)<sup>61-63</sup>. These alternatively spliced mRNA transcripts are called 'isoforms' (Figure 1.3A). Alternative splicing is ubiquitous in the human genome<sup>64</sup>; 86% of human genes have a minor isoform with a frequency of 15% or higher, relative to the major isoform<sup>61</sup>.

Correct splicing requires the coordination of large multi-protein/RNA spliceosomal complexes that identify and bind to the short (and degenerate) consensus splicing motifs amidst many kilobases of pre-mRNA, with high specificity. Alternative splicing additionally requires tight regulation of the relative abundances of different transcripts in different cellular contexts. Therefore, it's no surprise that the spliceosome commonly makes mistakes - referred to as 'mis-splicing'<sup>65</sup>.

The consequences of mis-splicing for the resultant protein are often dire. Due to the protein-code being written in 3 nt subunits called 'codons', the aberrant inclusion or exclusion of even 1 or 2 nt will disrupt the code for the whole remainder of the transcript, pushing it 'out of frame'. A common result of this is the introduction of a Premature Termination Codon (PTC) into the transcript, where one of the three Termination Codons (TAA, TAG, TGA) which signal that the end of the mRNA transcript has been reached is detected in an aberrant position. PTCs are detected by a cellular surveillance mechanism which degrades and eliminates these transcripts ('nonsense-mediated decay' or NMD) (Figure 1.3B, absent protein)<sup>66</sup>.

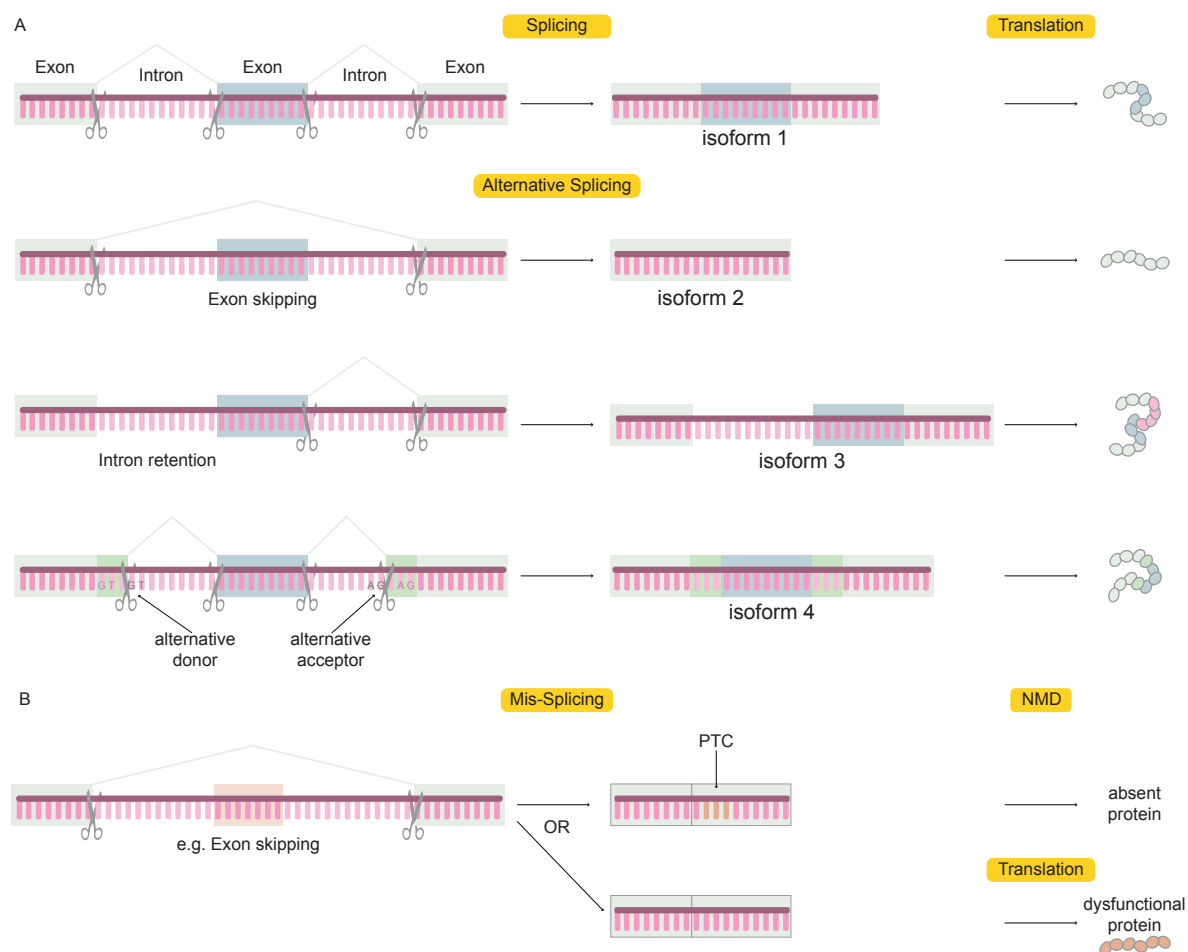
The most well-defined model of the molecular mechanism of NMD posits that as ribosomes scan the spliced mRNA, if a stop codon is detected more than ~50 bases upstream of an exon-exon junction complex (EJC, multi-protein assemblies which bind near exon-exon junctions after the splicing reaction), a protein complex called SURF is recruited<sup>66,67</sup>. This is detected as abnormal, as usually the stop codon appears downstream of any transcript bound EJCs. The binding of the SURF complex then triggers a series of steps which expose the mRNA molecule to degradation by exonucleases<sup>66,67</sup>. The efficiency of NMD, i.e. what proportion of aberrant transcripts are degraded, has been measured at variable rates in



different tissues and cellular conditions as well as according to the position of the PTC<sup>68-70</sup>, The variable efficiency of NMD makes it difficult to determine the true abundance of aberrant transcripts with PTCs produced, without inhibiting NMD.

In-frame events or encoded PTCs which do not trigger NMD may also nevertheless produce pathogenic transcripts. Even if the included or excluded length of DNA is divisible by 3 and the reading frame is retained, the missing amino acids may be crucial for protein function (Figure 1.3B, dysfunctional protein), or the added amino acids may sabotage the protein or also encode a PTC<sup>71,72</sup>.

The same mechanisms that produce alternative isoforms (i.e., Exon Skipping, Intron Retention, Alternative Donors/Acceptors) can occur aberrantly, resulting in a transcript degraded by NMD, or a dysfunctional protein missing or disrupting a functional domain (Figure 1.3B). Reference databases cataloguing detected RNA transcripts have been developed (for example the Ensembl<sup>73</sup> and Refseq<sup>74</sup> databases), however it remains difficult to ascertain if a given transcript, even an annotated one, constitutes a biologically functional alternate isoform or one aberrantly produced by a splicing error. Gene- and disease-specific expertise is often required to decisively call an isoform ‘alternate splicing’ or ‘mis-splicing’.



**Figure 3 Alternative Splicing and Mis-splicing.** A) the same pre-mRNA transcript can be alternatively spliced to produce several different mRNA transcripts and resultant polypeptides via the mechanisms of exon skipping, intron retention, and alternative splice-sites (donors and acceptors). B) Exon skipping (example shown), intron retention and alternative splice-site events (not shown) constitute mis-splicing if they produce transcripts which encode a Premature Termination Codon (PTC) and undergo nonsense-mediated decay (NMD) or else are translated into dysfunctional proteins.

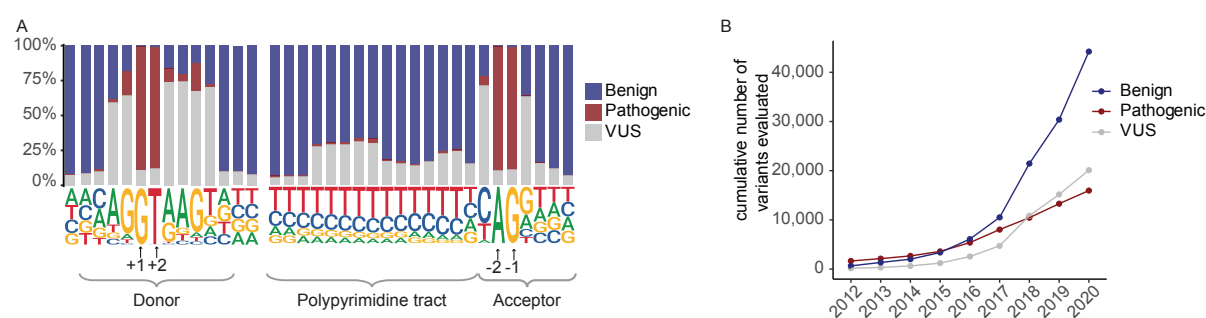
### 1.2.3 RNA splicing variants in Mendelian disorders

In Mendelian disease, pathogenic mis-splicing can be caused by genomic variants. A ‘splicing variant’ usually refers to a genomic variant affecting a splice site – either the donor, acceptor, polypyrimidine tract or branchsite<sup>26,75</sup>. When splice sites are modified by a genomic variant it can compromise their recognition by the spliceosome and subsequent intron excision, causing mis-splicing and an aberrant mRNA transcript that will commonly encode a PTC. Additionally, a genetic variant may introduce a new splice motif in a location where there wasn’t one before, or modify the sequences surrounding a naturally occurring decoy or cryptic ‘GT/C’ or ‘AG’ in the genome, so that the spliceosome mistakenly splices at these cryptic splice-sites instead of the bonafide splice-sites at the exon-intron junction<sup>76</sup>. In some

cases, genetic variants outside of the more well-defined splicing motifs and instead occurring in SREs can nevertheless disrupt splicing, however these remain difficult to study systematically and remain largely limited to individual reports of disease-associated SREs in individual exons/introns<sup>77–79</sup>.

### 1.3 Clinical interpretation of splicing variants

Many splicing variants remain classified as VUS due to the difficulty of interpreting them clinically<sup>20,80</sup>. Whilst the triplet amino acid code allows us to consider the biochemical effect that an exonic variant has on the encoded gene product (protein). In contrast, the impact of coding or noncoding variants affecting splicing motifs are much more difficult to interpret<sup>75,81</sup>, due to absence of a comprehensive equivalent ‘splicing code’ reflecting our incomplete understanding of splicing regulation. Outside the essential dinucleotides, a large proportion of splicing variants remain classified VUS in Clinvar (Figure 1.4A), and the number of VUS added has grown steadily over the past several years (Figure 1.4B).



**Figure 4 Variants at splicing motifs in Clinvar 2021.** A) Synonymous or intronic variants at the exon-intron boundary. Donor +1/+2 and Acceptor -1/-2 variants are overwhelmingly pathogenic, with a large proportion of VUS outside this ‘essential splice-site’ region. B) The number of VUS among the same set of variants has grown steadily over the last decade, along with Benign and Pathogenic classified variants.

In 2015, The American College of Medical Genetics and Genomics (ACMG) and Association of Molecular Pathologists (AMP), released a set of guidelines for the clinical interpretation of short sequence variants in Mendelian disease which have since become the internationally accepted standard for variant interpretation<sup>82</sup>. According to ACMG/AMP guidelines, clinical geneticists can apply the ‘pathogenic very strong’ criterion (PVS1) for predicted loss-of-function to any essential dinucleotide variant in a gene where loss-of-function is a known mechanism of disease: these nucleotides are essential to the splicing

reaction, and so virtually guaranteed to cause mis-splicing<sup>82,83</sup>. This criterion partly explains why most essential dinucleotide variants in ClinVar are classified pathogenic (Figure 1.4A).

However, even this criterion can only be applied after theoretical consideration of likely mis-splicing outcomes induced by the variant. Different mis-splicing outcomes will result in different aberrant transcripts, which may be degraded if they introduce a PTC. The outcome of intron retention, exon skipping, and cryptic splicing events must therefore be considered before even an essential dinucleotide variant can be confidently predicted to cause a loss-of-function<sup>83</sup>.

Beyond the application of the PVS1 criterion, the specific mis-splicing events induced by a splicing variant are crucial to pathology interpretation. For example, splicing mutations in the *DMD* gene can have two main outcomes: If the variant leads to mis-splicing events that disrupt the reading frame, no functional protein will be produced, resulting in Duchenne Muscular Dystrophy that is associated with loss of ambulation and use of a wheelchair around 11-13 years of age and a median life expectancy of 28 years<sup>84,85</sup>. If a variant in *DMD* leads to mis-splicing events that retain the reading frame, truncated or otherwise slightly modified dystrophin protein will be produced and lead to Becker muscular dystrophy, a milder form of disease that can have a later-onset and slower progression<sup>86-88</sup>.

For most variants affecting splicing motifs outside the essential dinucleotide, it is currently impossible to clinically interpret whether spliceosomal recognition will be impaired or ablated, and if so, exactly how splicing will be affected. Variants which modify splicing motifs without abolishing the splice site may cause partial mis-splicing, meaning that while aberrant transcripts are produced, some proportion of transcripts remains correctly spliced<sup>26,89,90</sup>. The amount of correctly spliced transcript that is required to avoid a pathogenic outcome varies from gene to gene, and even from exon to exon, and its determinants are largely unknown. In addition, common alternative splicing of genes in different tissues lends extra complexity: it is important to consider impact of variant-associated mis-splicing of a given gene in the manifesting tissue(s) relevant to that disorder, to confidently call a variant benign or pathogenic.

In ClinVar, 83.5% of splicing variants that have been classified pathogenic are essential dinucleotide variants<sup>91</sup> (Figure 1.4A). In contrast, a large scale unbiased analysis of splicing

variants in a cohort undergoing genetic sequencing found that up to 39% of splice-disrupting mutations were outside the essential dinucleotide, suggesting that ~35-40% of pathogenic splicing variants which don't modify the essential dinucleotide may be missed in routine clinical practice<sup>91</sup>. Many of these 'extended splice-site' variants likely remain classified as VUS, a class which has been steadily growing since the introduction of MPS into clinical practice (Figure 1.4B). These issues mean that functional studies of the effects of genetic variants on splicing are critical to assess variant pathogenicity<sup>92</sup>.

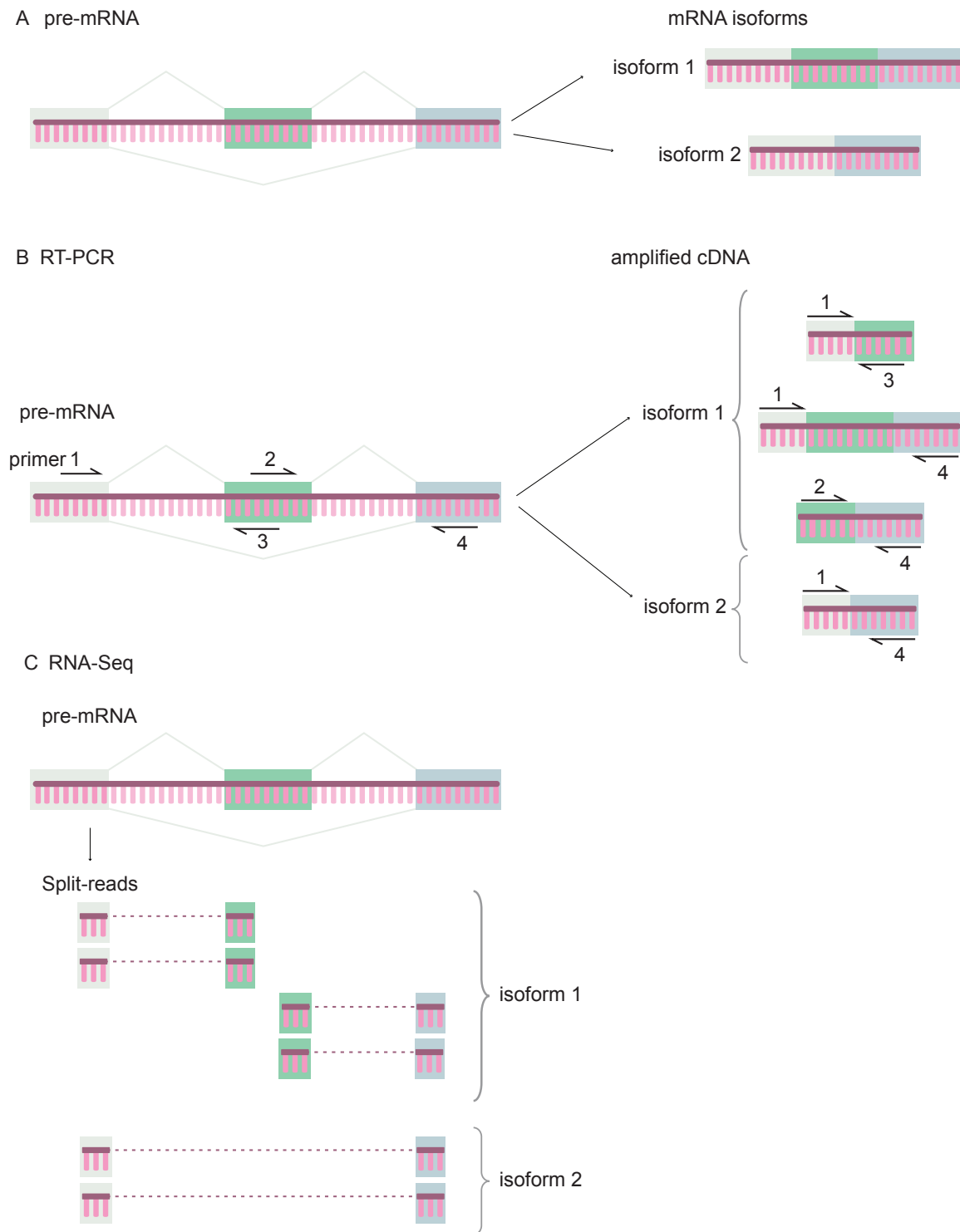
### 1.3.1 RNA functional studies for variant interpretation

The traditional workhorse of RNA functional studies is Reverse-Transcription PCR (RT-PCR), where mRNA derived from a patient sample is reverse-transcribed and amplified using 'primers' - short oligonucleotides (~18-24 nt) complementary to patient cDNA to selectively amplify sections using PCR cycling (Figure 1.5B). Primers can be designed to specifically amplify canonical splicing or mis-splicing to provide diagnostically useful information; for example, the end and beginning of subsequent exons which would indicate correct splicing. RT-PCR could demonstrate a splicing defect if this stretch of sequence is absent in the patient and present in controls. Researchers also design primers to amplify mis-splicing events likely to result from the variant, such as exon skipping, intron retention and possible cryptic sites, and see if they're amplified in the patient compared to controls.

RT-PCR is highly sensitive as PCR amplifies the specific targeted sequence by many orders of magnitude over 30 or 40 PCR cycles. This sensitivity can be very important, as many mis-splicing events are out-of-frame, meaning NMD will degrade the aberrant transcripts in patient RNA, making false negatives feasible.

An alternate approach is RNA-sequencing or 'RNA-Seq'. The technological advances that have made large-scale DNA sequencing feasible have equally applied to the sequencing of RNA. Modern sequencing methods involve fragmenting DNA/RNA into many short fragments which can be rapidly sequenced in parallel, and then reassembled by informatic alignment back to a reference sequence (much like assembling a jigsaw puzzle using the picture on the box). Researchers then inspect how the reads are mapped in the region

## Introduction



**Figure 5 Schematic representation of RT-PCR and RNA-Seq for the detection of two mRNA isoforms.** A) The same pre-mRNA transcript may yield two alternative isoforms, in this example through skipping of the central exon. B) RT-PCR requires the design of primers amplifying portions of the cDNA sequence. In this case three unique primer pairs amplify isoform 1, while only one primer pair amplifies isoform 2. C) RNA-Seq detection of splicing isoforms requires the alignment of short reads spanning exon-intron junctions ('split-reads').

surrounding the variant using ‘sashimi plots’ in the Integrative Genomics Viewer<sup>93</sup>, and compare with controls which don’t have the variant (Figure 1.5C).

While DNA has one reference sequence to align the sequence reads to, RNA sequencing requires alignment back to a complicated ‘transcriptome’ of alternative and tissue-specific splicing. This vastly different property of RNA means that functional studies must be carefully planned and interpreted, and standardised processes are only just beginning to be established<sup>89,94</sup>.

RNA-Seq is designed to capture the entire transcriptome, meaning primers targeting specific splicing events are not required, making it a more agnostic, hypothesis-free and generalisable approach. However, it requires adequate expression of a gene in the specimen the RNA is extracted from, to provide adequate sequencing read-depth of the region of interest<sup>95,96</sup>. As many genetic disorders are recessive, and correctly spliced transcripts are arising from the allele *in trans*, high sequencing depth is often needed to detect lowly expressed genes and/or to ensure mis-spliced transcripts being effectively targeted by NMD are detected. Standard guidelines for sequencing depth have not been set and in practice cost is usually the limiting factor<sup>97</sup>.

Another important consideration is the length of RNA sequenced in each read (‘read length’), which can affect the accuracy of alignment back to the reference transcriptome. The longer the reads are (or: the more information each puzzle piece has) the easier it is to trace them back to their originating location in the genome, especially for ‘split reads’ which cover the end and beginning of two exons spliced together (Figure 1.5C)<sup>98,99</sup>. A common read length is 150 nucleotides (nt). The ‘overhang’, or proportion of the read aligned to one or the other exon may be only a few nucleotides, and with the beginning and ends of exons being part of conserved splice motifs, this length may not be enough to unambiguously map a given short read to a specific exon-intron junction. Additionally, the popular alignment tool STAR allows a small number mismatches between the read and the reference genome, and also prioritises the alignment of annotated events (requires 1 nt overhang) over unannotated events (requires 5 nt overhang), introducing biases which may affect the veracity of aligned reads<sup>100,101</sup>. The shorter the sequencing read, the more likely it is to be mis-mapped to the wrong spot or discarded altogether.

While RNA-Seq has incredible potential as a diagnostic technology, the diagnostic ambiguity conferred by informatic shortcomings correctly mapping short-sequencing reads to a gene, no way to be sure in recessive disorders whether canonically spliced transcripts detected are arising from the splicing variant allele or the allele *in trans*, the comparatively high cost and lack of uniform quality standards mean its currently difficult to bring into standard practice.

RT-PCR is highly sensitive and is supported by an existing quality framework established for Sanger sequencing of PCR amplicons for segregation studies, enabling its use as an accredited RNA testing pathway in a diagnostic setting. For this reason, it's the current gold standard for RNA studies, does not require bioinformatic expertise in read alignment, quantification and analysis necessitated by RNA-Seq studies<sup>102,103</sup>. RT-PCR has its own technical caveats that can have great diagnostic impact: you only detect what your primers are capable of amplifying under the PCR cycling conditions employed. Therefore, the bespoke nature of RT-PCR design for each case, and the necessity of designing primers for all theorised mis-splicing events, makes RT-PCR a slow and low throughput functional RNA testing pipelines and consequently incompatible with routine diagnostic medicine.

Samples are therefore rarely sent for RNA diagnostic testing without well-founded suspicion of a defect in RNA splicing or transcription. *In silico* tools that predict the impact of variants on splicing ('splicing algorithms') have thus become instrumental in the process of narrowing down possible variants to make the most of expensive, time-consuming, and technically challenging RNA studies<sup>76</sup>. A plethora of splicing algorithms have been developed over the last few decades and have greatly aided in the prioritisation of splicing variants<sup>92,104</sup>. However, predictions of different algorithms are often discordant, greatly hampering their use in a clinical context. Optimised use of *in silico* splicing prediction tools requires a solid understanding of how each algorithm works and their strengths and weaknesses in different genetic contexts.

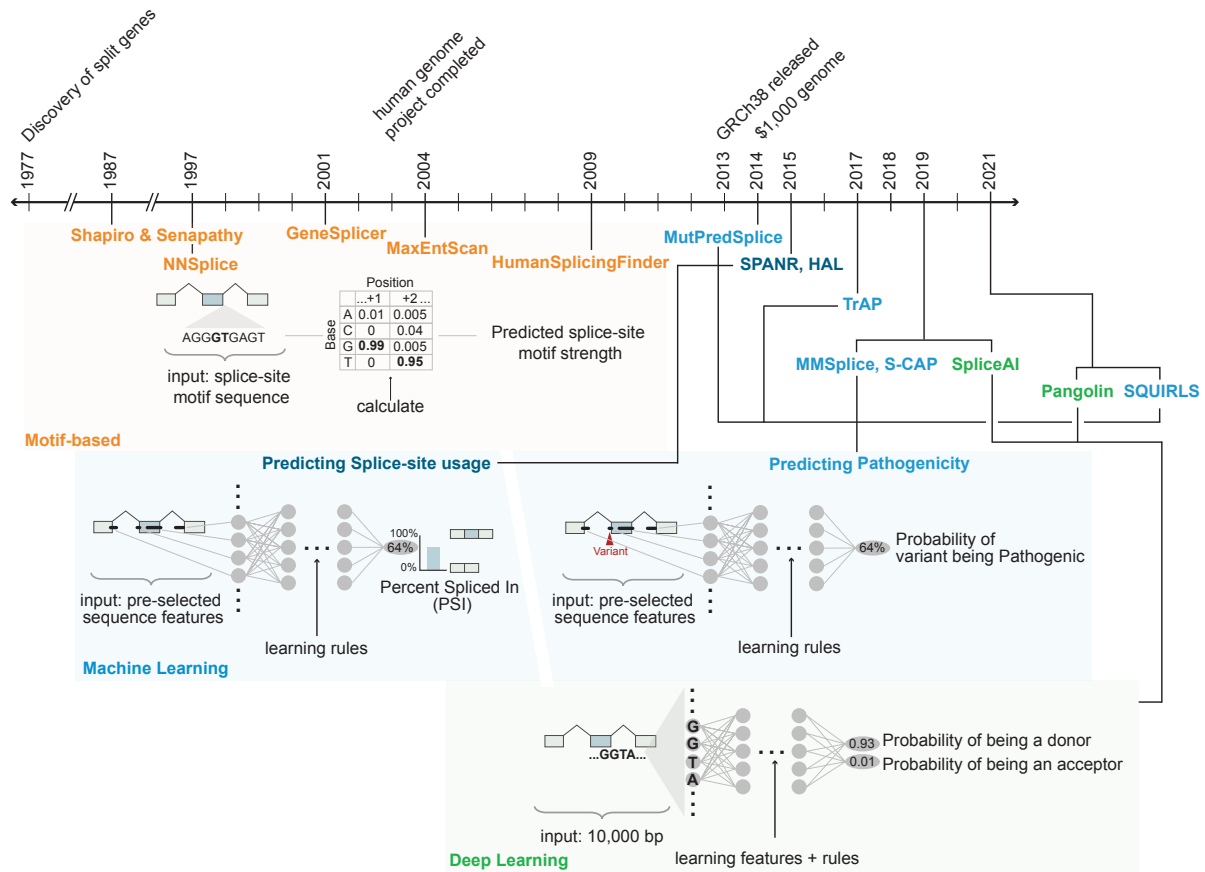


## 1.4 In silico tools used for the interpretation of splicing variants

### 1.4.1 Motif-based Algorithms

Before the completion of the human genome project in the early 2000s, the main purpose of *in silico* splicing prediction tools was as ‘gene finders’, to identify genuine exon-intron boundaries in the genome and annotate a complete list of transcribed human genes. To aid in annotation efforts, researchers developed statistical models to assist computational recognition of the degenerate, short donor and acceptor splicing motifs. These statistical models were based on known exon-intron boundaries (from humans and model organisms) and used to search for splicing motifs in the genome that flank exons, enabling mapping of genes and gene structures.

One of the earliest developed gene finders is the position weight matrix (PWM) model developed by Shapiro and Senapathy<sup>105</sup> (Figure 1.6, 1987). To score a sequence of nucleotides, the PWM simply multiplies the probability (‘weight’) of finding each nucleotide at its position in the sequence if that sequence were a donor or acceptor<sup>105</sup>. In the PWM model all positions of the motif are considered independent, meaning that for example finding a G at position -1 doesn’t affect the calculated probability of finding a G at position +5. While this makes the model easy to calculate and understand, it ignores important dependencies between nucleotides.



**Figure 6 Timeline of Splicing Algorithm development.** 10 years after the initial discovery that eukaryotic genes were split into exons and introns (1977), the Shapiro and Senapathy PWM splicing algorithm was published. **Orange:** Subsequent motif-based algorithms (1.4.1). These algorithms calculate the predicted strength of a splice-site motif, usually based on some statistical model of the ideal splice-site. **Blue:** Contemporary machine learning algorithms (1.4.2.1-2). *Left:* Predicting splice-site usage (1.4.2.1). These algorithms use sequence features to predict Percent Spliced In (PSI) *Right:* Predicting pathogenicity (1.4.2.2). These algorithms use sequence features to predict the probability of a variant being pathogenic (or benign). **Green:** Deep learning algorithms (1.4.2.3). These algorithms use a large sequence window to calculate the probability that that sequence is a donor.

In 1997 NNSplice was developed, a neural network trained to differentiate real from decoy splice sites, using short stretches of sequence<sup>106</sup> (Figure 1.6). Neural networks automate the process of finding patterns in training data, making it easier to analyse large datasets and identify real from decoy splicing motifs. In 2001 GeneSplicer was developed, similarly trained to differentiate between real and decoy splice sites using a statistical Markov model, which takes dinucleotide as well as nucleotide preferences into account<sup>107</sup>. In addition to short donor and acceptor motifs GeneSplicer created models of coding and noncoding sequences using 80 bp on either side of real splice sites, which helped it identify sequence contexts that are more typical in exons versus introns, to improve accuracy<sup>107</sup>. In 2004 Yeo and Burge created MaxEntScan, which uses a maximum entropy model to differentiate real

from decoy splice sites<sup>108</sup>. This model considers not only dinucleotide preference but other higher-order dependencies between non-adjacent positions in the splice site motif.

Due to their original purpose as gene finders, all four models are based on incomplete annotations of human transcripts. They are trained on between ~3,000 to ~17,000 annotated splice sites, excluding the majority of the almost half a million splice sites now annotated in the GRCh38 genome assembly (Figure 1.6, 2013). They also all exclude the small, but important, subset of real splice donors that use a ‘GC’ essential splice site rather than the far more common ‘GT’, and a rare class of essential splice motifs that recruit a specialised spliceosome called the minor spliceosome (U12) that splices only around 800 human introns<sup>109</sup>. Despite these caveats, these historical splicing algorithms are still commonly appropriated for use in clinical genetics to assess impact to splice sites by genetic mutations<sup>89</sup>. This is in part due to their integration into the commonly used program Alamut Biosoftware ® Rouen, France, which integrates predictions from multiple tools in a user-friendly data visualisation tool.

The program also integrates an additional two algorithms: 1) SpliceSiteFinder-like (SSF-like) which is a now deprecated and unavailable algorithm based on Shapiro and Senapathy’s PWMs<sup>105</sup>. 2) Human Splicing Finder (HSF)<sup>110</sup>, created in 2009 based on Shapiro and Senapathy’s PWMs includes PWMs based on the Branchsite and additional Splicing Regulatory Elements, and an additional constraint that an AG cannot appear in the polypyrimidine tract of an acceptor. Furthermore, There have been recent efforts to maximise the accuracy of these traditional motif-based algorithms by combining them: Splicing Prediction in Consensus Elements (SPiCE) combines predictions of SpliceSiteFinder-like and MaxEntScan<sup>108</sup> to train an algorithm on 142 BRCA1/BRCA2 variants they’d assayed *in vitro*<sup>111</sup>.

All these algorithms use short stretches of sequence to predict the strength of splice sites – between 7 and 16 nt for the donor and between 15 and 41 nt for the acceptor. While this encapsulates the regions with a high degree of sequence conservation and the motifs known to be necessary to complete splicing (Figure 1.2B), it does not sufficiently capture the constellation of sequence factors that regulate a splicing decision. This can be seen by the fact that sequences which either partially or perfectly match splicing motifs, yet which are not used commonly as splice sites (‘decoy splice sites’) are common throughout the human

genome - 98% of the true donor sequences used by MaxEntScan are also contained in the false donor (decoy) set<sup>112</sup>. In addition, it's been known for decades that usage of a splice site by the spliceosome does not necessarily correlate with how well its sequence agrees with the consensus<sup>113</sup>. These conundrums are likely due to ill-understood long-range sequence regulatory elements (SREs) that influence splicing decisions. For example, SREs have been identified throughout the exon<sup>40,42,44</sup>, surrounding the exon-intron border<sup>43</sup> as well as deep into the intron<sup>114</sup>.

In the last decade, several splicing algorithms have been developed using machine learning methods able to integrate these additional sequence factors outside short splice-motifs, with a variety of approaches and use-cases (Figure 1.6, blue).

### **1.4.2 Contemporary Machine Learning Algorithms**

In machine learning, instead of explicitly programming or defining the rules of a statistical model, researchers provide a 'machine' with many examples of what they're trying to predict, as well as information or 'features' they think can discriminate between these two groups (Figure 1.6, blue). The machine then tries to iteratively learn rules to allow this discrimination and minimise mis-classification. Machine learning algorithms can learn and incorporate rules based from numerous discriminating features, making them a promising tool for representing splice site strength more comprehensively than statistical models of short splice site motifs.

There have been three main approaches to training machine learning algorithms for splicing prediction.

#### *1.4.2.1 Predicting splice site usage*

The concordance of a splice site motif with a consensus sequence strength is known to incompletely correlate with experimentally determined splice site usage, so one group of algorithms instead uses experimental measures of splice site usage to train algorithms to provide a more functional prediction of splice site strength. These algorithms are trained to predict a measure of splice site usage called PSI (Percent Spliced In), using pre-selected

sequence features. PSI is the fraction of transcripts where the splice site is used, and is measured using RNA-seq splice-junction data (Figure 1.6, ‘Predicting Splice Site usage’). SPANR was trained using RNA-seq from 16 human tissues in the Illumina human body Map 2.0 project<sup>115</sup> (Figure 1.6, 2015) and Hexamer Additive Linear (HAL) was trained using data from a Massively Parallel Reporter Assay (MPRA) wherein they measured PSI after introducing millions of random sequences into an exon-intron junction in a plasmid<sup>43</sup> (Figure 1.6, 2015).

As the name implies, Modular Modelling of Splicing (MMSplice) is made up of several modules each trained to predict some dimension of splicing<sup>116</sup>. One module predicts PSI, and was trained using the same MPRA assay data generated by Rosenberg et al. (2015)<sup>43</sup>, and another predicts the impact of variants on exon skipping trained using a dataset called Vex-seq, where relative splicing efficiencies were measured after the introduction of 2,059 variants into reporters<sup>117</sup>. The MMSplice authors have since expanded their algorithm to predict PSI of exons in a tissue-specific manner across 56 tissues (MTSplice)<sup>118</sup>

### 1.4.2.2 *Predicting splicing variant pathogenicity*

Given that the most common usage of splicing algorithms is to predict the outcome of genetic variants in a clinical setting, several algorithms are trained to directly predict the pathogenicity of splicing variants using clinical variant information (Figure 1.6, ‘Predicting Pathogenicity’).

Most algorithms of this class use a largely overlapping set of features: strength of donor and acceptor motifs, SRE motifs and conservation, intron/exon length and the distance of the variant to an annotated splice site. These features are used in MutPred Splice (Figure 1.6, 2014)<sup>119</sup>, Transcript-inferred Pathogenicity (TrAP) (Figure 1.6, 2017)<sup>120</sup>, Splicing Clinically Applicable Pathogenicity Prediction (S-CAP) (Figure 1.6, 2019)<sup>121</sup> and Super Quick Information-content Random-forest Learning of Splice Variants (SQUIRLS) (Figure 1.6, 2021)<sup>122</sup>. Most algorithms use one or more unique features in addition to these, such as:

- > TrAP, which uses the number of transcripts which overlap the variant<sup>120</sup>.

- > S-CAP, which uses SPANR PSI predictions, measures of gene and regional constraint and an additional measure of ‘splice site constraint’ made by measuring the number of rare vs common variants in the splicing motif region observed in population databases<sup>121</sup>.
- > SQUIRLS, which notes whether an AG is introduced in the AGEZ (defined as -50 to +3 relative to the acceptor).

In addition, MMSplice contains a module trained using the predictions from its other modules (donor and acceptor strength and delta-PSI) as features to predict pathogenicity<sup>116</sup> (Figure 1.6, 2019).

Effective machine learning requires a large amount of training data, so this class of algorithms require a large and comprehensive dataset of both pathogenic and benign splicing variants.

The training set of pathogenic splicing variants for most algorithms is derived from public databases of clinically classified variants; ClinVar<sup>21</sup> (MMSplice<sup>116</sup>), The Human Gene Mutation Database (HGMD)<sup>123</sup> (MutPred Splice<sup>119</sup>), or a combination of the two (S-CAP<sup>121</sup>). TrAP is trained on pathogenic variants curated from literature alone<sup>120</sup>. Due to the extreme bias in classification of essential dinucleotide variants among pathogenic splicing variants in clinical databases (Figure 1.4A), many algorithms are severely underpowered in their training data for pathogenic non-essential splice site variants. While the authors of SQUIRLS tried to counter this bias by filtering out essential dinucleotide variants<sup>122</sup>, the ability to reliably assess extended splice site variants continues to be a major issue in clinical practice.

Forming a large, high-quality training set of benign splicing variants is also difficult – as costly RNA functional studies are performed rarely for variants deemed as unlikely to affect splicing. Nevertheless, diverse approaches have been taken to create a training set of benign splicing variants. While MMSplice and SQUIRLS used ClinVar variants classified as benign and within 100 nt of splice sites<sup>116,122</sup>, MutPred Splice uses HGMD variants classified as pathogenic that have published functional data showing no detectable disruption in splicing, as well as high frequency SNVs from the 1000 Genomes Project<sup>119</sup>, deemed as too common to be likely to affect splicing (as mis-splicing induces a frameshift in two-thirds of instances).

Similarly, S-CAP uses a set of synonymous SNVs observed in at least 1% of the population<sup>121</sup> and TrAP uses a set of *de novo* synonymous variants observed anywhere within a coding transcript in healthy individuals<sup>120</sup>.

While high frequency in the general population is suggestive that a variant is not pathogenic, it is no guarantee. The authors of S-CAP show that 18% of presumed benign variants in their training set were predicted to affect splicing by MaxEntScan or LaBranchoR (a branchsite prediction algorithm)<sup>121</sup>, suggesting these training sets of ‘benign variants’ are highly compromised.

Importantly, confirmed variant-associated mis-splicing can be functionally benign. While a synonymous or intronic variant proximal to a splice site classified pathogenic can be inferred as causing pathogenic mis-splicing, a synonymous or intronic variant classified benign can either be associated with no change in splicing, or the variant can cause some degree of mis-splicing that nevertheless has benign clinical consequences for the encoded protein.

### 1.4.2.3 Predicting splice site recognition with deep learning

To avoid training data compromised by mis-annotation and ascertainment bias, newer algorithms have been trained to predict the outcome of splicing variants directly from the DNA sequence, using deep learning. Deep learning is an extension of the idea of machine learning. Instead of the researcher deciding on input features, the algorithm instead breaks up the input itself and finds features that can be used in prediction (Figure 1.6, green). In this way the algorithm can perhaps learn features directly from the DNA sequence that are unaccounted for by current research data.

Deep-learning algorithm SpliceAI (Figure 1.6, 2019) is transforming splicing variant prediction<sup>124</sup>. Splice AI is trained only on the raw sequence of 20,287 human pre-mRNA transcripts. It assesses up to 10,000 nt of sequence surrounding each splice site, with unprecedented accuracy in correctly identifying real, annotated splice sites in the human genome. Splice AI shows a 95% top-k accuracy - which is the proportion of correct predictions at the threshold where the predicted number of splice sites equals the actual number of splice sites.

The top-k accuracy increased from 57% when using 80 nt of sequence context flanking each splice-site, to 95% with 10,000 nt, lending further credence to the importance of long-range sequence determinants for spliceosomal selection of splice-site motifs. The authors showed SpliceAI had independently learnt to recognise splicing regulatory element motifs directly from the DNA sequence, such as the SR-protein motif and branchsite sequence motifs<sup>124</sup>.

As it is based on transcript annotations rather than transcript abundance, SpliceAI primarily predicts splice site recognition rather than a quantitative prediction of splice site usage and is not tissue-specific. Pangolin<sup>125</sup> uses a deep learning model to quantitatively predict splicing across four tissues (heart, liver, brain and testis) (Figure 1.6, 2021). The authors also trained Pangolin using 10,000 nt of sequence around splice sites however included sequences from rhesus, mouse and rat genomes in addition to human genomic sequence<sup>125</sup>. Pangolin marginally outcompeted SpliceAI on top-k accuracy<sup>125</sup>.

### **1.5 Integrating *in silico* splicing tools into clinical practice**

Contemporary machine learning splicing algorithms hold great promise as tools to prioritise clinical variants likely to affect RNA splicing. *In silico* algorithms are one of the eight evidence sources recommended for variant interpretation by the ACMG/AMP, and can currently be used as supporting evidence to classify a splicing variant as benign (criteria BP4) or pathogenic (criteria PP3)<sup>82</sup>. Importantly, ACMG/AMP criteria carry the condition that all *in silico* predictions must be concordant for computational evidence to be used at the level of supporting evidence<sup>82</sup>. For variants outside essential dinucleotides, historical algorithms implemented into Alamut® (which is commonly used by diagnostic laboratories) rarely produce concordant predictions<sup>89</sup>, however they remain some of the most cited splicing algorithms each year (Figure 1.7, orange).

With the explosion of machine learning and deep learning methods developed since the ACMG/AMP guidelines were published in 2015<sup>82</sup>, all trained using different source data and designed to predict subtly but importantly different metrics, this ACMG/AMP stipulation has become increasingly difficult to apply meaningfully and subsequently decreasingly relevant.



A more robust approach is to apply in silico tools according to independently benchmarked thresholds set using large sets of clinical variants with known, experimentally studied outcome at the RNA level<sup>126</sup>.

Clinical uptake of contemporary machine learning methods has thus far been limited (Figure 1.7, blue), due primarily to absence of clinical validation for pathology use. Benchmarking studies have begun being conducted on the newer machine learning tools<sup>92,127</sup>, and many expert ClinGen panels (e.g. Clinical Genome Resource, ClinGen<sup>128</sup>) are now shifting to the use of more contemporary algorithms such as MMSplice and SpliceAI based on published evidence of their improved accuracy over historical algorithms<sup>92</sup>. In particular, SpliceAI is having a significant impact and is already heavily cited (Figure 1.7, green).

Additionally, a recent preprint evaluated missense pathogenicity predictors within a statistical framework that aligns thresholds specifically with ACMG levels of evidential strength (supporting, moderate, strong and very strong)<sup>129</sup>. A similar approach which assesses in silico splicing prediction tools considering ACMG levels of evidence would likely be beneficial.

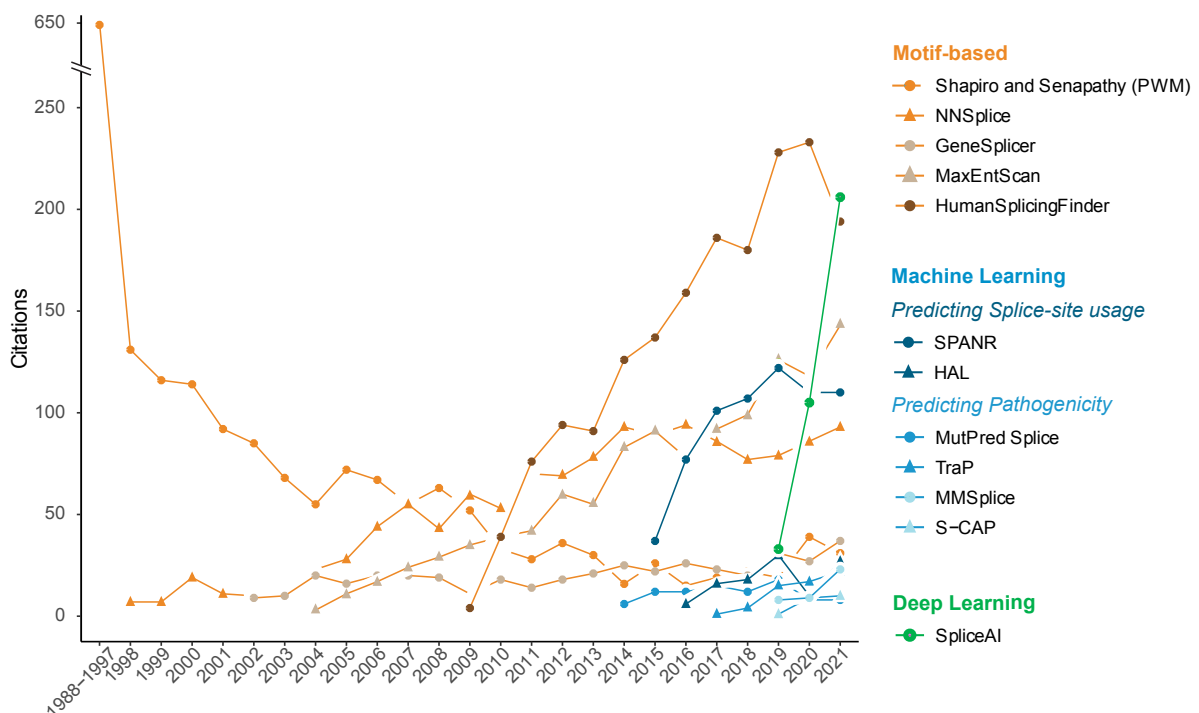


Figure 7 Citations by year of Splicing Algorithms published prior to 2021.

Despite the growing sophistication of splicing algorithms over the last few decades their implementation into diagnostic practice remains fraught with issues and filled with caveats. Splicing variants continue to be classified as VUS in ClinVar at a very high rate (Figure 1.4B). With the advent of targeted splice-modifying therapeutics that have the capacity to correct splice defects in patient cells<sup>130-132</sup>, genetic and transcriptomic diagnosis of patients with Mendelian disease is becoming even more crucial and time-sensitive. The diagnostic potential of under ascertained splicing variants must be met to benefit patients and families and integrate splicing diagnostics and therapeutics into clinical care<sup>97</sup>.

### 1.6 Thesis aims and outline

In the last decade, the clinical uptake of MPS for rare disease diagnosis has generated massive levels of sequencing data, motivating the parallel development of software tools to assist pathology interpretation of vast numbers of genetic variants. Simultaneously, awareness of RNA splicing variants as a major causal basis of disease for patients undiagnosed by DNA testing has been increasing, calling for consensus among the international clinical genetic community on a quality standard framework for RNA diagnostic testing. Rapid development of new splicing algorithms, each using different methods, and often still offering discordant predictions for many variants, presents the incipient challenge of how best to integrate contemporary machine-learning and deep-learning splicing prediction tools into standard diagnostic care.

The aim of my PhD thesis was to develop pragmatic, ‘big data’ informatic solutions that bridge the gap between the data science of splicing algorithms, and genetic pathology interpretation of splicing variants; specifically, **by developing methods based on empirical evidence to predict *if* and *how* a DNA variant will disrupt RNA splicing in rare disease.**

**Chapter two** analyses 5,145 cryptic-donors activated by 4,811 variants to determine features that dictate which cryptic donor(s) are most likely to be activated when a genetic variant compromises an annotated donor splice-site, among the plethora of decoy donors in the genome. We show that while splice site strength (as measured by six predictive algorithms) and distance from the annotated donor splice-site are influential, these features alone do not dictate spliceosomal selection of a cryptic donor splice site. We found the answer lies in

natural splicing mistakes mined from 40,233 RNA-sequencing samples (40K-RNA database). Simply, ranking the most common cryptic splice sites used in natural splicing mistakes around each annotated splice-site predicts with 87% accuracy the cryptic-donor(s) activated by genetic variants.

In **Chapter three** we extend our empirical method for predicting cryptic-donor activation to predict exon-skipping and cryptic splice-site activation arising from donor or acceptor variants. We update our 40K-RNA database to source unannotated splicing events (mis-splicing mistakes) from 335,301 publicly available RNA-Seq samples (300K-RNA) and create an easy-to-use web portal called *SpliceVault* that hosts both 40K-RNA and 300K-RNA. We show that the same ranking method applied to 300K-RNA accurately predicts 96% of exon-skipping events and 86% of cryptic splicing events activated by 88 variants across 74 genes and 140 affected individuals or heterozygotes subject to RNA Diagnostics in our lab. We also adapt SpliceAI to offer predictions of how RNA splicing will be disrupted and compare the accuracy of SpliceAI's deep-learning with SpliceVault's empirical evidence.

## **Empirical prediction of variant-activated cryptic splice donors using population-based RNA-Seq data**

### **2.1 Overview**

One significant barrier to clinical interpretation of splicing variants has been the fact that they often activate / upregulate one or more of the latent ‘decoy’ splice sites scattered throughout the human genome (‘cryptic’ splicing). The activation of an out of frame cryptic splice site will cause degradation of mRNA transcripts by NMD, whereas the activation of an in-frame cryptic splice site can produce aberrant protein, with vastly different implications for pathology interpretation.

Whether a cryptic splice site will be activated and if so, which of the many decoy splice sites present will be activated, has been an intractable question in pathology interpretation of splicing variants. In Chapter 2 I undertook an analysis of the empirical features which distinguish 5,145 cryptic-donors activated in the event of a splicing variant, from the 86,963 decoy-donors which were not used surrounding those same splice sites.

I assessed the ability of three motif-based measures of splice-site strength (including a novel method developed by our lab), and one deep-learning method (SpliceAI) to distinguish cryptic splice sites from decoy splice sites and found that while cryptic donor activation may be influenced by these measures of splice-site strength, there are other factors likely not being accounted for. Notably, a genome-wide analysis of the depletion of competitive decoy-donors surrounding annotated-donors, as well as the rare stochastic use of decoy-donors across 40,233 publicly available RNA-seq samples (40K-RNA), revealed that the distance between the cryptic-donor and annotated-donor is likely a key determinant, as well as sequence features differentiating the ends of exons and the beginning of introns.

Strikingly, we found that ranking the most common cryptic donors surrounding each splice-site in 40K-RNA provided potent predictive information for which cryptic-donors may be activated in the event of a variant at that splice-site. 40K-RNA showed higher sensitivity than SpliceAI in predicting cryptic-donor activation in the event of a variant at the annotated splice site. We’ve defined an accurate, evidence-based method to predict cryptic-donor

Empirical prediction of variant-activated cryptic splice donors using population-based RNA-Seq data


activation in the context of a variant affecting the annotated-donor, substantially improving the interpretability of donor splicing variants.

This chapter was published as an article for which I was first author:

**Dawes, R.**, Joshi, H. & Cooper, S.T. Empirical prediction of variant-activated cryptic splice donors using population-based RNA-Seq data. *Nat Commun* **13**, 1655 (2022).

<https://doi.org/10.1038/s41467-022-29271-y>

# Empirical prediction of variant-activated cryptic splice donors using population-based RNA-Seq data

Ruebena Dawes <sup>1,2</sup>, Himanshu Joshi<sup>1</sup> & Sandra T. Cooper <sup>1,2,3</sup> 

Predicting which cryptic-donors may be activated by a splicing variant in patient DNA is notoriously difficult. Through analysis of 5145 cryptic-donors (versus 86,963 decoy-donors not used; any GT or GC), we define an empirical method predicting cryptic-donor activation with 87% sensitivity and 95% specificity. Strength (according to four algorithms) and proximity to the annotated-donor appear important determinants of cryptic-donor activation. However, other factors such as splicing regulatory elements, which are difficult to identify, play an important role and are likely responsible for current prediction inaccuracies. We find that the most frequently recurring natural mis-splicing events at each exon-intron junction, summarised over 40,233 RNA-sequencing samples (40K-RNA), predict with accuracy which cryptic-donor will be activated in rare disease. 40K-RNA provides an accurate, evidence-based method to predict variant-activated cryptic-donors in genetic disorders, assisting pathology consideration of possible consequences of a variant for the encoded protein and RNA diagnostic testing strategies.

<sup>1</sup>Kids Neuroscience Centre, Kids Research, Children's Hospital at Westmead, Sydney NSW2145, Australia. <sup>2</sup>Discipline of Child and Adolescent Health, Faculty of Health and Medicine, University of Sydney, Sydney NSW2006, Australia. <sup>3</sup>The Children's Medical Research Institute, 214 Hawkesbury Road, WestmeadNSW 2145 Sydney, Australia. ✉email: [sandra.cooper@sydney.edu.au](mailto:sandra.cooper@sydney.edu.au)

Genetic variants affecting the conserved sequences of the consensus splicing motifs can alter binding of spliceosomal components and induce mis-splicing of precursor messenger RNA (pre-mRNA)<sup>1</sup>, making them a common cause of inherited disorders<sup>2–5</sup>. Splicing variants can simultaneously induce different mis-splicing outcomes, including skipping of one or more exons, activation of a cryptic splice-site(s), and/or retention of one or more introns<sup>1</sup>. Whether induced mis-splicing disrupts the reading frame or affects a region of known functional (and clinical) importance, has significant diagnostic implications. Therefore, knowing the specific mis-splicing outcome of genetic variant is necessary to conclusively link it to a disease. While the accuracy of in silico algorithms in predicting whether a variant will cause mis-splicing has been comprehensively assessed<sup>6–9</sup>, there is currently no reliable means to predict which mis-splicing event(s) may occur in response to a variant that activates mis-splicing. As a result of this and other factors, the vast majority of splice site variants are classified as variants of uncertain significance (VUS); a non-actionable diagnostic endpoint in genomic medicine<sup>10</sup>.

We recently evaluated the accuracy and concordance of SpliceAI (SAI)<sup>11</sup> and algorithms within Alamut Visual® (Interactive Biosoftware, Rouen, France)<sup>12,13</sup>, to predict splicing outcomes arising from genetic variants identified in 74 families with monogenetic conditions subject to RNA diagnostic studies (79 variants; 19% essential GT-AG splice-site variants and 71% extended splice-site variants)<sup>14</sup>. Algorithmic predictions of the strengths of activated cryptic splice sites were highly discordant, especially for cryptic donors. SAI's deep learning showed the greatest accuracy in predicting activated cryptic splice-site(s) (66% true positive with 34% false negative), whereas historical algorithms within Alamut Visual® resulted in 34–69% false negatives<sup>14</sup>.

In this study we focus on determining empirical features that inform prediction of variant-associated spliceosomal selection of a cryptic-donor, in preference to the annotated-donor and other nearby decoy-donors (any GT or GC not used by the spliceosome). Through analysis of 4811 variants in 3399 genes, we show that while splice-site strength and proximity to the annotated-donor strongly influence spliceosomal selection of a cryptic-donor, these factors alone are not sufficient for accurate prediction. Importantly, we show that the most common mis-splicing events seen at each exon-intron junction across 40,233 publicly available RNA-seq samples compiled within the 40K-RNA database, predict with accuracy which cryptic-donor will be activated in rare disease.

## Results

**Reference database of variants activating a cryptic-donor.** We collate a database of cryptic-donor variants, defined as variant-associated erroneous use of a donor other than the annotated-donor. Variants were derived from several sources<sup>11,15,16</sup> (Fig. 1a, see methods). The genomic locations and extended sequences of the annotated-donor, cryptic-donor(s), as well as any decoy-donors (any GT/GC motif within 250 nucleotides (nt) of the annotated-donor), were compiled for analysis. We define the extended donor splice-site region as spanning the fourth to last nucleotide of the exon ( $E^{-4}$ ,  $E$  = exon) to the eighth nucleotide of the intron ( $D^{+8}$ ;  $D$  = donor), as constraint on sequence diversity eases beyond this point (supplementary Fig. 1).

Cryptic-donor variants fall into three categories (Fig. 1b, Box 1): A) Annotated-Modified (AM): a genetic variant modifies the annotated-donor resulting in activation one or more unmodified cryptic-donors ( $n = 2186$ ) (Fig. 1c–e). AM-variants which are SNVs and DNA insertions commonly affect the  $E^{-1}$ ,

$D^{+1}$ ,  $D^{+2}$  and  $D^{+5}$  positions of the annotated-donor (Fig. 1c), and AM-variants which are DNA deletions ranged from 1 to 57 nts in length (Fig. 1d). 89% of AM-variants result in use of a single cryptic-donor, 9% activate 2 cryptic-donors and 2% activate 3 or more cryptic-donors (Fig. 1e).

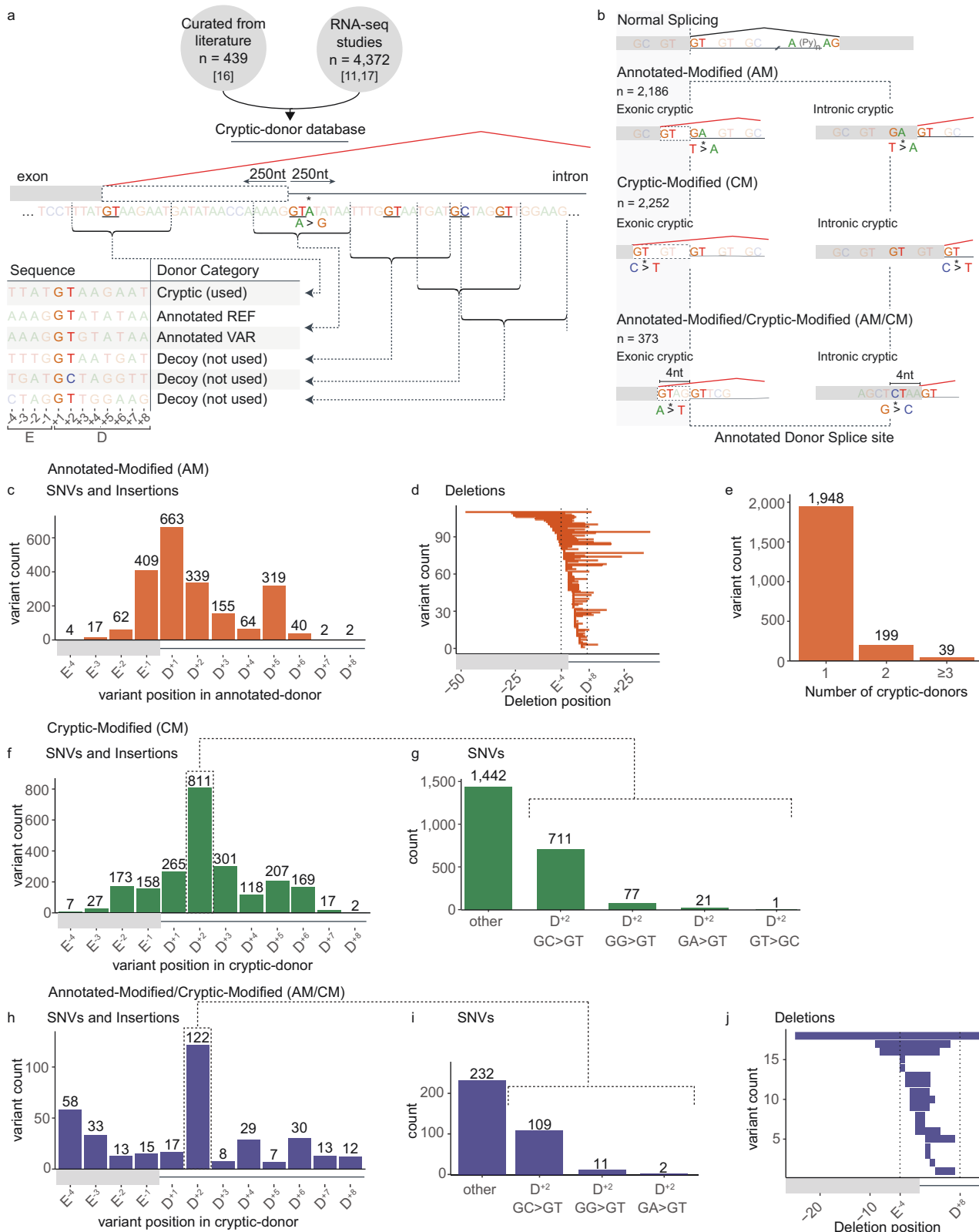
B) Cryptic-Modified (CM): a genetic variant modifies a cryptic-donor and does not affect the annotated-donor ( $n = 2252$ ) (Fig. 1f, g). CM-variants most frequently affect the  $D^{+2}$  position of the cryptic-donor (Fig. 1f), with 32% of all CM SNVs changing the cryptic-donor essential splice motif from GC to GT (Fig. 1g).

C) Annotated-Modified/Cryptic-Modified (AM/CM): a genetic variant that simultaneously modifies the annotated-donor and nearby cryptic-donor ( $n = 373$ ) (Fig. 1h–j). AM/CM-variants which are SNVs and DNA insertions also most frequently affect the  $D^{+2}$  position (122/373) of the cryptic-donor (Fig. 1h), with 31% of SNVs converting a GC to GT (Fig. 1i). AM/CM-variants which are DNA deletions range from 1 to 36 nts in length (Fig. 1j).

**87% of cryptic-donors lie within 250 nt of the annotated-donor.** 99% of cryptic-donors activated by AM-variants and 71% of cryptic-donors activated by CM-variants, lie within 250 nt of the annotated-donor (87% collectively, Fig. 2a, b). By definition, AM/CM-variants activate a cryptic-donor that spatially overlaps the annotated-donor; 26% of AM/CM cryptic-donors lie at either the  $E^{-4}$  or  $D^{+5}$  position (Fig. 2c). For exonic cryptic-donors activated at  $E^{-4}$ , the GT at  $D^{+1/+2}$  of the annotated-donor becomes  $D^{+5/+6}$  of the cryptic-donor; conversely for intronic cryptic-donors activated at  $D^{+5}$ , the GT at  $D^{+5/+6}$  of the annotated-donor becomes  $D^{+1/+2}$  of the cryptic-donor.

While decoy-donors are present everywhere, which ones are selected as cryptic-donors by the spliceosome in the context of a genetic variant appears strongly influenced by their proximity to the annotated-donor (Fig. 2a, b), as shown by their enrichment at proximal locations relative to all decoys present in the genome (Fig. 2d). The steep decline in exonic decoys (Fig. 2d, left) is due to the shorter lengths of exons limiting their frequency at these distances (50th and 90th percentile for exon length shown). Notably, each annotated-donor has on average 36 decoy-donors within  $+/-250$  nt not used by the spliceosome – indicating that features other than proximity to the annotated-donor define a usable cryptic-donor (Fig. 2e).

**Only 31–67% of cryptic-donors are stronger than the annotated-donor.** We examined the ability of four algorithmic measures of splice-site strength to predict cryptic-donor activation (Fig. 3). We compared the performance of MaxEntScan (MES)<sup>13</sup>, NNSplice (NNS)<sup>12</sup> and SpliceAI (SAI)<sup>11</sup> as well as our own method termed Donor Frequency (DF) (see methods and supplementary Fig. 1 for details, supplementary Fig. 2a–c for full plots). DF measures donor strength based on how many annotated-donors in the human genome have the exact same sequence. DF calculates the median frequency of four consecutive windows of nine nucleotides in length (between  $E^{-4}$  and  $D^{+8}$ ) among all annotated-donors, converted to a cumulative frequency distribution. For example, if an  $E^{-3}$  to  $D^{+6}$  sequence has a raw frequency of 222, this combination of nine bases occurs at the analogous position for 222 annotated-donors, corresponding to the 35th percentile of a cumulative frequency distribution across the human genome (see supplementary Fig. 1c). For these and all further analyses, we excluded the 1113 cryptic variants in the database derived from SAI predictions already validated on GTEx RNA-seq data<sup>11</sup>. Our nomenclature of REF and VAR corresponds to the reference (REF) or variant (VAR) donor sequence.



**Fig. 1 Reference database of variants activating a cryptic-donor.** **a** Schematic of the cryptic-donor database. E = Exon, D = Donor. See methods. **b** Three categories of cryptic-donor variants in the database: Annotated-Modified (AM-variants), Cryptic-Modified (CM-variants) and Annotated-Modified and Cryptic-Modified (AM/CM-variants). **c-e** Characteristics of AM-variants ( $n = 2186$ , orange). Positions of AM (**c**) Single Nucleotide Variants (SNVs), insertions and (**d**) deletions relative to the annotated-donor. In (**d**) each of the horizontal bars represents one deletion variant showing the position and width of each deletion relative to the annotated-donor. **e** The number of cryptic-donors activated by each AM-variant. **f-g** Characteristics of CM-variants ( $n = 2255$ ; green). **f** Positions of CM SNVs and insertions relative to the cryptic-donor. **g** Frequency of SNVs resulting in cryptic activation, highlighting the prevalence of GC > GT D<sup>+2</sup> variants. **h-j** Characteristics of AM/CM-variants ( $n = 373$ , blue). **h, i** As in **f, g**. **j** As in **d**.



**Box 1 | Glossary**

*Annotated-donor*: A donor in an ensembl-annotated transcript.

*Decoy-donor*: Any essential donor dinucleotide (GT/GC) that is not an annotated-donor.

*Cryptic-donor*: A decoy-donor shown to be activated (i.e. used by the spliceosome) by a genetic variant.

*Annotated-Modified (AM)*: A genetic variant modifies the annotated-donor resulting in activation one or more unmodified cryptic-donors.

*Cryptic-Modified (CM)*: A genetic variant modifies a cryptic-donor and does not affect the annotated-donor.

*Annotated-Modified/Cryptic-Modified (AM/CM)*: A genetic variant that simultaneously modifies the annotated-donor and nearby cryptic-donor.

*Donor Frequency (DF)*: A measure of donor strength based on how many annotated-donors in the human genome have the exact same sequence.

*Competitive decoy-donor*: A decoy-donor with a DF score at least 10% the score of the nearby annotated-donor.

*40K-RNA*: An aggregated database of splice-junctions detected across 40,233 publicly available RNA-seq samples.

The four algorithms use different methods to measure the intrinsic strength of a given donor splice-site. In the following discussion we use the term stronger and weaker to denote a donor that has a higher or weaker score, respectively, according to that algorithm. Comparisons such as weaker by >50% denote that the donor score has been reduced by more than half by the variant.

For AM-variants, activation of a cryptic-donor typically occurs in the context of a variant that weakens the annotated-donor to less than half of its original strength (Fig. 3a, dark blue). While many AM cryptics are stronger than the annotated<sub>VAR</sub> (Fig. 3c, example shown in Fig. 3b), a substantial subset are not the strongest decoy-donor within 250nt (Fig. 3d). In fact, many activated cryptic-donors are not recognised as bona fide donors by the respective algorithms, notably NNS (Fig. 3e).

Intuitively, for most CM-variants the cryptic is strengthened by the variant (Fig. 3f, orange, example shown in Fig. 3g). However, less than half of activated cryptics are stronger than the annotated-donor (Fig. 3h). Along similar lines, for a majority of AM/CM-variants the annotated-donor is weakened (Fig. 3i, blue) while the adjacent cryptic is strengthened by the variant (Fig. 3j, orange, example shown in Fig. 3k). However, only 29–67% of AM/CM-cryptics<sub>VAR</sub> are stronger than the annotated-donor<sub>VAR</sub> (Fig. 3l). Despite similar overall performance for each algorithm, they showed discordance in variant outcome predictions (Fig. 3M, N) and measures of splice-site strength (Supplementary Fig. 2d).

In summary, four independent algorithms concur that cryptic-donor activation typically occurs in response to weakening of the annotated-donor (85–99% of variants) or strengthening of the cryptic-donor (67–98% of variants). However, only 35–70% of activated cryptic-donors are stronger than the annotated-donor<sub>VAR</sub>, and for unmodified cryptic-donors, 29–62% are not the strongest decoy-donor within 250 nt. Thus, while relative strength of the annotated- and cryptic-donor influence spliceosomal use, there are other factors at play.

**Competitive decoy-donors are depleted close to annotated-donors.** Decoy-donors of comparable or greater strength to the annotated-donor rarely occur within 150 nt (Fig. 4a, red). However, exonic and intronic regions around donors have characteristic single and dinucleotide frequencies which may contribute to the rarity of decoy-donors (supplementary Fig. 3). In particular, the first 50 nt of the intron often shows enrichment in G and T dinucleotides, with distinct patterns: 1) G repeats are enriched in the shortest of introns and T repeats in the longest (supplementary Fig. 3c); 2) Introns with G (or C) at the D<sup>+</sup> position are enriched in G dinucleotides whereas introns with A (or T) at the D<sup>+</sup> position are enriched in T dinucleotides (supplementary Fig. 3d); 3) Introns with rare donors (low DF) are enriched in T-repeats compared to introns with the most common donors (supplementary Fig. 3e). Therefore, we adapted a previously used method<sup>17</sup> which takes dinucleotide preferences into account, to assess if decoy-donors occur more or less

commonly than expected by random chance (see Methods and supplementary Fig. 4).

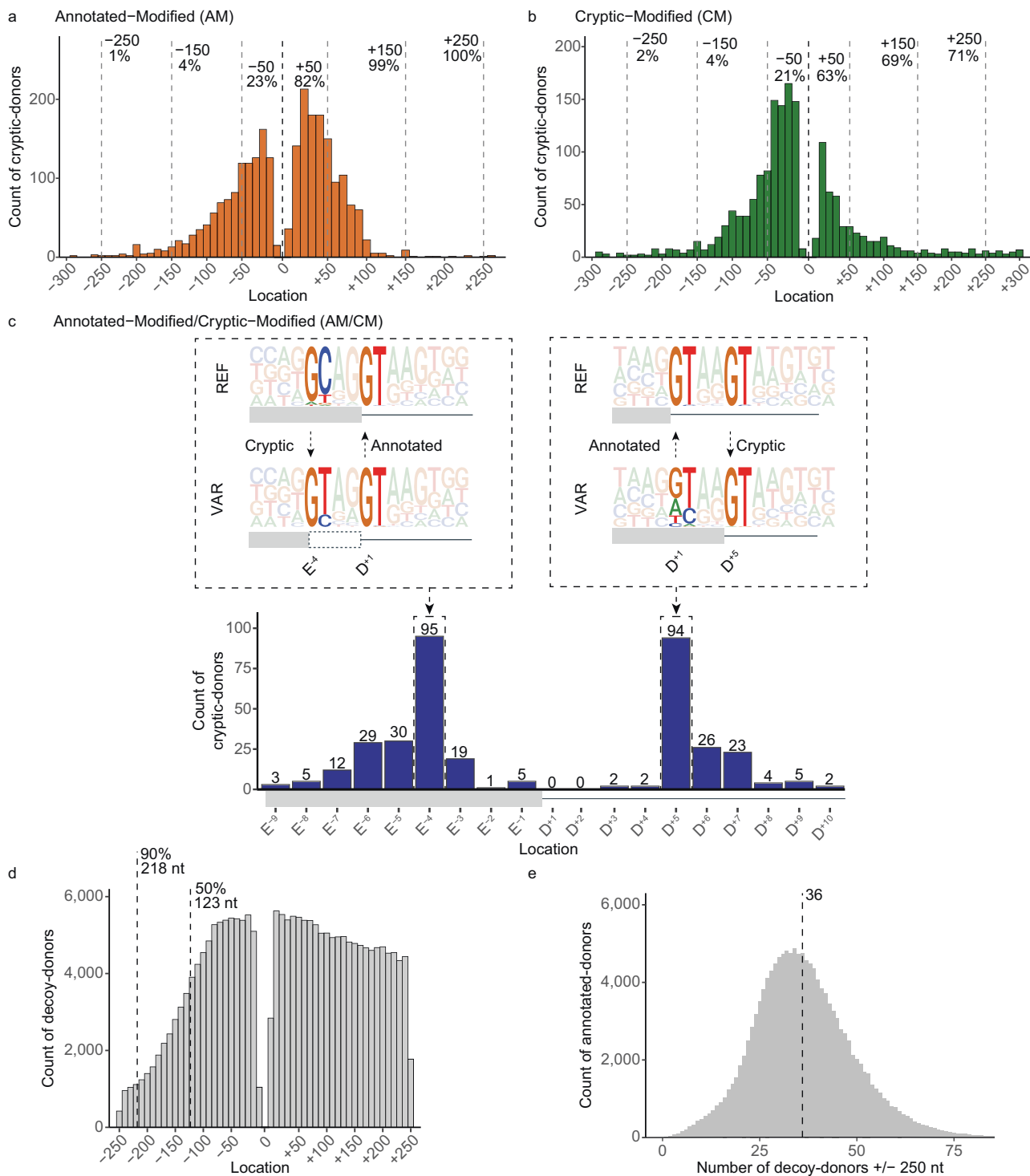
GT decoy-donors show increasing exonic depletion approaching the annotated-donor, with out-of-frame decoys (red) depleted more than in-frame decoy-donors (orange), while showing negligible depletion in the intron (Fig. 4b). GC decoy-donors show no depletion in either the exon or the intron (supplementary Fig. 5a).

We next assessed what proportion of decoy-donors in the genome are used, albeit rarely, via unannotated splice-junctions detected across 40,233 publicly available RNA-seq samples from GTEx<sup>18</sup> and Intropolis<sup>19</sup> (40K-RNA). Unannotated splice-junctions (representing stochastic mis-splicing), seen rarely in RNA-seq samples aggregated across a population, constitute empirical evidence that both splicing reactions can be executed using a decoy-donor. Therefore, we mined 40K-RNA for splice-junctions representing the use of cryptic-donors within 250 nt of any annotated-donor, and ranked them according to the number of samples they were present in (see methods). Overall, ~7% of all unannotated decoy-donors are in fact present as rare, stochastic mis-splicing events in 40K-RNA.

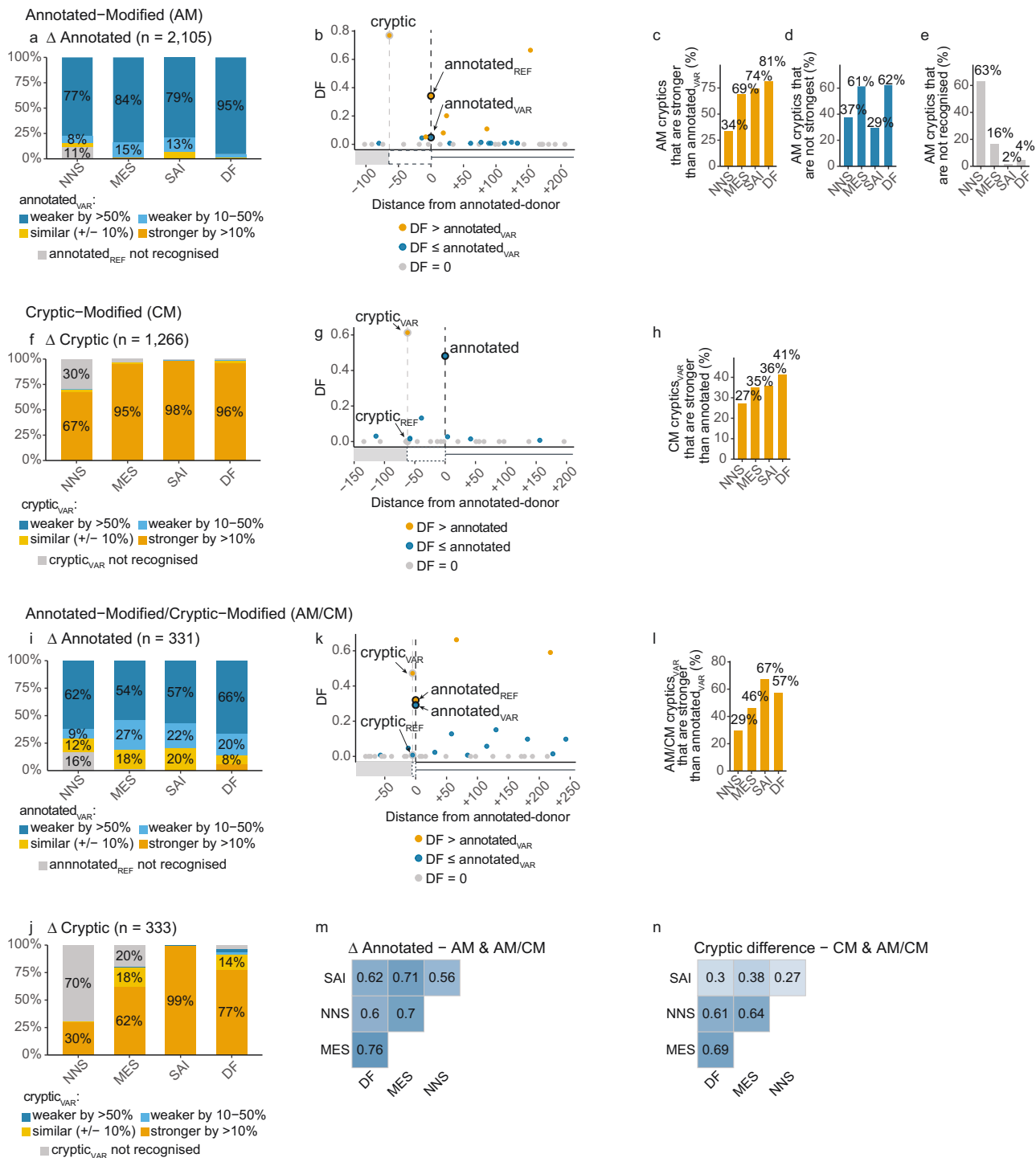
The proportion of exonic GT decoy-donors present in 40K-RNA (relative to all decoys) dramatically increases with proximity to the annotated-donor, with intronic decoys showing only a modest change (Fig. 4c). This mirrors depletion patterns (Fig. 4b) and confirms that decoy-donors closer to the annotated-donor are inherently more likely to be used by the spliceosome. Less than 4% of exonic GC decoy-donors are present in 40K-RNA, even very close to the exon/intron junction, in line with their observed lack of depletion (Supplementary Fig. 5b).

The ability of DF to measure donor strength is evidenced by Fig. 4d, e. While there is negligible depletion of decoy-donor sequences that do not exist as a bona fide donor at any exon-intron junction in GRCh37 (DF = 0, grey), there is clear depletion of exonic decoy-donors closer in DF (50–90% DF, mid-blue), or of similar or greater DF ( $\geq 90\%$  DF, dark blue) (Fig. 4d, left), relative to the annotated-donor. Depletion is even evident for decoy-donors that have DF of only 10% relative to the annotated-donor, and so we define a competitive decoy-donor as one above this threshold. Interestingly, except for the most competitive decoy-donors ( $\geq 90\%$  DF; Fig. 4d, right, dark blue), decoy-donors show no depletion in the intron. Concordantly, the proportion of exonic decoy-donors present in 40K-RNA increases with increasing relative DF, and to a lesser extent at the start of the intron (Fig. 4e).

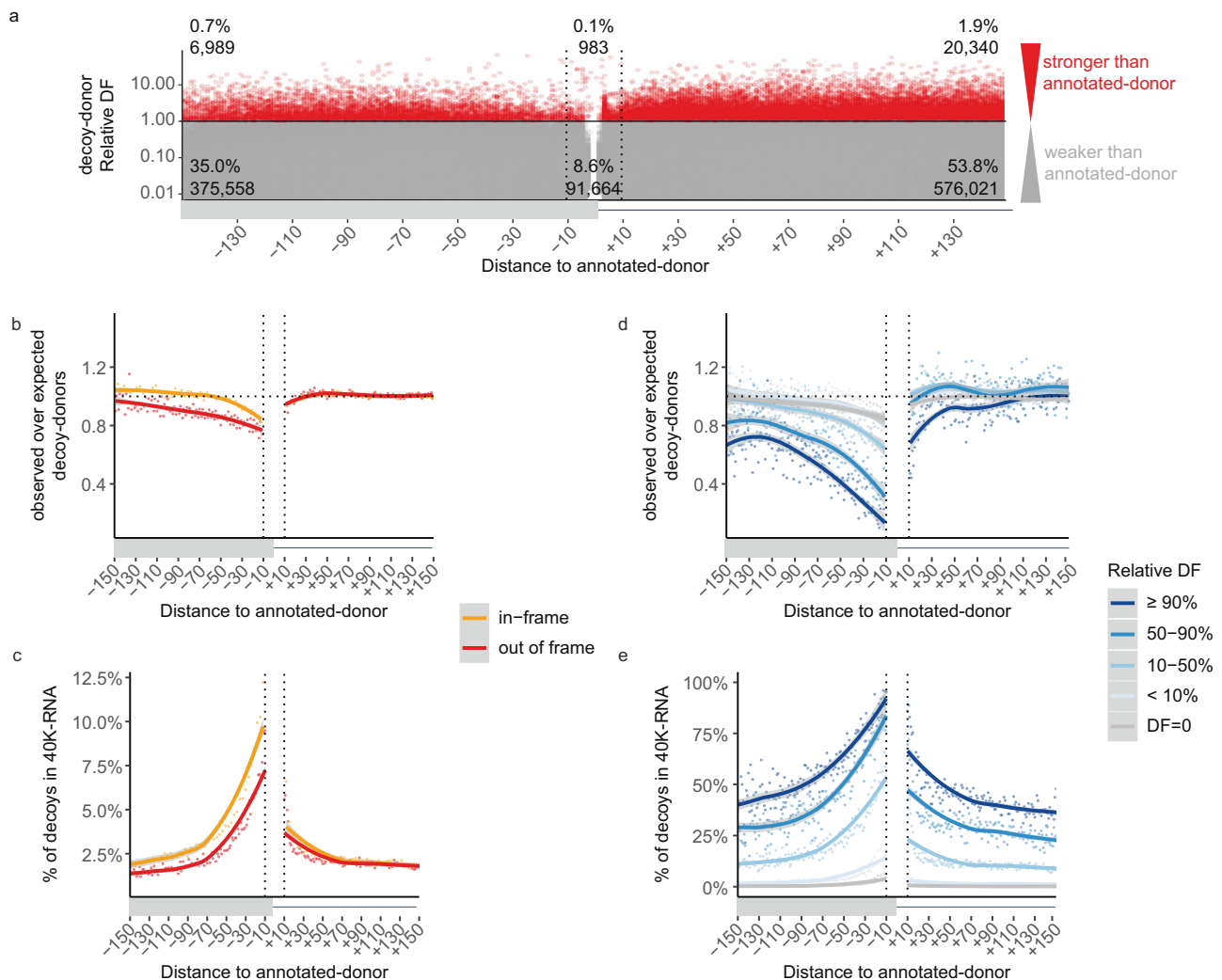
**Why are intronic decoy-donors less likely to be used by the spliceosome?** The fact that intronic decoy-donors are less depleted and less likely to be seen in 40K-RNA (Fig. 4b–e) was initially perplexing, given that cryptic-donors are just as common in the intron as in the exon (Fig. 2a, b). However, we reasoned distinctive nucleotide preferences in the first ~50 nt of the intron



**Fig. 2 Cryptic-donor activation is influenced by proximity to the annotated-donor.** **a, b** Distribution of cryptic-donors activated by (a) AM-variants and (b) CM-variants. Location (x-axis) denotes the distance of the cryptic-donor from the annotated-donor, with negative values upstream into the exon and positive values downstream into the intron. **c** (Bottom) Distribution of cryptic-donors activated by AM/CM-variants. (Top) Pictograms showing the Reference (REF) and Variant (VAR) sequences for AM/CM-variants. Activated cryptic-donors are prevalent at E<sup>-4</sup> (left) and D<sup>+5</sup> (right) due to conserved GTs at D<sup>+1/+2</sup> and D<sup>+5/+6</sup> of the conserved donor splice-site sequence. **d** Frequency of naturally occurring decoy-donors (any GT or GC) +/- 250 nt of annotated-donors in the human genome. Dashed lines indicate the 50th and 90th percentile for exon length. The decline in exonic donors is due to relatively fewer longer exons. **e** Distribution of the number of decoy-donors in the +/- 250 nt surrounding each annotated-donor in the human genome. Dashed line shows that there are an average of 36 decoy-donors within 250 nt of each annotated-donor.



**Fig. 3 Cryptic donor activation is influenced by relative strength. a-e** Assessment of algorithmic scores of splice-site strength for AM-variants. NNS NNSplice, MES MaxEntScan, SAI SpliceAI, DF Donor Frequency. Categories such as weaker by >50% are assigned based on how the score has been impacted by the variant (i.e., more than halved). **a** Proportion of variants with annotated-donor  $\Delta$  scores (Annotated<sub>VAR</sub>/Annotated<sub>REF</sub>) in each of the categories shown in the figure key. Most AM-variants weaken the annotated-donor by >50% (dark blue). See supplementary Fig. 2 for full plots. **b** Example variant showing the Donor Frequency (DF) scores (see supplementary Fig. 1) for the cryptic-donor (DF = 0.77), versus the reference (REF = 0.34) and variant (VAR = 0.05) annotated-donor, as well as surrounding decoys not used. Vertical dotted lines indicate position of annotated- and cryptic-donors. Donors coloured according to the figure key. **c** Percent of cryptic-donors stronger than the annotated<sub>VAR</sub>. **d** Percent of cryptic-donors that are not the strongest donor splice-site within 250 nt. **e** Percent of AM-variant activated cryptic-donors that are not recognised by each algorithm (i.e., score of 0). **f-h** Strength measures for CM-variants. **f** Proportion of variants with cryptic-donor  $\Delta$  scores (Cryptic<sub>VAR</sub>/Cryptic<sub>REF</sub>) in each of the categories shown in legend. Most CM-variants strengthen the cryptic-donor by >10% (dark yellow). See supplementary Fig. 2 for full plots. **g** As in **b**. **h** As in **c**. **i, j** Strength measures for AM/CM-variants. **i** As in **a**. **j** As in **f**. **k** As in **b**. **l** As in **c**. **m, n** Pearson correlation of strength measures. **m**  $\Delta$  Annotated (VAR/REF) for AM & AM/CM-variants (all variants which affect the annotated-donor). **n** Cryptic difference (VAR - REF) for CM & AM/CM-variants (all variants which affect the cryptic-donor).



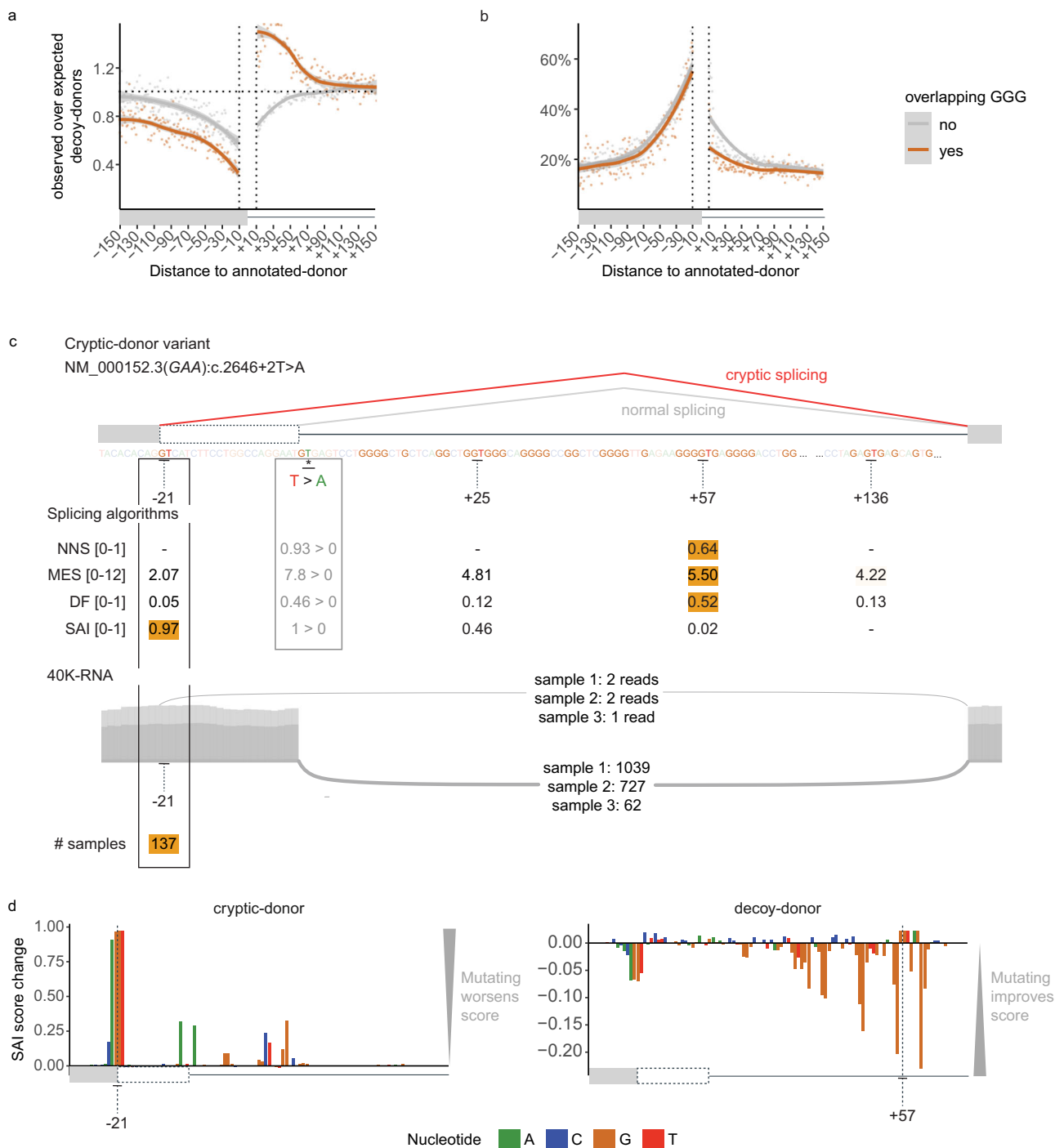
**Fig. 4 Competitive decoy-donors are specifically surrounding annotated-donors.** **a** The relative Donor Frequency (DF) of all decoy-donors within 150 nt of annotated-donors (decoy-donor DF / annotated-donor DF). Plots are shown for +/−150 nt of annotated-donors due to the steeply declining number of exons longer than this. Decoy-donors with a stronger DF score than the annotated-donor are shown in red, otherwise grey. **b** Depletion of GT decoy-splice sites (observed/expected) (see Methods and Supplementary Fig. 4). Exonic donors where use of the decoy-donor would be in-frame are shown in orange, whereas those out-of-frame (or intronic) are shown in red. GT decoy-donors show increasing exonic depletion approaching the annotated-donor, and negligible depletion in the intron. **c** Decoy-donors in-frame and closer to the annotated-donor are more likely to be present in 40,233 publicly available RNA-seq samples (40K-RNA). At each distance from the annotated splice-site, the number of decoy-donors present in 40K-RNA is divided by the total number of naturally occurring decoy-donors at that position **d** depletion of GT decoy-donors as in **b**, split according to decoy-donor DF relative to the annotated-donor (decoy-donor DF/annotated-donor DF). There is negligible depletion of decoy-donor sequences that do not exist as a bonafide donor in GRCh37 (DF = 0, grey), with increasing depletion of exonic decoy-donors closer in DF to the annotated donor (blue gradient). **e** Proportion of GT decoy-donors seen in 40K-RNA as in **c**, split as in **d**. Decoy-donors closer to the annotated-donor and with higher DF relative to the annotated-donor are more likely to be present in 40K-RNA. Lines show LOESS smoothing (locally weighted smoothing i.e., trendlines) with confidence bands in grey.

could affect measures of depletion, and/or, influence the usability of decoy-donors in this region. For example, G-repeat splicing regulatory elements (SREs) are abundant within the first ~50 nt of the intron<sup>20–22</sup>.

We defined competitive decoy-donors as those with a DF of at least 10% that of the associated annotated-donor (see Fig. 4d, e). In the first 50 nt of the intron, competitive decoy-donors overlapping G-triplets show no depletion and conversely appear enriched (Fig. 5a, intron- orange). In contrast competitive decoy-donors not overlapping G-triplets are depleted (Fig. 5a, intron- grey). Additionally, a higher proportion of intronic decoy-donors not overlapping G-triplets are seen in 40K-RNA than those overlapping G-triplets (Fig. 5b, intron- grey). The reciprocity in these data is consistent with a masking effect of intronic G-repeat

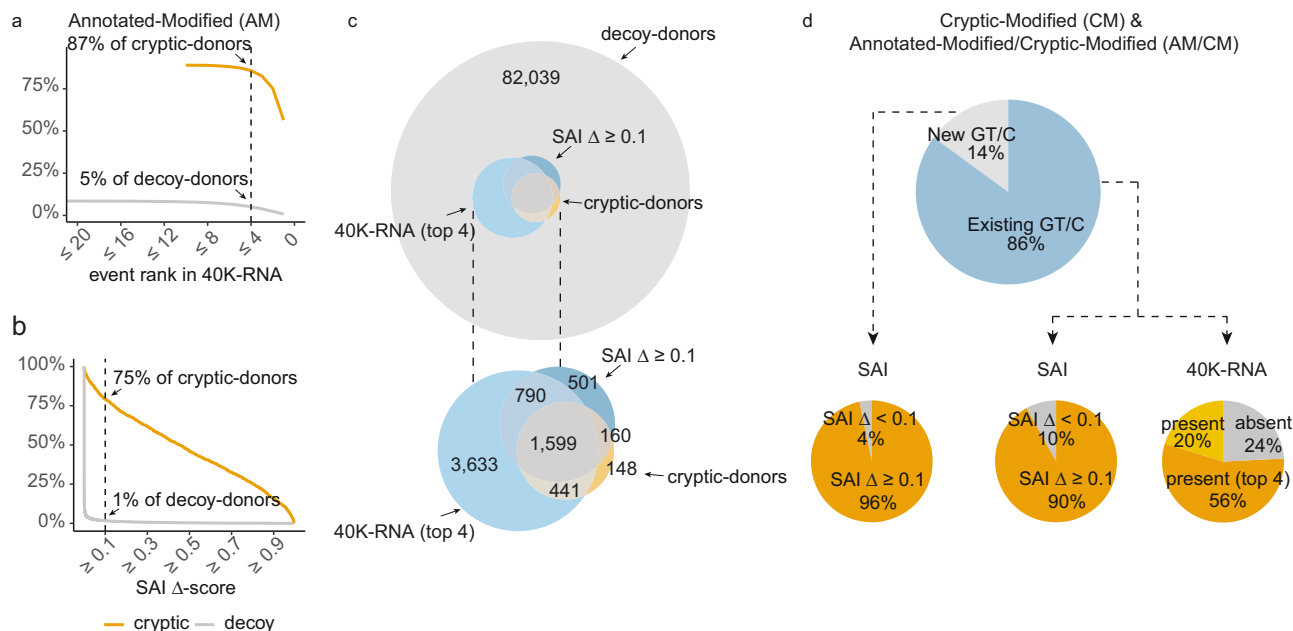
motifs on (competitive) decoy-donors, likely due to RNA secondary structure and/or RNA binding proteins preventing their use.

Figure 5c shows an example variant in gene *GAA* (NM\_000152.3:C.2646 + 2 T > A) identified in an individual affected with glycogen storage disease type II<sup>23</sup> that induces splicing to an exonic cryptic-donor 21 nt upstream of the annotated-donor. NNS, MES, and DF rank the decoy-donor at +57 as the strongest donor - however this donor is enveloped within a G-repeat rich region which may mask it, and accordingly is not present in 40K-RNA. SAI instead predicts use of the cryptic-donor at −21. Notably, this cryptic-donor is present in 137 samples in 40K-RNA, providing empirical evidence that despite its weak primary motif, it can be used by the spliceosome.



**Fig. 5** Utility of 40K-RNA to identify decoy-donors able to be used by the spliceosome. **a**, **b** Depletion (**a**) and proportion in 40K-RNA (**b**) of GT decoy-donors that do or do not overlap G-triplets. Calculated as in Fig. 4b, c. Decoy-donors overlapping G-triplets are depleted in the exon but not in the intron, where they show enrichment because: 1) G repeats are enriched in the first 50 nt of the intron (see supplementary Fig. 3) and 2) donor sequences are commonly G-rich. Plots are limited to decoys with relative DF > 0.1 (defined as competitive with the annotated-donor, see Fig. 4d, e). Lines show LOESS smoothing (locally weighted smoothing i.e. trendlines) with confidence bands in grey. **c** Top: Schematic of an AM-variant identified in an individual with glycogen storage disease type II associated with gene GAA (NM\_000152.3:C.2646+2 T > A)<sup>23</sup> with algorithm scores for annotated- (REF > VAR) decoy- and cryptic-donors. NNS NNSplice, MES MaxEntScan, DF Donor Frequency, SAI SpliceAI. The strongest donor for each algorithm (score range shown in square brackets) is coloured orange. Below: Sashimi plot from three GTEx RNA-seq samples identifying use of the -21 cryptic-donor as present in 40K-RNA (At least 1 read is detected in 137/30753 samples with detectable expression of transcript). SpliceAI (SAI) correctly scores the -21 cryptic-donor as the most likely cryptic-donor. **d** Result of SAI in silico mutagenesis showing the bases contributing to predicted strength of the -21 cryptic-donor (left) and +57 decoy-donor (right). SAI score change denotes the decrease (if positive) or increase (if negative) on the predicted strength of the donor when that nucleotide is mutated (see methods). Note that the presence of the cryptic-donor at -21 and intronic G-repeats negatively impact the score of the +57 decoy-donor according to SAI.





**Fig. 6 40K-RNA potentially informs cryptic-donor activation.** **a, b** The percentage of Authentic-Modified (AM) cryptic-donors correctly predicted, and decoy-donors incorrectly predicted at different cut-offs for **(a)** event rank in 40K-RNA and **(b)** SpliceAI (SAI)  $\Delta$  score ( $\text{donor}_{\text{VAR}} - \text{donor}_{\text{REF}}$ ). **a** The dotted line denotes cryptic- and decoy-donors predicted using a cut-off of events ranked 4 or less. 87% of cryptic-donors activated by AM-variants are among the top 4 events present in 40K-RNA within 250 nt of the annotated-donor, with 95% of decoy-donors not in the top 4 events. **b** While SpliceAI (SAI) outperforms other algorithms using a cut-off of  $\Delta$  scores 0.1 and above (see supplementary Fig. 6a), it predicts only 75% of cryptic-donors ( $\text{SAI } \Delta \geq 0.1$ ). SAI accurately excludes 99% of decoy-donors. **c** (top) Venn diagram showing the overlap of cryptic-donors with those predicted by SAI or our 40K-RNA method, among the entire pool of decoy-donors. (bottom) magnification of internal Venn. **d** Cryptic-donors activated by CM- and AM/CM-variants. (top) New GT/C (Light grey) denotes variants creating a GT or GC essential splice-site motif (40K-RNA is inherently unsuitable for these variants). Existing GT/C (blue) denotes variants that modify a decoy-donor with a pre-existing GC or GT essential splice-site. (bottom) For 40K-RNA, the orange segment denotes cryptic-donors in the top 4 events, yellow segment denotes where the cryptic-donor is present in 40k-RNA but is not in the top 4 events, and grey denotes cryptic-donors absent from 40K-RNA.

SAI in silico mutagenesis of the cryptic-donor at -21 and decoy-donor at +57 show that SAI deep-learning perceives the negative impact of the G-repeats on the usability of the +57 cryptic-donor (Fig. 5d). Intronic G-repeats are known examples of SREs<sup>20,24</sup> (see Fig. 5 and additional examples supplementary Fig. 5c–f). Whether or not a cryptic donor can be used is influenced by a constellation of features: the consensus donor sequence, as well as proximal and more distal splicing regulatory elements. Regulatory elements are not factored by many algorithms, though may be identified by SAI, likely underpinning its enhanced capabilities in recognition of usable (cryptic) splice-sites. In contrast, 40K-RNA uses empirical evidence from RNA-Seq data that reveals which cryptic splice-sites are usable in the context of the specimens tested.

**90% of cryptic-donors in AM-variants are present in 40K-RNA.**

We assessed whether 40K-RNA provides a viable means to prioritise cryptic-donors likely to be activated in the context of a genetic variant affecting the annotated-donor (i.e. AM-variants). 90% of AM-variant activated cryptic-donors are present in 40K-RNA, while 91% of unused decoy-donors are absent. Therefore, 40K-RNA provides potent predictive information with respect to both true positives (cryptic-donors) and true negatives (decoy-donors). Notably, while cryptic-donors were observed in multiple independent samples across 40K-RNA, they were typically very low frequency splice-junctions (44% had a maximum of 4 reads or less in any one sample, supplementary Fig. 6b).

We chose the top 4 40K-RNA events at each splice-junction (or all events if there were less than 4 detected) as our predicted

cryptic-donors as this maximised sensitivity (87%) without compromising specificity (95%) (Fig. 6a). Use of 40K-RNA had a higher sensitivity than all four algorithms assessed (Fig. 6a, b, supplementary Fig. 6a). The sensitivity of 40K-RNA is inherently influenced by read-depth of the target transcript: more than 85% of cryptic-donors are detected in transcripts with a read depth of >250 for the annotated exon-exon splice-junction (normal splicing); whereas only 29% of cryptic-donors are detected in 40K-RNA in transcripts where normal splicing had a maximum read count of <100 (supplementary Fig. 6c). Consequently, we assessed SAI as a complementary approach for situations where our empirical method is underpowered or not well suited.

We define SAI prediction of cryptic-donor activation as a donor-gain  $\Delta$ -score of 0.1 or greater, which accurately predicts 75% of cryptic-donors and inaccurately predicts only 1% of decoy-donors (Fig. 6b). SAI showed higher sensitivity than NNS, and comparable sensitivity to MES and DF, while greatly improving on their specificity (supplementary Fig. 6a). However, the sensitivity of SAI is compromised for cryptics at increasing distance from the annotated-donor - only 55% of cryptic-donors further than 100 nt from the annotated splice site had a  $\Delta$ -score above 0.1 (supplementary Fig. 6d, e). If we take the union of SAI and 40K-RNA cryptic-donor predictions (i.e., cryptics predicted by either of the two methods), we accurately predict 2210/2389 (93%) of cryptic-donors (Fig. 6c) and inaccurately predict 6% of unused decoy-donors.

Use of 40K-RNA has caveats for CM-variants and AM/CM-variants, and cannot be used for variants that create a GT (or GC) motif. However, for the subset of CM-variants and AM/CM-variants where the variant modifies the extended splice site region

of an extant GT/C decoy-donor (1525 variants, Fig. 6d, top- blue), 76% are present in 40K-RNA, with 56% in the top 4 events.

40K-RNA is least sensitive for variants that most significantly impact the strength of the cryptic-donor: For D<sup>+2</sup> CM-variants, only 32% of the cryptic-donors are present in the top 4 events, as compared to 85% for E<sup>-3</sup> variants (supplementary Fig. 6f; E<sup>-3</sup> is the third to last exonic base). Accordingly, even if a GC decoy-donor is not present in 40K-RNA, conversion by a variant to a GT donor presents high risk for cryptic-activation. SAI performed well for CM-variants and AM/CM-variants, correctly predicting 96% of variants that created an essential donor motif and 90% which modified an existing essential motif (Fig. 6d).

## Discussion

The ultimate goal of splicing predictions is to determine if and how a genetic variant will induce mis-splicing of pre-mRNA. Even for essential splice-site variants that almost invariably cause mis-splicing, consideration of probable consequences of the variant is critical for pathology application of the ACMG-AMP code PVS1<sup>25</sup> (null variant due to presumed mis-splicing of the pre-mRNA) and of equal importance to strategise functional testing for RNA diagnostics<sup>14</sup>. While activation of a cryptic-donor 6 nucleotides away will remove or insert two amino-acids, activation of a cryptic-donor 4 nucleotides away will induce a frameshift, with vastly different implications for pathology interpretation.

We learned five key lessons from our analyses of 4811 cryptic-donor variants in 3399 genes: (1) Decoy-donors that show evidence of natural stochastic use by the spliceosome in population-based RNA-Seq data (i.e., are present in 40K-RNA) have the greatest probability of activation as cryptic-donors. (2) Decoy-donors closer to the annotated splice site are inherently more likely to be used by the spliceosome, likely due to the presence of all required sequence features that are facilitating use of the annotated donor.

(3) Cryptic-donors do not necessarily need to be stronger than the annotated-donor to present substantive risk for mis-splicing, with decoy-donors only 10% of the strength of the annotated-donor able to compete for spliceosomal binding. (4) Intronic G-repeats can diminish the likelihood of spliceosomal recognition and use of intronic decoy splice sites. (5) SAI's deep-learning appreciates the broader sequence context determining spliceosomal usability of a cryptic-donor, though less accurately predicts activation of more distal cryptic-donors (>100 nt from the annotated-donor).

SAI's deep learning presents a major improvement in predicting cryptic-donor activation. However, use of SAI in a pathology context is limited by the challenge of deriving a clinically-meaningful interpretation of a number on a 0–1 scale, without independently verifiable evidence. In contrast, 40K-RNA provides an accurate, evidence-based means to rank cryptic-donors likely to be activated by genetic variants.

Brandão et al.<sup>26</sup> used deep sequencing of twelve major cancer susceptibility genes to catalogue all alternative and aberrantly spliced transcripts. They found variant-activated cryptic splicing was often seen at much lower levels in disease controls, suggesting that the dominant transcript in rare disease may be seen as a stochastic mis-splicing event in other samples. We use this insight, mining the breadth of publicly available RNA-seq data across numerous tissues to comprehensively catalogue stochastic cryptic splicing events across all genes.

The heightened sensitivity and empirical nature of using 40K-RNA is of vital importance for pathology assessment of variants affecting the essential donor splice-site, as not considering a likely cryptic-donor activated can lead to profound complications in

variant interpretation. Prospectively, the sensitivity of 40K-RNA can be enhanced by ultra-deep sequencing. It is also easy to envisage extending the method to predict other mis-splicing events such as exon skipping, and mis-splicing events at the acceptor splice site. 40K-RNA can reliably identify distal cryptic-donors with high likelihood of activation, which may not be identified by SAI. Conversely, SAI can reliably identify cryptic donors with high likelihood of activation not detected in 40K-RNA, due to low read depth of the target gene.

In conclusion, we define an accurate, evidence-based method to predict cryptic-donor activation in the context of a variant affecting the annotated-donor, based on stochastic mis-splicing events observed in 40,233 publicly available RNA-seq samples (40K-RNA). We provide a web resource that reports and ranks the most commonly (mis)used cryptic donors proximal to every ensemble annotated-donor<sup>27</sup> (<https://kidsneuro.shinyapps.io/splicevault-40k/>). Our research establishes that for AM-variants, if a cryptic-donor is activated, in 87% of cases it will be among the top 4 events. We hope this evidence-based method may improve clinical interpretability of donor variants.

## Methods

**Creating a database of cryptic-donor variants.** Variants were derived from several sources: (1) 439 variants curated from literature, predominantly comprised of 364 variants in DBASS<sup>15</sup> and supplemented by curation from published literature of 75 additional variants<sup>28,29</sup> (2) 4372 variants derived from RNA-seq studies: Variant-associated aberrant cryptic-donor activation detected from RNA-seq data identified by SAVnet in somatic tumor samples ( $n = 3259$ )<sup>16</sup> and 1113 variants identified in GTEx samples by spliceAI and verified using RNA-seq data<sup>11</sup>. The following inclusion criteria applied: (1) Variants had to occur within E<sup>-4</sup>-D<sup>+8</sup> of the annotated or the cryptic-donor, otherwise they were excluded as outside the bounds of this analysis. (2) annotated cryptic-donors were within the same exon/intron as the variant (i.e., between the 5' end of the exon and 3' end of the intron surrounding the affected donor). (3) The annotated cryptic-donor VAR sequence had to have an essential GT/GC dinucleotide at D<sup>+1</sup>/D<sup>+2</sup>, to minimise mis-annotated variants being included.

**Annotating variant categories.** We annotated variants with categories we defined— if the variant occurred within E<sup>-4</sup>-D<sup>+8</sup> of the annotated-donor, it was an AM-variant, if it occurred within E<sup>-4</sup>-D<sup>+8</sup> of the cryptic-donor it was a CM-variant, and if it occurred within E<sup>-4</sup>-D<sup>+8</sup> of both the annotated- and cryptic-donor it was an AM/CM-variant. For 37/373 of AM/CM-variants, an additional unmodified cryptic-donor was activated, in addition to the cryptic-donor modified by the variant— these were excluded from analyses.

**Compiling annotated-, cryptic- & decoy-donor sequences.** The R package BSgenome.Hsapiens.1000genomes.hs37d5<sup>30</sup> was used to extract (up to) 500 nt of genomic sequence preceding and succeeding the annotated-donor (GRCh37). For each variant in the cryptic-donor database, we extracted up to 250 exonic nucleotides in the 5' direction (i.e., if the exon was only 50 nt the window of analysis would be 50 nucleotides), and up to 250 intronic nucleotides in the 3' direction, in the same fashion (Fig. 1a).

From the (up to) 500 nt of sequence we pulled E<sup>-4</sup>-D<sup>+8</sup> sequences for the annotated- and cryptic-donor before and after each variant (REF and VAR respectively). We also identified any other essential donor dinucleotides (i.e., GT or GC) which were present in the sequence and extracted the E<sup>-4</sup>-D<sup>+8</sup> sequence surrounding them. These sequences we define as decoy-donor- sequences containing the essential donor dinucleotides (i.e., a GT or a GC) but which weren't utilised by the spliceosome as a result of the variant (Fig. 1a, lower). For intronic decoy-donors, we excluded any which would result in an intron too short to be spliced (as defined by the 1<sup>st</sup> percentile for intron length in the human genome = 80 nt)<sup>31</sup>. Importantly, without additional filtering, no cryptic-donors in the database violated this rule.

**Algorithms for splice site strength.** We retrieved predicted scores for annotated-donors, cryptic-donors and decoy-donors in the database in both the REF and VAR sequence context, for four algorithms. (1) MaxEntScan (MES)<sup>13</sup> scores were retrieved using the perl script provided at [http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan\\_scoreseq.html](http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html). MES scores below 0 were standardised to 0 as predicted non-functional splice sites (2) NNSplice (NNS)<sup>12</sup> scores were retrieved using the online portal ([https://www.fruitfly.org/seq\\_tools/splice.html](https://www.fruitfly.org/seq_tools/splice.html)), set to human, with default settings (i.e., a minimum score of 0.4, with any scores below predicting a non-functional splice site) (3) SpliceAI (SAI)<sup>11</sup> scores were retrieved using a script adapted from the SAI GitHub repository (<https://github.com/illumina/SpliceAI>) which allows spliceAI to score custom sequences. We rounded

the scores to three decimal places, and scores at 0 when rounded as such ( $i.e., < 0.01$ ) were termed non-functional splice site predictions. (4) Donor Frequency (DF) calculates the median frequency among four 9 nt windows of sequence spanning the donor (see supplementary Fig. 1b, c) converted to a cumulative percentile distribution scale. DF measures donor strength based on how many annotated-donors in the human genome have the exact same sequence. An example of a DF calculation is shown in supplementary Fig. 1c, where a median DF raw value of 179 lies at the 31st percentile of a cumulative frequency distribution. After assessing several window lengths (supplementary Fig. 1a) we used 9nt windows as optimally encompassing the splice site.

**Naturally occurring decoy-donors.** Our set of naturally occurring human decoy-donors were derived from the set of all canonical Ensembl transcripts (Release 75)<sup>27</sup>, with first and last introns and single exon transcripts removed. We filtered the set so that junctions were within the open reading frame for the given gene, so we knew that cryptic splicing here would affect the protein. We also removed exons with alternative 5' or 3' ends already annotated in different transcripts. We used these criteria to form a set of 142,014 canonical exon-intron junctions that are not alternatively spliced (or at least not annotated as such). We extracted sequences surrounding annotated-donors and extracted all decoy-donors just as for the cryptic database (see methods section creating a database of cryptic-donor variants).

**Decoy-donor depletion.** Decoy-donor depletion was calculated using a method we adapted from a previous study<sup>17</sup> that controls for dinucleotide frequencies (supplementary Fig. 4). For exonic sequences, we took up to 150 nt or the maximum length of the exon, whichever was shorter (and similarly for the intron, stopping 50nt from the acceptor). We limited analysis to 150nt of exonic sequence as the majority of exons are shorter than this. We then shuffled exonic and intronic sequences separately, maintaining dinucleotide frequencies (using shuffle\_sequences with euler method from the universal motif R package<sup>32</sup>). We performed the shuffle 15 times and took the average count of decoy-donors at each nucleotide position as our expected count at this position. The observed count of decoy-donors was then divided by the expected count at each position.

**Creating 40K-RNA.** We had two sources of data for 40K-RNA- RNA-seq data from Intropolis<sup>19</sup> and GTEx<sup>18</sup>. Intropolis is a set of ~42 M splice-junctions found across 21,504 human RNA-seq samples from the Sequence Read Archive (SRA). Samples were aligned using Nellore et al. annotation-agnostic aligner Rail-RNA<sup>33</sup>. Intropolis was downloaded from its dedicated github repository (<https://github.com/nellore/intropolis>). Per sample splice-junction files were obtained from GTEx (phs000424.v8.p2 [[https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000424.v8.p2](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2)]). Using Datamash<sup>34</sup>, splice-junction read counts were summarised across all samples for each unique splice-junction and translated from GRCh38 to GRCh37 using liftOver<sup>35</sup>.

For each set of splice-junctions (Intropolis and GTEx), we cross-referenced and located junctions within ensembl transcripts. We filtered to cryptic-donor events by scanning for any unannotated donors used between the 5' end of the exon and the 3' end of the intron for that respective exon-intron junction, where the junction also spliced to the next annotated acceptor. Events from the two sources were merged, sample counts were tallied across the two datasets, and splice-junctions present in at least 3 samples and representing cryptic-donor use within 250 nt of any annotated-donor were retained.

**Sashimi plots.** For Fig. 5c, and S6b, c sashimi plots were generated using 3 GTEx bam files for each example, each from the tissue with the highest TPM for the respective gene. Sashimi plots were created using ggsashimi<sup>36</sup>.

**SpliceAI in silico mutagenesis plots.** For Fig. 5d and S6b, c we performed the in silico mutagenesis method described by Jaganathan et al<sup>11</sup>. That is, the importance score of each nucleotide was calculated as:

$$s_{actual} = \frac{s_A + s_C + s_G + s_T}{4} \quad (1)$$

where  $s_{actual}$  is the score calculated on the genuine sequence, and  $s_A$ , for example, is the score calculated when an A is substituted at this position.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The variants used in the cryptic-donor database are provided in the Source Data file. 40K-RNA is available as a web-resource at: <https://kidsneuro.shinyapps.io/splicevault-40k/>. Additionally, the full dataset is available under restricted access to limit hosting costs. Access can be obtained by creating a google cloud billing account and downloading at this link using google cloud tools- [https://storage.googleapis.com/misspl-db-data/misspl\\_events\\_40k\\_hg19.sql.gz](https://storage.googleapis.com/misspl-db-data/misspl_events_40k_hg19.sql.gz). The GTEx v8 data used in this study were obtained from

dbGaP accession number phs000424.v8.p2 [[https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000424.v8.p2](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2)]. Intropolis data used in this study were obtained from the dedicated GitHub repository <https://github.com/nellore/intropolis>. Source data are provided with this paper.

## Code availability

All code required to replicate figures in the study are available in a GitHub repository: [https://github.com/kidsneuro-lab/cryptic\\_donor\\_prediction](https://github.com/kidsneuro-lab/cryptic_donor_prediction). Additionally, code required to create 40K-RNA is available in a separate repository <https://github.com/kidsneuro-lab/40K-RNA>.

Received: 18 July 2021; Accepted: 1 March 2022;

Published online: 29 March 2022

## References

- Anna, A. & Monika, G. Splicing mutations in human genetic disorders: Examples, detection, and confirmation. *J. Appl. Genet.* **59**, 253–268 (2018).
- López-Bigas, N., Audit, B., Ouzounis, C., Parra, G. & Guigó, R. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.* **579**, 1900–1903 (2005).
- Ars, E. Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. *Hum. Mol. Genet.* **9**, 237–247 (2000).
- Esquerra-Inchausti, M. et al. High prevalence of mutations affecting the splicing process in a Spanish cohort with autosomal dominant retinitis pigmentosa. *Sci. Rep.* **7**, 39652 (2017).
- Teraoka, S. N. et al. Splicing defects in the ataxia-telangiectasia gene, ATM: Underlying mutations and consequences. *Am. J. Hum. Genet.* **64**, 1617–1631 (1999).
- Colombo, M. et al. Comparative in vitro and in silico analyses of variants in splicing regions of BRCA1 and BRCA2 genes and characterization of novel pathogenic mutations. *PLoS ONE* **8**, e57173 (2013).
- Houdayer, C. et al. Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. *Hum. Mutat.* **33**, 1228–1238 (2012).
- Jian, X., Boerwinkle, E. & Liu, X. In silico tools for splicing defect prediction: A survey from the viewpoint of end users. *Genet. Med.* **16**, 497–503 (2014).
- Tang, R., Prosser, D. O. & Love, D. R. Evaluation of Bioinformatic Programmes for the Analysis of Variants within Splice Site Consensus Regions. *Adv. Bioinforma.* **2016**, 5614058 (2016).
- Truty, R. et al. Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation. *Am. J. Hum. Genet.* **108**, 696–708 (2021).
- Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
- Reese, M. G., Eeckman, F. H., Kulp, D. & Haussler, D. Improved splice site detection in genies. *J. Comput. Biol.* **4**, 311–323 (1997).
- Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA Splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
- Bournazos, A. M. et al. Standardized practices for RNA diagnostics using clinically accessible specimens reclassifies 75% of putative splicing variants. *Genet. Med.* (2021) <https://doi.org/10.1016/j.gim.2021.09.001>.
- Buratti, E., Chivers, M., Hwang, G. & Vorechovsky, I. DBASS3 and DBASS5: databases of aberrant 3'- and 5'-splice sites. *Nucleic Acids Res.* **39**, D86–D91 (2011).
- Shiraishi, Y. et al. A comprehensive characterization of cis-acting splicing-associated variants in human cancer. *Genome Res.* **28**, 1111–1125 (2018).
- Iacono, M., Mignone, F. & Pesole, G. uAUG and uORFs in human and rodent 5'untranslated mRNAs. *Gene* **349**, 97–105 (2005).
- GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
- Nellore, A. et al. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biol.* **17**, 266 (2016).
- McCullough, A. J. & Berget, S. M. G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol. Cell. Biol.* **17**, 4562–4571 (1997).
- Caputi, M. & Zahler, A. M. Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 Family. *J. Biol. Chem.* **276**, 43850–43859 (2001).
- Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711 (2015).



23. Huie, M. L., Anyane-Yeboah, K., Guzman, E. & Hirschhorn, R. Homozygosity for multiple contiguous single-nucleotide polymorphisms as an indicator of large heterozygous deletions: Identification of a novel heterozygous 8-kb intragenic deletion (IVS7–19 to IVS15–17) in a patient with glycogen storage disease Type II. *Am. J. Hum. Genet.* **70**, 1054–1057 (2002).
24. Xiao, X. et al. Splice site strength-dependent activity and genetic buffering by poly-G runs. *Nat. Struct. Mol. Biol.* **16**, 1094–1100 (2009).
25. Abou Tayoun, A. N. et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum. Mutat.* **39**, 1517–1524 (2018).
26. Brandão, R. D. et al. Targeted RNA-seq successfully identifies normal and pathogenic splicing events in breast/ovarian cancer susceptibility and Lynch syndrome genes. *Int. J. Cancer* **145**, 401–414 (2019).
27. Howe, K. L. et al. Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
28. Leman, R. et al. Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: An international collaborative effort. *Nucleic Acids Res.* **46**, 7913–7923 (2018).
29. Pros, E. et al. Nature and mRNA effect of 282 different NF1 point mutations: Focus on splicing alterations. *Hum. Mutat.* **29**, E173–E193 (2008).
30. Gehring, J. *BSgenome.Hsapiens.1000genomes.hs37d5: 1000genomes Reference Genome Sequence (hs37d5). R package version 0.99.1.* (2016).
31. Bryen, S. J. et al. Pathogenic abnormal splicing due to intronic deletions that induce biophysical space constraint for spliceosome assembly. *Am. J. Hum. Genet.* **105**, 573–587 (2019).
32. Tremblay, B. *universalmotif: Import, Modify, and Export Motifs with R. R package version 1.8.4.* (2021).
33. Nellore, A. et al. Rail-RNA: Scalable analysis of RNA-seq splicing and coverage. *Bioinformatics* **33**, 4033–4040 (2016).
34. Free Software Foundation, I. *GNU Datamash, Available at: <https://www.gnu.org/software/datamash/>.* (2014).
35. Hinrichs, A. S. et al. The UCSC genome browser database: Update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
36. Garrido-Martin, D., Palumbo, E., Guigó, R. & Breschi, A. ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. *PLOS Comput. Biol.* **14**, e1006360 (2018).

## Acknowledgements

This project was supported by a National Health and Medical Research Council of Australia Senior Research Fellowship (S.T.C. APP1136197) and Ideas Grant (S.T.C. APP1186084). R.D. is supported by a University of Sydney Research Training Program Scholarship and Merit Award Supplementary Scholarship. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS.

## Author contributions

Data curation and analysis: R.D. and H.J. Funding acquisition and supervision: S.T.C. Visualization: R.D. Writing – original draft: R.D. Writing review and editing: R.D. and S.T.C.

## Competing interests

S.T.C. and H.J. are named inventors of Intellectual Property (IP) described in part within this manuscript owned jointly by the University of Sydney and Sydney Children's Hospitals Network. S.T.C. is director of Frontier Genomics Pty Ltd (Australia) who have licenced this IP. S.T.C. receives no payment or other financial incentives for services provided to Frontier Genomics Pty Ltd (Australia). Frontier Genomics Pty Ltd (Australia) has no existing financial relationships that will benefit from publication of these data. The remaining co-authors declare no conflicts of interest.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-29271-y>.

**Correspondence** and requests for materials should be addressed to Sandra T. Cooper.

**Peer review information** *Nature Communications* thanks Graziano Pesole and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

## Supplementary Information

Empirical prediction of variant-activated cryptic splice donors using population-based RNA-Seq data

Ruebena Dawes<sup>1,2</sup> Himanshu Joshi<sup>1</sup>, and Sandra T. Cooper<sup>1,2,3\*</sup>

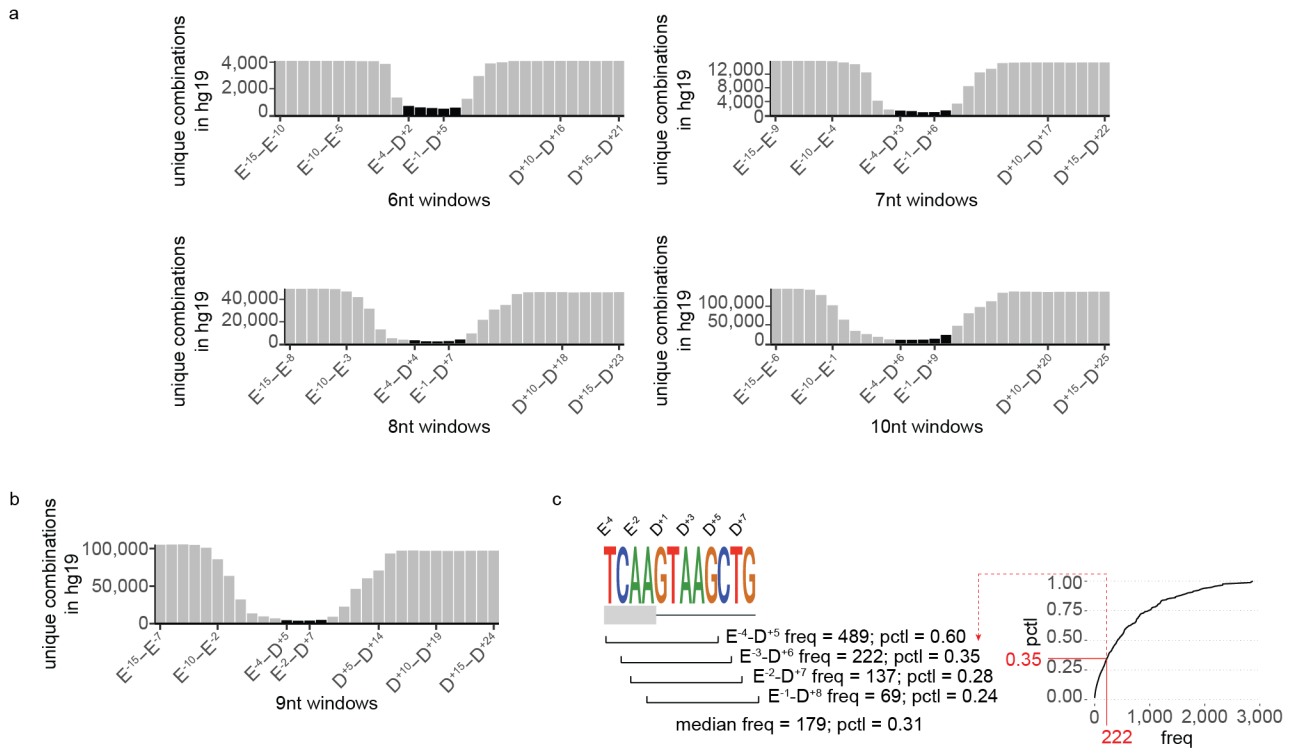
<sup>1</sup>Kids Neuroscience Centre, Kids Research, Children's Hospital at Westmead, Sydney, NSW2145, Australia

<sup>2</sup>Discipline of Child and Adolescent Health, Faculty of Health and Medicine, University of Sydney, Sydney, NSW2006, Australia

<sup>3</sup>The Children's Medical Research Institute, 214 Hawkesbury Road, Westmead NSW 2145, Sydney, Australia

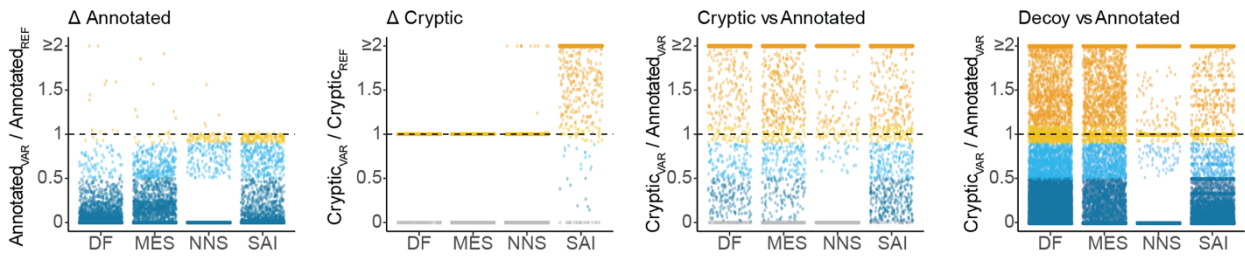
### **This PDF File Includes:**

Supplementary Figures 1-6

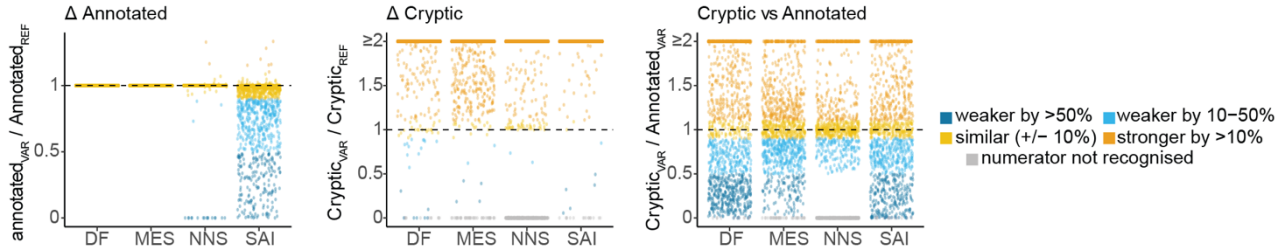


**Supplementary Fig. 1 Calculation of Donor Frequency (DF) as a measure of donor strength. a-b)** Frequency of unique combinations of donor sequences at each position of the exon-intron junction, spanning 6, 7, 8, 10 (**a**) or 9 (**b**) consecutive nucleotides. Black bars denote windows overlapping the  $E^4-D^8$  donor sequence window. Four sliding windows of 9 nt spanning the annotated-donor (coloured black), spanning 12 nt from the fourth-to-last exonic base ( $E^4$ ; E = exon) to the eighth intronic base ( $D^8$ ; D = donor), were used for DF calculation. These windows were chosen due to the jump in sequence diversity seen with windows upstream of/including E-5, and downstream of/including E+9. **c)** Donor Frequency is calculated as the median frequency across each 9nt window, converted to a cumulative percentile distribution. DF measures donor strength by how many annotated-donors in the human genome have the exact same sequence. In this example, a median DF raw value of 179 lies at the 31st percentile of a cumulative frequency distribution.

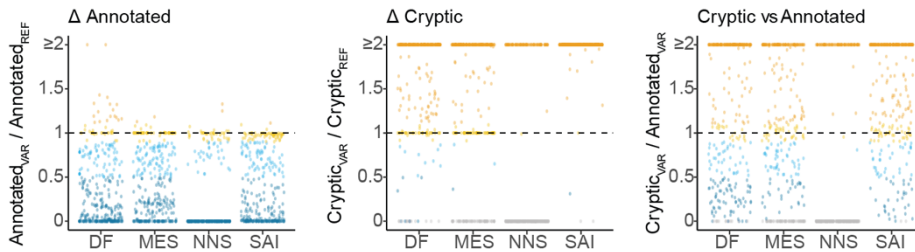
a AM-variants



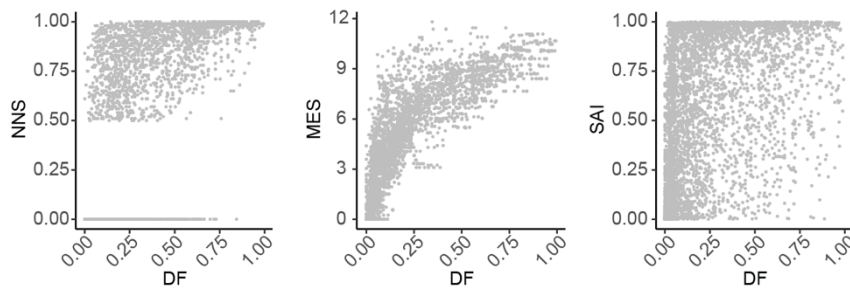
b CM-variants



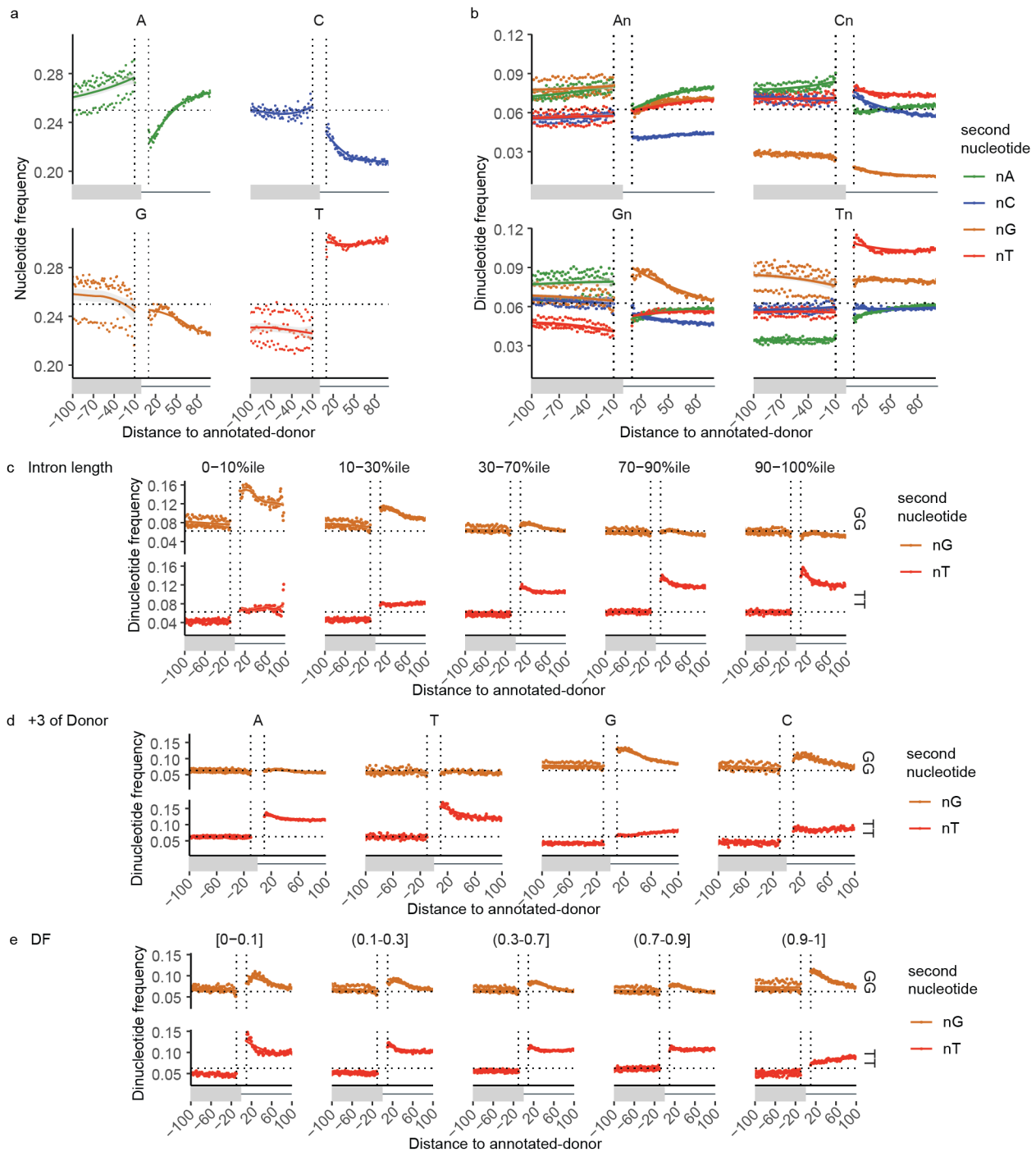
c AM/CM-variants



d



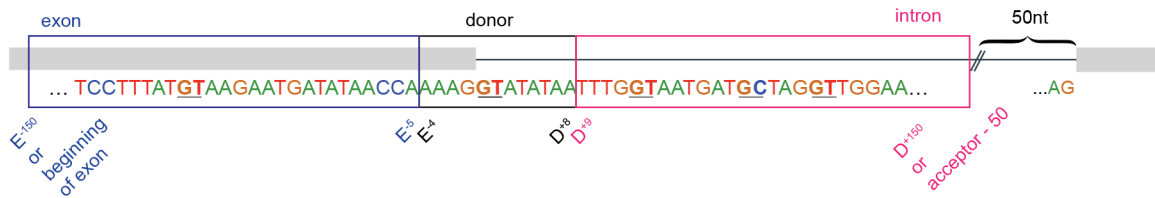
**Supplementary Fig. 2 Algorithmic prediction of cryptic-activation. a-c)** DF (Donor Frequency), MES (MaxEntScan), NNS (NNSplice) and SAI (SpliceAI) scores for **(a)** AM-variants **(b)** CM-variants and **(c)** AM/CM-variants. Colour coding is explained in the Figure key. When a donor strength score of 0 is returned, we set it to  $1 \times 10^{-6}$  to allow for the  $\Delta$  calculations (VAR/REF; VAR = variant; REF = reference). **d)** Comparison of NNS, MES and SAI scores with DF for all cryptic-donors (scores for VAR sequence) in our Cryptic-Donor database. DF shows strongest correlation with MES. NNS does not recognise a subset of human donors to offer a strength prediction.



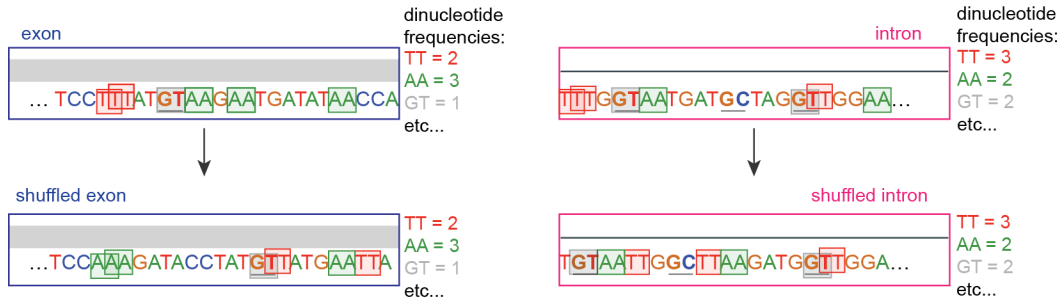
**Supplementary Fig. 3 G- and T- dinucleotide repeats show distinct patterns of enrichment in different introns.**

**a-b)** frequencies of each **(a)** nucleotide and **(b)** dinucleotide at each position surrounding annotated-donors. Vertical dotted lines denote boundaries at -10 and +10 where calculations start (i.e. excluding the conserved extended splice-site region). Horizontal dotted lines denote a random frequency of **a)** 1/4 for single nucleotides and **b)** 1/16 for dinucleotides. Lines show LOESS smoothing (locally weighted smoothing i.e. trendlines) with confidence bands in grey. In **(b)** Panels are according to the first nucleotide and colours are according to the second nucleotide in the dinucleotide. Note enrichment of G- and T- dinucleotides in the first 50 nt of the intron. **c-e)** frequencies of dinucleotides GG, and TT at each position surrounding annotated-donors. Vertical dotted lines denote boundaries at -10 and +10 where calculations start, horizontal dotted line denotes a random frequency of 1/16. Lines show LOESS smoothing (locally weighted smoothing i.e. trendlines) with grey confidence bands. **a)** G-repeats are enriched in the shortest human introns whereas T-repeats are enriched in longer introns. Length bins: < 149 nt (< 10<sup>th</sup> percentile), 149-627 nt (10 - 30<sup>th</sup> percentile), 628-3010 nt (30 - 70<sup>th</sup> percentile), 3011-9270 nt (70 - 90<sup>th</sup> percentile), > 9270 nt (> 90<sup>th</sup> percentile). **b)** Annotated donors with D<sup>+3</sup> G (or C) are enriched in G-dinucleotides whereas donors with D<sup>+3</sup> A (or T) are enriched in T-dinucleotides. **c)** Rare donors (low Donor Frequency (DF)) show greater enrichment for T dinucleotide repeats compared with common donors (high DF).

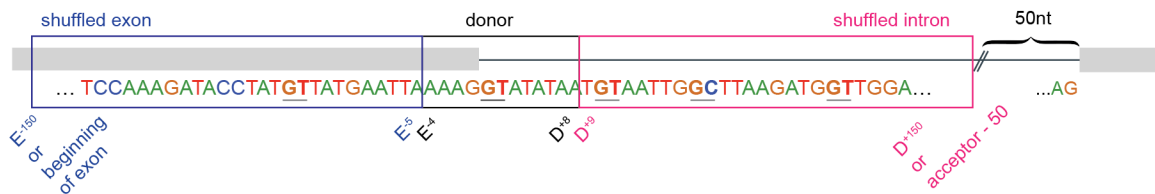
1. Partition sequences, into exonic, intronic, and donor



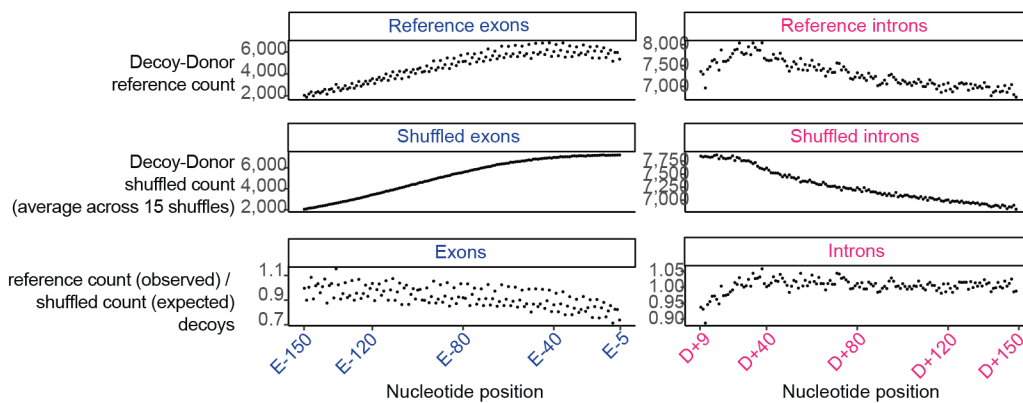
2. Shuffle exons and introns separately, maintaining dinucleotide frequencies



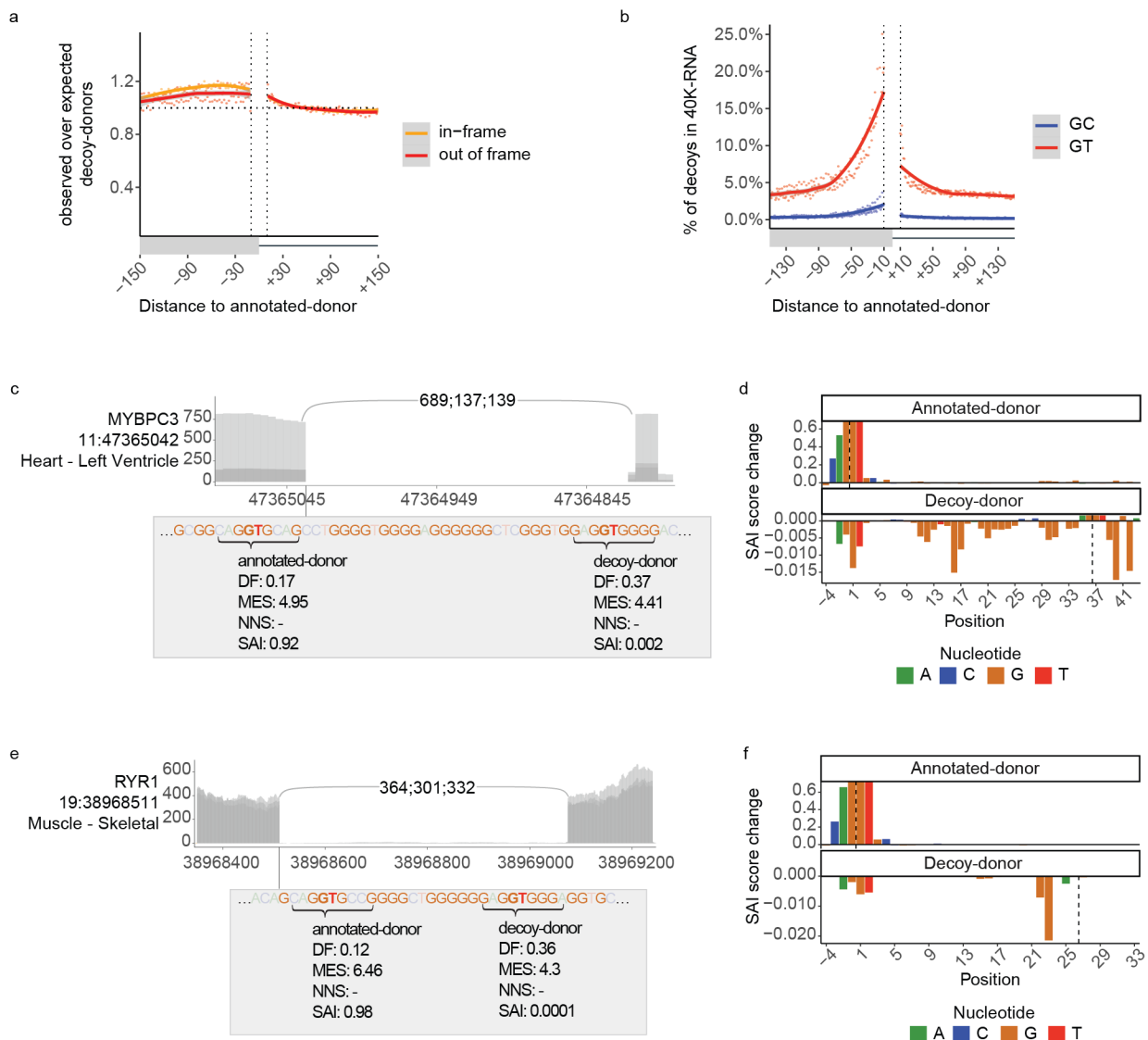
3. create set of shuffled exon-intron junction sequences



4. Tally decoy-donors at each nucleotide in reference & shuffled sequence sets

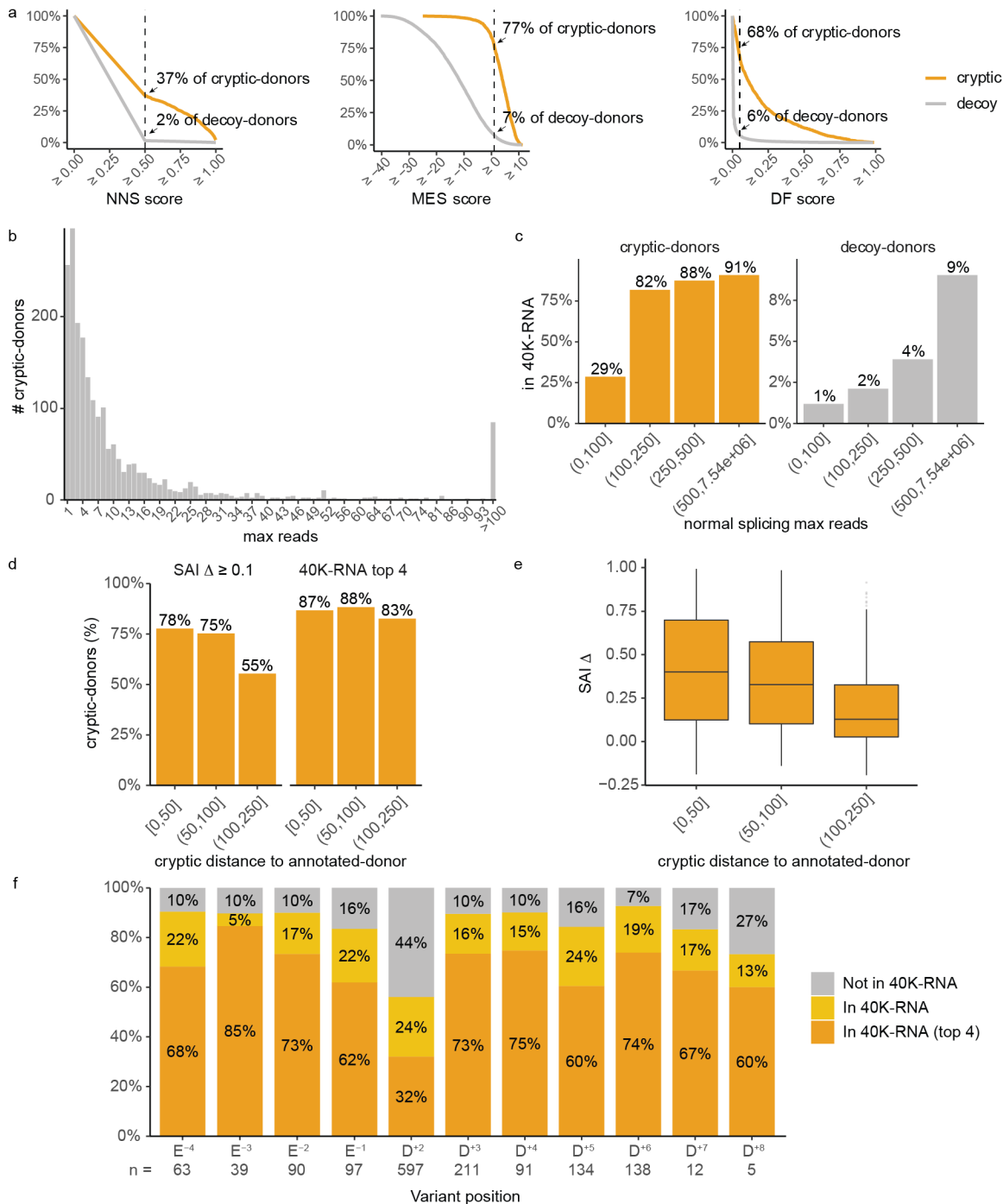


**Supplementary Fig. 4 Schematic representation of how decoy depletion is calculated.** Sequences in blue 'exon' and pink 'intron' boxes are shuffled separately (maintaining dinucleotide frequencies) and the number of actual decoy-donors at each position is divided by the number in the shuffled sequence set.



**Supplementary Fig. 5 40K-RNA provides insight into natural use of decoy-donors by the spliceosome.** **a)** The ratio of observed to expected ‘GC’ decoy-donors at each position +/-150 nt of annotated-donors. Yellow: In-frame exonic decoy-donors. Red: out-of-frame decoy-donors. **b)** The percent of decoys present in 40K-RNA +/-150 nt of annotated-donors. Red: ‘GT’ decoy-donors closer to the annotated-donor are more likely to be present in 40K-RNA. Blue: ‘GC’ decoy-donors are rarely seen in 40K-RNA. At each position the number of decoy-donors present in 40K-RNA is divided by the total number of naturally occurring decoy-donors at that position. Lines show LOESS smoothing (locally weighted smoothing i.e. trendlines) with grey confidence bands. **c-f)** two examples of decoy-donors overlapping G-repeats that outcompete the annotated-donor according to Donor Frequency (DF) and MaxEntScan (MES), but are not present in 40K-RNA, and which SpliceAI (SAI) correctly identifies as non-functional donors. NNS = NNSplice **c,e)** Shows overlays of 3 GTEx RNA-seq samples from the tissues with the highest TPM for that gene. The numbers (e.g., 689;137;139) denote the detected reads in each respective sample for that splice-junction. Algorithmic strength scores for annotated- and decoy-donors are boxed. **d,f)** Result of SAI in silico mutagenesis showing the bases contributing to predicted strength of the annotated-donor (top) and decoy-donor (bottom). ‘SAI score change’ denotes the decrease (if positive) or increase (if negative) on the predicted strength of the donor when that nucleotide is mutated (see methods). Black dashed line indicates the position of the annotated- or decoy-donor and position on the x-axis denotes the position of the nucleotide relative to the annotated-donor.





**Supplementary Fig. 6 Effectiveness of 40K-RNA and SAI for prediction of cryptic-donor selection. a)** Sensitivity (orange) and specificity (grey) of NNSplice (NNS) using a cut-off of 0.5, MaxEntScan (MES) using a cut-off of 1 and Donor Frequency (DF) using a cut-off of 0.05 to predict cryptic-donor activation in AM-variants. **b)** The maximum number of reads detected in any one RNA-seq sample across 40K-RNA for each cryptic-donor activated by an AM-variant. **c)** Only 29% of cryptic-donors for target genes with < 100 max reads corresponding to normal splicing are present in 40K-RNA, rising sharply to > 82% sensitivity for transcripts with more than 100 max reads corresponding to normal splicing. **d)** Percent of AM-variant cryptic-donors with SAI  $\Delta$  scores greater than or equal to 0.1 (left) or in the 40K-RNA top 4 (right) in different bins according to cryptic distance to the annotated-donor. SpliceAI's sensitivity for AM-variants drops to 55% for cryptic-donors more than 100 nt from the annotated-donor **e)** SAI  $\Delta$  scores for AM cryptic-donors (n = 2,348) relative to their distance from the annotated-donor. Internal lines denote the median value, and the lower and upper limits of the boxes represent 25<sup>th</sup> and 75<sup>th</sup> percentiles. The whiskers extend to the smallest and largest values no further than 1.5 x inter-quartile range (IQR). **f)** The percent of CM- and AM/CM-variant cryptics detected in 40K-RNA, according to the position of the SNV within the extended splice-site region of the activated cryptic-donor.



SpliceVault predicts the precise nature of variant-associated mis-splicing.

## **SpliceVault predicts the precise nature of variant-associated mis-splicing.**

### **3.1 Overview**

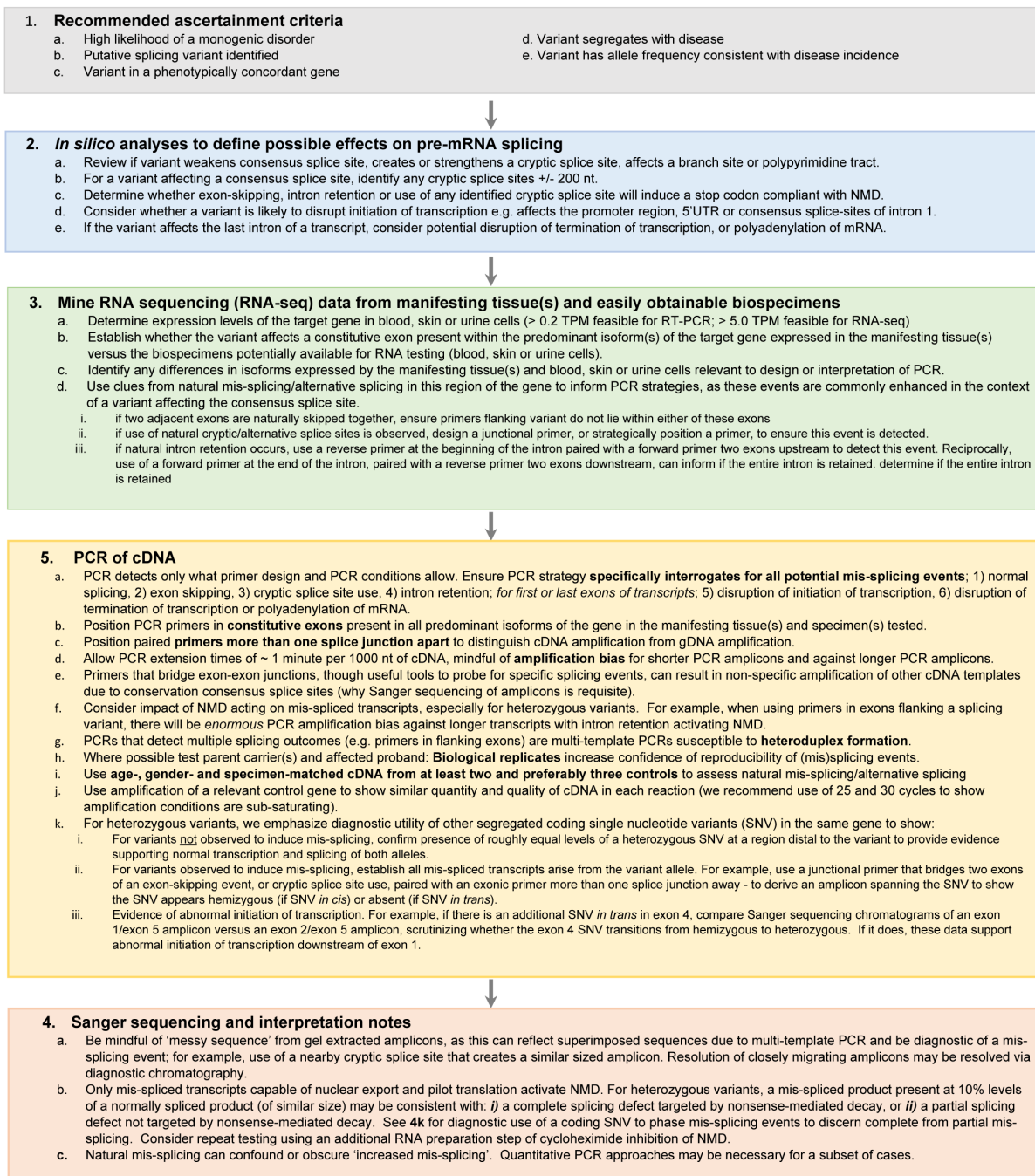
In Chapter 3, we extend the method developed in Chapter 2 for predicting cryptic-donor activation, to predict exon skipping events, as well as mis-splicing events at the acceptor splice site. We greatly expand the breadth of sequencing data used in predictions from 40,233 (40K-RNA) samples to 335,663 (300K-RNA), and house them both in a resource we term ‘SpliceVault’.

To validate 300K-RNA we used a cohort of experimentally verified splice-altering variants, with RNA diagnostics performed by our lab, most of which were initially published in a 2022 paper from our lab, led by Adam Bournazos<sup>89</sup>, which was outside the scope of this thesis. To aid in the readability of my thesis I’ve re-presented Figure S5 from Bournazos et al. 2022<sup>89</sup>, and the source of variants and strategy of RNA diagnostic studies in producing this set of splice-altering variants is summarised here:

74 families were recruited from local area health districts across Australia and New Zealand according to if they had splicing variants with high clinical suspicion of causality, using the following criteria: (1) a likely monogenic Mendelian disorder, (2) variant allele frequency in public databases consistent with disease incidence, (3) putative splicing variant in a clinician-defined, phenotypically concordant gene, and (4) preferably segregating with disease). RNA was derived either from clinically accessible tissues (blood, skin fibroblasts or urothelial cells) or biopsy specimens if available.

The study established comprehensive procedural guidelines for RNA diagnostics via RT-PCR (Figure 3-1): In essence, patient cDNA and age-, sex- and specimen-matched cDNA from at least two controls were interrogated for normal splicing as well as any potential exon skipping, intron retention, and cryptic splice site usage. Potential cryptic splice sites were identified using *in silico* algorithms.

## SpliceVault predicts the precise nature of variant-associated mis-splicing.



**Figure 3-1 Re-presented Figure S5 from Bournazos et al. 2022<sup>89</sup>.** Procedural guidelines for RNA Diagnostics via RT-PCR and Sanger sequencing endorsed by Clinical Variant Curators (genetic pathologists and qualified diagnostic scientists).

Using this high-quality set of variants with RNA diagnostic data with which we were intimately familiar, as well as two additional variant cohorts curated from literature, we assessed the sensitivity and positive predictive value of 300K-RNA predictions of cryptic activation and exon skipping events, across donor and acceptor variants. Additionally, we applied custom interpretive rules to SpliceAI output allowing it to offer predictions of exon skipping and intron retention events in addition to cryptic activation, to allow comparison

SpliceVault predicts the precise nature of variant-associated mis-splicing.

with 300K-RNA (no other algorithm currently predicts both exon skipping and cryptic activation).

300K-RNA outperformed SpliceAI on average across three independent cohorts of variants, with a mean sensitivity of 92%. Additionally, RNA reanalysis of several variants where 300K-RNA Top-4 events were not detected revealed that they had been missed upon initial analysis, due to the difficulties of RNA studies. This highlights the danger of false negatives in RNA diagnostics and potential impact to variant interpretation, examined in my introduction.

We found that despite the heterogeneity of 300K-RNA, top-ranked events were highly concordant across its component tissue-types and data sources. We propose new draft recommendations for the application of PVS1 to essential splice site variants, recommending consideration of 300K-RNA Top-4 events and intron retention.

This chapter was accepted at *Nature Genetics* as an analysis article for which I am joint first author. My contributions were creation of the 300K-RNA database and cross-referencing predictions with experimental data generated by Adam Bournazos. I prepared Figure 1C, Figures 2-4, Extended Data Figures 1-6, and write and edited the manuscript along with co-authors. I created the accompanying 300K-RNA web portal, with assistance from Himanshu Joshi with database management and tissue-specific capabilities.

**Dawes R**, Bournazos AM, Bryen SJ et al. SpliceVault predicts the precise nature of variant-associated mis-splicing. *Nat Genet (NG-AN59054R3)*. (2022).

SpliceVault predicts the precise nature of variant-associated mis-splicing.

**SpliceVault predicts the precise nature of variant-associated mis-splicing.**

Ruebena Dawes<sup>1,2,3,a</sup>, Adam M. Bournazos<sup>1,2,3a</sup>, Samantha J. Bryen<sup>1,2,3</sup>, Shobhana Bommireddipalli<sup>1,3</sup>, Rhett G. Marchant<sup>1,2,3</sup>, Himanshu Joshi<sup>1,3,b</sup> and Sandra T. Cooper<sup>1,2,3,b\*</sup>

<sup>1</sup>Kids Neuroscience Centre, Kids Research, Children's Hospital at Westmead, Sydney, NSW 2145, Australia

<sup>2</sup>Discipline of Child and Adolescent Health, Faculty of Health and Medicine, University of Sydney, Sydney, NSW 2006, Australia

<sup>3</sup>The Children's Medical Research Institute, 214 Hawkesbury Road, Westmead NSW 2145, Sydney, Australia

a Ruebena Dawes and Adam Bournazos contributed equally as joint first authors

b Himanshu Joshi and Sandra Cooper contributed equally as joint senior authors.

**Corresponding author:**

Professor Sandra Cooper

Joint Head, Scientific Director, Kids Neuroscience Centre, The Children's Hospital at Westmead, Locked Bag 4001, Sydney, NSW 2145, Australia.

Discipline of Child and Adolescent Health, Faculty of Health and Medicine, University of Sydney, NSW 2006, Australia.

Telephone: (+61) (02) 9845 1455

E-mail: [sandra.cooper@sydney.edu.au](mailto:sandra.cooper@sydney.edu.au)

SpliceVault predicts the precise nature of variant-associated mis-splicing.

### **Abstract**

Clinical interpretation of splicing variants depends critically upon the nature of variant-associated mis-splicing and consequence(s) for the encoded gene product. Arrestingly, ranking the four most common unannotated splicing events across 335,663 reference RNA-sequencing samples (300K-RNA Top-4), identifies the nature of variant-associated mis-splicing with remarkable prescience. 300K-RNA Top-4 correctly identifies 96% of exon-skipping events and 86% of cryptic splice-sites induced by 88 variants across 74 genes and 140 affected individuals or heterozygotes subject to RNA Diagnostics. 300K-RNA shows higher sensitivity and positive predictive value than SpliceAI in predicting exon-skipping and cryptic-activation events. Importantly, RNA re-analyses showed we had missed 300K-RNA Top-4 events for several clinical cases tested prior to 300K-RNA. In conclusion, 300K-RNA provides an evidence-based method that predicts with high sensitivity the nature of variant-associated mis-splicing. The SpliceVault web portal allows users easy access to 300K-RNA, to augment both pathology consideration of PVS1 and RNA diagnostic investigations.

SpliceVault predicts the precise nature of variant-associated mis-splicing.

## Introduction

Genetic variants that induce mis-splicing of precursor messenger RNA (pre-mRNA) are a common cause of inherited disorders<sup>1,2</sup>. Interpreting pathogenicity of a splicing variant depends on the nature of detected mis-splicing, relative to the known pathogenetic mechanism(s) of disease for that gene and disorder (i.e., loss-of-function/gain-of-function)<sup>3-8</sup>.

Variants impacting essential splice-sites, the almost invariant GT-AG flanking each intron, are virtually guaranteed to induce mis-splicing. Due to triplet codons, mis-splicing of pre-mRNA commonly induces a frameshift or encodes a premature termination codon (PTC), supporting rationale for consideration of essential splice-site variants under the PVS1 Null Variant (Very Strong evidence level) criterion of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP) guidelines<sup>9</sup>. In 2018, revised PVS1 guidelines recommended application of the PVS1 code for essential splice-site variants, at varying strengths, based upon theoretical consideration of consequences from exon-skipping, intron retention and use of any cryptic splice-site within 20 nucleotides (nt)<sup>10</sup>. While only ~20% of variant-activated cryptic donors are within 20 nt<sup>11</sup>, consideration of a larger window is unfeasible in diagnostic genetic pathology, due to the large number of potential cryptic splice-sites present in the genome. In addition, factors that induce multi-exon skipping (or retention of multiple introns) associated with some splice-altering variants are unknown.

For RNA Diagnostic testing performed in our laboratory<sup>3</sup>, we routinely interrogate RNA sequencing (RNA-Seq) data from control specimens (in house or from GTEx<sup>12</sup> or ENCODE<sup>13</sup>) to assess patterns of alternative splicing of the target gene between the manifesting tissues and clinically accessible specimens. We observed that the predominant, variant-associated mis-spliced transcript(s) identified in specimens from affected individuals and heterozygotes were often observed as rare, stochastic splice-junctions in control RNA-Seq data. Brandão and colleagues detailed a similar finding, with dominant variant-induced mis-spliced *BRCA1* or *BRCA2* transcripts often seen as rare events in disease controls<sup>14</sup>. Additionally, Kremer *et al.*, found that splicing ‘noise’ often forecast the location of variant-activated pseudoexons and reasoned that a population-based RNA-seq compendium could aid in variant prioritisation<sup>15</sup>.

SpliceVault predicts the precise nature of variant-associated mis-splicing.

In Dawes et al., 2022<sup>11</sup>, we analysed 5,145 variants activating cryptic splice-sites and established that 87% of activated cryptic splice-sites are those detected as rare, unannotated splice junctions in 40,233 RNA-Seq samples from GTEx<sup>12</sup> and Intropolis<sup>15</sup> (40K-RNA database<sup>11</sup>). The key insight that cryptic donors activated by genetic variants are also seen as rare events in population-based RNA-Seq data, led us to explore whether other forms of variant-associated mis-splicing may be predicted by quantifying the relative prevalence of stochastic, natural, unannotated splicing events (referred to hereafter as mis-splicing events).

We therefore created 300K-RNA, an expanded resource detailing the most common unannotated splicing events local to each exon-intron junction of Ensembl<sup>16</sup> and RefSeq<sup>17</sup> transcripts, based on splice-junctions detected across 335,663 publicly available RNA-Seq samples from Genotype Tissue Expression dataset (GTEx)<sup>12</sup> and Sequence Read Archive (SRA)<sup>18</sup>, processed in the recount3 project<sup>19</sup> (300K-RNA). 300K-RNA is updated to the GRCh38 genome assembly and is hosted in a web resource called SpliceVault, together with 40K-RNA (GRCh37)<sup>11</sup>. Unannotated splice-junctions in 300K-RNA constitute evidence that a splicing event is biophysically possible and possesses the requisite constellation of features for the splicing reactions to be executed. Our central hypothesis is that a genetic variant impeding or precluding spliceosomal use of an annotated splice-site is most likely to enhance or activate stochastic mis-splicing events that occur naturally.

Herein we demonstrate that 300K-RNA Top-4 ranked events correctly identifies 96% of exon-skipping events (including multi-exon skipping) and 86% of activated cryptic splice-sites induced by 88 variants in 74 genes for 140 affected individuals or heterozygotes subject to RNA Diagnostics. We additionally provide a comparison with SpliceAI<sup>20</sup>, applying custom interpretive rules to SpliceAI  $\Delta$ -scores +/- 5000 nt of variants to infer predictions of exon skipping, intron retention and cryptic splice-site activation.

SpliceVault predicts the precise nature of variant-associated mis-splicing.

## Results

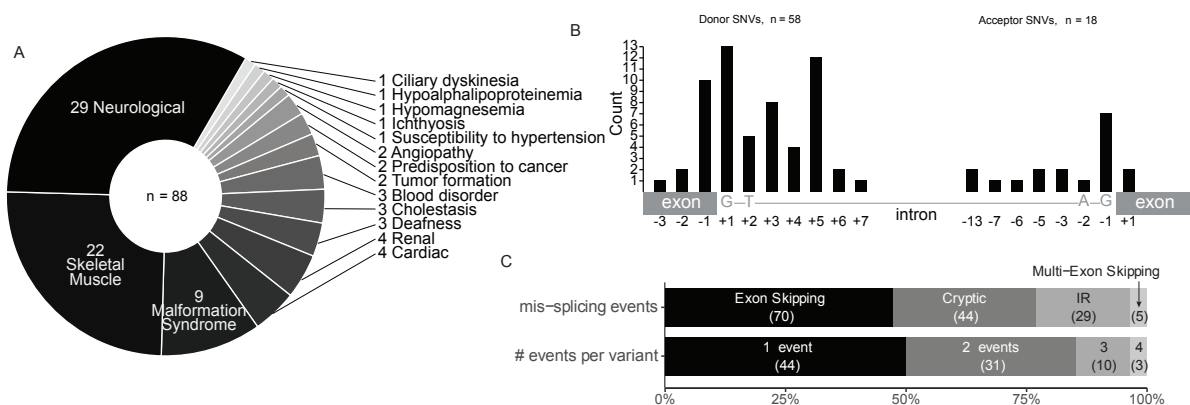
### *A set of experimentally verified splice-altering variants*

We performed retrospective analysis of 88 splice-site variants across 74 genes that are confirmed by RNA diagnostics<sup>3</sup> to disrupt pre-mRNA splicing (Supplementary Table 1).

*Inclusion Criteria:* Variants affecting an annotated splice-site and demonstrated to activate exon-skipping or cryptic splice-site use. *Exclusion Criteria:* Variants creating or modifying a cryptic splice-site (inappropriate for our method<sup>11</sup>).

Reverse transcription PCR (RT-PCR) and/or RNA-Seq were performed on RNA isolated from clinically accessible specimens from 140 affected individuals or heterozygotes with diverse Mendelian conditions<sup>3,21-24</sup> (Figure 1A and Supplementary Table 1). The majority of probands had neurological ( $n=29$ ), skeletal muscle ( $n=22$ ), or malformation syndrome ( $n=9$ ) phenotypes. 32% of variants affect the essential GT ( $n=19$ ) or AG ( $n=9$ ) splice-sites and 68% affect the extended donor or acceptor splice-site regions. The dataset included 76 single nucleotide variants (SNVs), 4 insertions, 6 deletions and 2 deletion-insertion variants (Figure 1B).

Half of the variants (44/88) induced two or more mis-splicing events (Figure 1C). Variants most frequently caused skipping of a single exon (70/148 total events, 47%), followed by cryptic activation (44/148 events, 30%) and intron retention (29/148 events, 20%), and rarely caused multi-exon skipping (5/148 events, 3%) (Figure 1C).



**Figure 1. Variant cohort details** **A**) Phenotypes associated with 88 experimentally-verified clinical splicing variants and **B**) Position of the 76/88 variants that are single nucleotide variants (SNVs) relative to the essential splice-sites. **C**) Nature of 148 unannotated splicing events (mis-splicing) induced by the 88 variants. IR = Intron Retention.



SpliceVault predicts the precise nature of variant-associated mis-splicing.

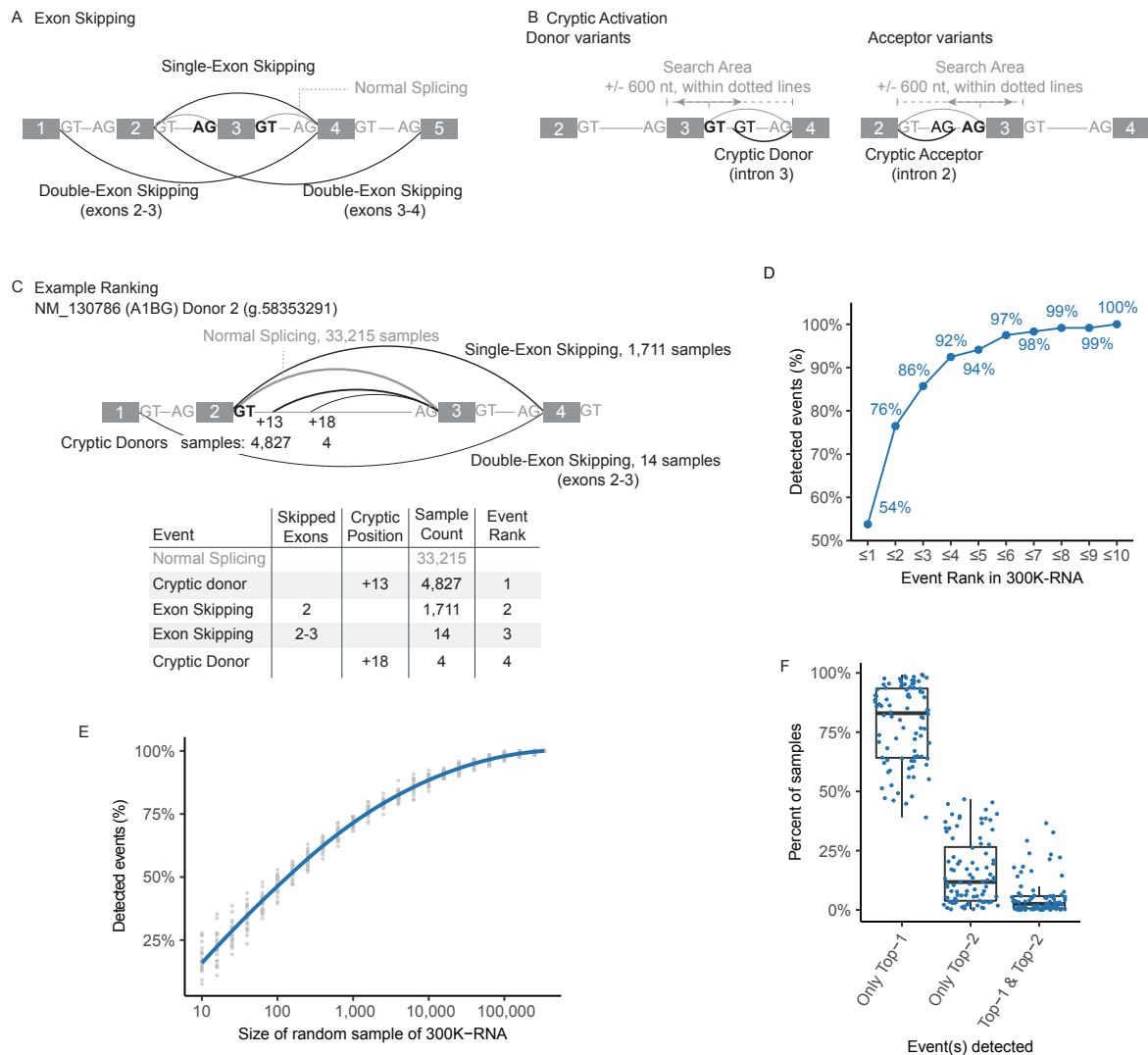
### *Unannotated splicing events in 300K-RNA*

The 300K-RNA database describes natural variation in splicing among 335,663 publicly available RNA-Seq samples from GTEx<sup>12</sup> and SRA<sup>18</sup>, collected in the recount3 resource<sup>19</sup> (see Methods). For each donor and acceptor in Ensembl<sup>16</sup> and RefSeq<sup>17</sup> transcripts, we collate all unannotated, stochastic splicing events surrounding that splice-site (Figure 2A-C), detected in RNA-Seq samples processed in a unified pipeline in the recount3 resource<sup>19</sup>. Wilks *et al.*, use splicing-aware alignment in an annotation-agnostic fashion, preventing bias against detection of unannotated events<sup>19</sup>.

These splice-junctions provide experimental evidence for an executed splicing reaction using: **a)** a paired donor and acceptor from different introns, reflecting skipping of one or more consecutive exons normally present in that transcript (Figure 2A, *exon skipping*); or **b)** an annotated donor or acceptor, paired with an unannotated acceptor or donor, respectively, indicating cryptic splicing (Figure 2B, *cryptic splicing*). Mis-splicing events detected at each splice-site in 300K-RNA are ranked by the number of samples in which at least 1 splice-junction read was detected (Figure 2C). Highest sensitivity and PPV for 300K-RNA predictions were obtained using the Top-4 ranked events as a prediction of the nature of mis-splicing - applying a filter of a maximum of two exons skipped and cryptic splice-sites within 600 nt (Extended Data Figure 1; filter demarked hereafter by an asterisk). Using this filter, 300K-RNA Top-10\* events identified all 119 exon-skipping and cryptic splicing events induced by the 88 variants (Figure 2D) with 64/119 (54%) the Top-ranked\* event for that splice-site (Figure 2D, Extended Data Figure 2).

Figure 2E shows the importance of sequencing breadth in 300K-RNA for detection of all 119 true positive exon-skipping and cryptic splicing events. Taking random subsets within the 335,663 source specimens shows sensitivity only begins to maximise with ~100,000 samples. Deeper scrutiny of the 119 true positive events shows, on average, each event is detected as a single splice-junction read in 78% samples with this event – underpinning why all single read events are catalogued in 300K-RNA. Figure 2F reinforces the stochastic nature of these mis-splicing events, showing the Top-1 and Top-2 events around the splice sites affected by our 88 variants typically occur in mutually exclusive specimens – with both events seen, on average, in only 5% of samples where either event was seen.

## SpliceVault predicts the precise nature of variant-associated mis-splicing.



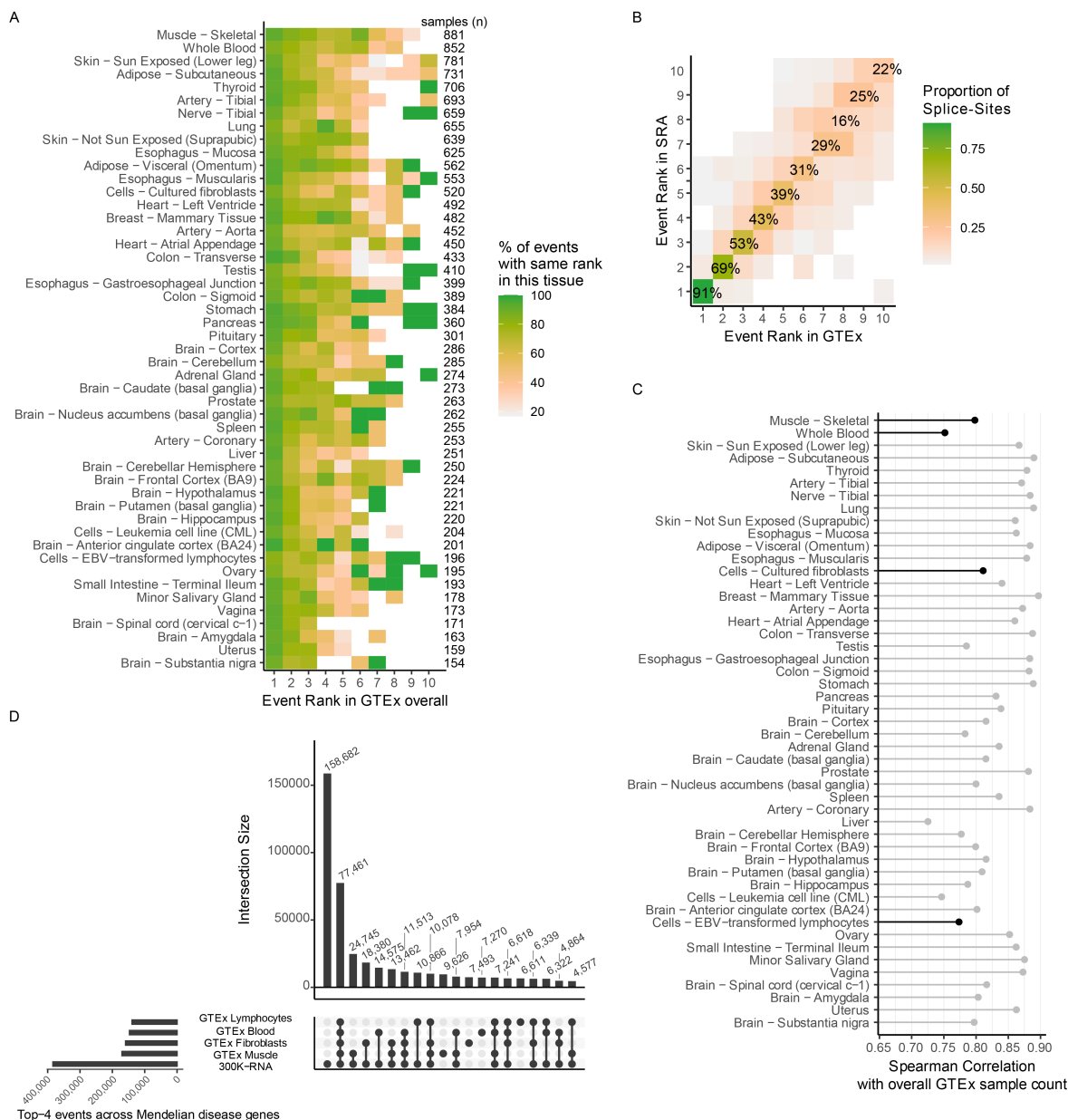
**Figure 2. Unannotated splicing events seen in 300K-RNA** **A)** Exon Skipping events are evidenced by split-reads spanning non-consecutive exons within the transcript. Splice-sites (GT/AG motifs) shown bold and in black are those for which events are being ranked. **B)** Cryptic Activation events are evidenced by split-reads spanning: *i)* an annotated acceptor and an unannotated donor or *ii)* an annotated donor and an unannotated acceptor. **C)** Example showing the Top-4\* events for NM\_130786 (A1BG) Donor 2 (g.58353291). Exon/Intron lengths are not drawn to scale. Arc thickness corresponds to event rank. **D)** 119/119 (100%) exon-skipping and cryptic activation events detected across 88 variants are present in 300K-RNA, and 92% are in the Top-4\* events for their respective splice-site. **E)** Percent of the 119 true positive events detected within random subsets of the 335,663 source specimens in 300K-RNA. Grey dots show proportion across 20 random samples, blue line shows mean proportions with LOESS smoothing. **F)** Top-1\* and Top-2\* events around the splice-sites affected by our 88 variants typically occur in mutually exclusive specimens – with both events seen, on average, in only 5% of total samples where either event was detected. Internal lines of boxplot denote the median value, and the lower and upper limits of the boxes representing 25<sup>th</sup> and 75<sup>th</sup> percentiles. Whiskers extend to the largest and smallest values at most 1.5IQR. \* = skipping one or two exons and cryptic activation within 600 nt of the annotated splice-site.

SpliceVault predicts the precise nature of variant-associated mis-splicing.

#### *Concordance of 300K-RNA events between tissues and datasets*

The ranking of the most common, natural mis-splicing events detected around each splice-site in 300K-RNA is highly concordant between each tissue within GTEx and between GTEx and SRA (Figures 3A-B, our 88 variants: Extended Data Figure 3A-B, 98,810 annotated splice-sites in the Mendeliome). This tells us that the spliceosome reproducibly makes the same mistakes across a diverse repertoire of tissues and cell lines. Additionally, sample counts of all 300K-RNA events were strongly correlated between different GTEx tissues (Figure 3C) and between GTEx and SRA (Extended Data Figure 3C,  $R = 0.91$  for our 88 variants,  $R = 0.84$  for the Mendeliome). However only 20% of 300K-RNA Top-4\* events are present in all four clinically accessible tissues in GTEx (Figure 3D, blood, fibroblasts, EBV-LCL, muscle) – reflecting low expression of many Mendelian genes in these tissues. Sequencing breadth in GTEx (i.e., the number of specimens available for that tissue sub-type) increases the number of unannotated splice-sites detected ( $R = 0.91$ ,  $p < 2.2e^{-16}$ ) (Extended Data Figure 3D).

Importantly, GTEx Muscle Top-4\* did not provide improved sensitivity over 300K-RNA Top-4\* for 19/88 of our variants associated with muscle disorders subject to RNA Diagnostics on muscle samples (Extended Data Figure 3E). Therefore, we recommend use of 300K-RNA Top-4\* as prediction of the probable nature of variant associated mis-splicing until we have great enough breadth and/or depth of RNA-Seq data to evaluate a tissue-specific approach. We advise caution for genes with known tissue-specific or developmental alternative splicing where RNA-Seq from the relevant tissue is not represented, or poorly represented, within the 300K-RNA data sources.



**Figure 3. 300K-RNA event rankings across tissues and data-sources. A)** Heatmap showing the proportion of mis-splicing events\* with the same event rank in each GTEx tissue subtype, as compared to all GTEx tissue subtypes combined - for the 88 splice-sites affected by our cohort of variants. Only tissues with  $\geq 100$  GTEx samples are shown. The Top-1\* event in individual tissues is concordant with the Top-1\* event in ‘all GTEx tissues’ for  $\geq 80\%$  of splice-sites. **B)** Concordance of top-ranked mis-splicing events\* in GTEx versus SRA. The Top-1\* event in GTEx is the Top-1\* event in SRA for 80/88 (91%) splice-sites. **C)** Spearman correlation of all mis-splicing events\* across 98,810 annotated splice-sites in the Mendeliome (see methods) in each GTEx tissue subtype versus GTEx overall. Only tissues with  $\geq 100$  GTEx samples are shown. *Black*: clinically accessible tissues. **D)** Upset plot<sup>27</sup> showing 300K-RNA Top-4\* across the Mendeliome (4 events per splice-site,  $n = 383677$  in total), versus Top-4\* specific to four clinically accessible tissues in GTEx. 77461/383677 (20%) of all 300K-RNA Top-4\*-across the Mendeliome are captured as Top-4\* events among all four clinically accessible tissues (blood, fibroblasts, EBV-LCL, muscle). \* = skipping one or two exons and cryptic activation within 600 nt of the annotated splice-site.

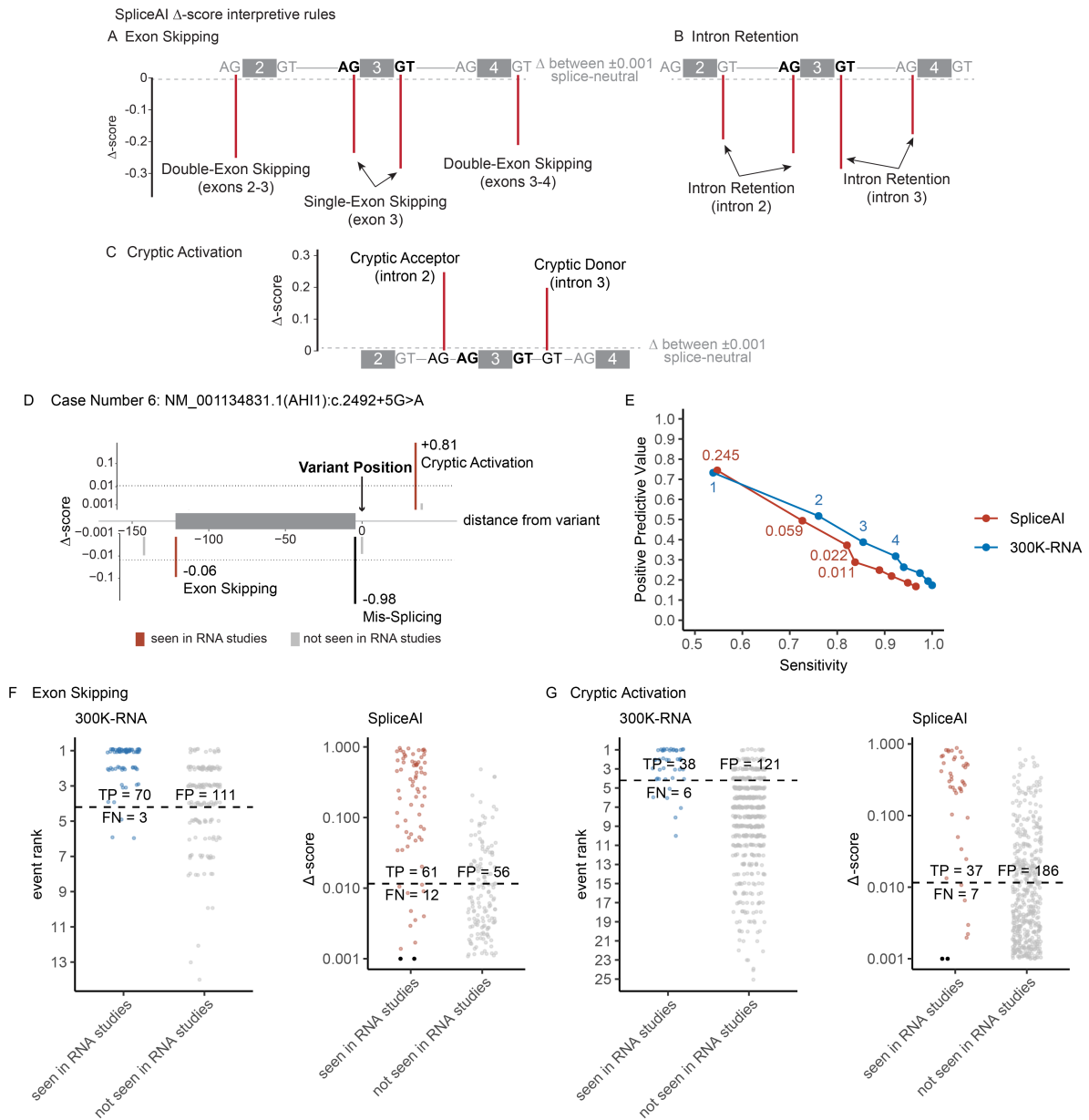
SpliceVault predicts the precise nature of variant-associated mis-splicing.

#### *SpliceAI predictions using custom interpretive rules*

SpliceAI author's <sup>20</sup> high sensitivity threshold of  $\Delta \geq 0.20$  predicted mis-splicing for 76/86 (88%) variants and for 63/86 (73%) variants with the high specificity threshold of  $\Delta \geq 0.50$ . Preliminary investigations using the default SpliceAI window of +/- 50 nt and high sensitivity threshold of  $\Delta \geq 0.20$  identified 13/44 variant-activated cryptic splice-sites (only 19/44 activated cryptics lie within 50 nt) and 3/73 exon skipping events. Therefore, we adapted SpliceAI to offer a fuller prediction of the nature of mis-splicing, by assessing all  $\Delta$ -scores generated by inputting pre-mRNA sequence +/- 5000 nt of the variant (Figure 4A-C).

Two insertion-deletion variants could not be assessed by SpliceAI (Supplementary Table 1, Case 50 and 66).  $\Delta$ -scores  $< 0.001$  were excluded as nominal predictions of neutral impact, and all scores above this threshold were retained for subsequent precision-recall analysis (see Methods). The remaining 2,836  $\Delta$ -scores returned for the 86 variants were interpreted according to the following rules:  $\Delta$ -loss scores at the annotated splice-site constituted a prediction of mis-splicing. *Exon Skipping* is inferred if both splice-sites flanking an *exon* have a  $\Delta$ -loss score above threshold (Figure 4A). *Double-Exon Skipping* is inferred if the relevant splice-site of the upstream or downstream intron also has a  $\Delta$ -loss score above threshold (Figure 4A). *Intron Retention* is inferred if both splice-sites flanking an *intron* have a  $\Delta$ -loss score above threshold (Figure 4B). *Cryptic Activation* is predicted by  $\Delta$ -gain score above threshold for any unannotated donor or acceptor within the bounds of the exon and intron flanking the variant splice-site (Figure 4C).  $\Delta$ -scores that do not fall into these categories are annotated as 'other' (see Methods). According to these rules, SpliceAI predicts at least one mis-splicing event for all 86 variants and up to 31 predictions for a single variant (Figure 4D, Extended Data Figure 4). 139/145 mis-splicing events elicited by the 86 variants lie within the maximum +/- 5000 nt SpliceAI window (44/44 cryptics, 68/68 single-exon skipping events, 4/5 double-exon skipping events and 7/12 intron retention events).

# SpliceVault predicts the precise nature of variant-associated mis-splicing.



**Figure 4. Comparison of 300K-RNA Top-4\* with SpliceAI.** **A-C)** Custom interpretive rules applied to SpliceAI  $\Delta$ -scores to predict the nature of mis-splicing. Heights of red lines denote example  $\Delta$ -scores that predict mis-splicing events according to our rules. **A)** *Single-Exon Skipping* is predicted if both splice-sites flanking the *exon* have a donor and acceptor loss  $\Delta$ -scores above threshold, and *Double-Exon Skipping* was inferred if the splice-site of the upstream or downstream intron also had donor loss or acceptor loss  $\Delta$ -score above threshold. **B)** *Intron Retention* was predicted if both splice-sites flanking an *intron* had donor loss and acceptor loss  $\Delta$ -score above threshold. **C)** *Cryptic Activation* was predicted by donor gain or acceptor gain  $\Delta$ -scores above threshold for any unannotated donor or acceptor. **D)** Example showing SpliceAI predictions of exon skipping and cryptic activation in Case Number 6. **E)** Sensitivity and PPV of 300K-RNA and SpliceAI for exon-skipping and cryptic activation prediction at different thresholds, for the 86/88 variants that can be scored by SpliceAI. Points on the 300K-RNA curve (blue) show metrics when using Top-1\*, 2\*, 3\*, 4\* etc events as prediction of the nature of mis-splicing. Points on the SpliceAI curve (red) show metrics at  $\Delta$ -scores that predict the same number of exon-skipping and cryptic-activation as 300K-RNA top-1\*, 2\*, 3\*, 4\* and so on. **F)** 300K-RNA and SpliceAI predictions of exon skipping (seen/not seen in RNA studies across 86 cases). **G)** 300K-RNA and SpliceAI predictions of cryptic splice-site activation (seen/not seen in RNA studies across 86 cases). Dotted lines indicate the threshold of Top-4\* and SpliceAI  $\Delta$ -score  $\geq 0.011$  identified in (E). *Black dots*: mis-splicing events seen in RNA studies but not meeting the  $\Delta$ -score threshold of 0.001. TP = True Positives, FN = False Negatives, FP = False Positives. \* = filtering events to those skipping one or two exons and cryptic activation within 600 nt of the annotated splice-site.

SpliceVault predicts the precise nature of variant-associated mis-splicing.

### *Comparing predictive performance of 300K-RNA with SpliceAI*

Precision-recall curves show the sensitivity and positive-predictive value (PPV) of 300K-RNA and SpliceAI to predict the 119 exon-skipping and cryptic-activation events induced by 86/88 variants assessable by both methods (Figure 4E). Top-4\* showed higher sensitivity (92%) than SpliceAI (84%) at a  $\Delta \geq 0.011$  threshold – this low threshold was selected because it identifies the same total number of mis-splicing events as Top-4\* (Figure 4E) to compare true/false positive and negative rates.

Top-4\* correctly identifies 96% (70/73) of detected exon skipping events while SpliceAI  $\Delta \geq 0.011$  predicts 85% (61/73) (Figure 4F). 86% (38/44) of activated cryptics are in Top-4\*, with SpliceAI  $\Delta \geq 0.011$  predicting 84% (37/44) (Figure 4G). Both methods show a low PPV: 39% (Top-4\*) and 52% (Splice AI  $\Delta \geq 0.011$ ) for exon-skipping (Figure 4F) and 24% (Top-4\*) and 17% (Splice AI  $\Delta \geq 0.011$ ) for cryptic-activation (Figure 4G). For intron retention events, which cannot be predicted using 300K-RNA, SpliceAI shows a sensitivity of 31% (9/29) and PPV of 36% (9/25) at the  $\Delta \geq 0.011$  threshold.

Sensitivity (S) and PPV of Top-4\* and SpliceAI (using  $\Delta \geq 0.011$  threshold) were similar across two additional sets of variants curated from literature; 58 variants tested in patient specimens (Top-4\* S = 91%, PPV = 26%; SpliceAI S = 94%, PPV = 33%) and 63 variants tested using midi-gene assays (Top-4\* S = 92%, PPV = 29%; SpliceAI S = 86%, PPV = 29%) (Extended Data Figure 5 and Supplementary Tables 2,3).

Analysis of features of Top-4\* events reveals that relative to false positive, true positives tend to be: identified in more samples (Extended Data Figure 6A); represented by more unannotated splicing reads (Extended Data Figure 6B-C); with a higher maximum ratio of the unannotated event relative to read-depth for annotated splicing in any one sample (Extended Data Figure 6E). However, there was no significant difference between true and false positives in the mean reads of annotated splicing in samples where events were detected (Extended Data Figure 6D) or in the mean ratio between unannotated and annotated splicing (Extended Data Figure 6F); showing these are trends rather than rules. Double exon-skipping is rarely activated by splice-site variants (Extended Data Figure 6G). 4/5 detected double exon skipping events were Top-1 (among 14 cases with double skipping ranked Top-1) and 1/5 Top-2; 3) show no difference in the length of the spliced-out region for exon-skipping

SpliceVault predicts the precise nature of variant-associated mis-splicing.

events (Extended Data Figure 6H) or cryptic distance (Extended Data Figure 6I). We identified one false positive that may be due to alignment issues in the short-read splice junction dataset: the first few nucleotides of two sequential exons are identical and split reads with only a few post-junction nts can map to exon-skipping or normal splicing.

We also emphasise that while a SpliceAI threshold of 0.011 was effective for this bespoke application to forecast the likely nature of any variant-elicited mis-splicing, we do not recommend use of a 0.011 delta score threshold as a prediction for mis-splicing generally, as our evidence from experimentally confirmed splice neutral variants (<sup>3</sup> and cases studied since) indicates this will yield a high false positive rate of > 50%.

#### *RNA re-analysis uncovers previously undetected Top-4\* events*

We noted many Top-4\* events not detected in our early RNA diagnostics cases (prior to 40K-RNA or 300K-RNA) involved double-exon skipping events (42%) or cryptic activation events further than 250 nt from the annotated splice-site (11%), which may have been missed on initial analysis. Prior to the development of 40K-RNA<sup>11</sup>, our laboratory practice included critical review of all cryptic splice-sites within 250 nt of the annotated donor<sup>3</sup>. Scrutiny of 24/88 variants where the top-1 event was not detected by RNA studies (Extended Data Figure 2) showed: 16/24 were detectable via the RNA Diagnostic strategy deployed, but not observed, due to: a) the event was not activated by the variant; b) low expression of the target gene, potentially limiting sensitivity; c) the variant simultaneously weakened an annotated splice-site and spatially overlapping cryptic splice-site comprising a Top-1\* event. 6/24 events were observed, though available RNA assay data did not confidently establish elevated levels relative to controls. 1/24 event was detected upon review of Sanger sequencing trace file. We performed RNA reanalysis for 1/24 case with an undetected Top-1\* event and three additional cases where multiple Top-4\* events were not detected (A024-*OPHN1* c.702+4A>G; A060-*GSDME* c.1183+5G>A; A014-*SPG11* c.2317-13C>G; A205-*EMD* c.266-3A>G)<sup>3</sup>. We identified or clarified variant-associated enhanced use of 1/4 multi-exon skipping events (Extended Data Figure 7A,B, *SPG11*, red), 4/4 cryptic splice-sites (Extended Data Figure 7C,D, *GSDME* and *EMD*, red), and 1 single-exon skipping event (Extended Data Figure 7D, *EMD*, red). Skipping of multiple exons associated with *SPG11* c.2317-13C>G was not detected initially by RT-PCR due to primer placement in exons too proximal to the splice variant, and undetected by RNA-Seq due to low read depth exacerbated by NMD. Activation of two cryptic donors and two cryptic acceptors associated



SpliceVault predicts the precise nature of variant-associated mis-splicing.

with *GSDME* c.1183+5G>A and *EMD* c.266-3A>G, respectively, were missed initially due to competition inherent with multi-template PCRs, heteroduplex formation and challenges resolving multi-trace chromatograms by Sanger sequencing.

#### *Half of essential GT-AG variants induce $\geq 1$ in-frame events*

49% (27/55) of essential splice-site (ES) variants across the three variant sets induced at least one in-frame event, with a similar proportion of 51% (45/88) for all variants in our cohort (Extended Data Figure 7E). When considering Top-4\* and intron retention as a prediction of variant-induced mis-splicing, the number of ES variants with  $\geq 1$  in-frame event is 80% (45/55) and 81% (71/88) for all variants in our cohort.

## **Discussion**

Clinical interpretation of splicing variants relies on predicting, or experimentally verifying, the nature of variant-induced mis-splicing to confirm variant impact on the encoded protein. This is of particular importance when applying the PVS1 (null variant) criterion to essential splice-site variants<sup>10</sup>. While the impact of exon skipping and intron retention on protein reading frame can be theorized, it has remained difficult to predict whether exon-skipping or cryptic splice-site activation will occur - and if a cryptic splice-site is activated, which one of the many potential sites present in the vicinity will be selected by the spliceosome.

Our empirical method of using 300K-RNA Top-4\* accurately predicts the nature of variant associated mis-splicing with 92% sensitivity for 88 variants across a broad range of genes and disorders, outperforming SpliceAI on average to correctly predict exon-skipping, double-exon skipping and cryptic splice site activation. We emphasize that Top-4\* cannot be used for variants creating or modifying the essential splice-site motif of a cryptic splice-site and recommend use of SpliceAI for this category of variant<sup>11</sup>. In addition, though intron retention cannot yet be quantified and ranked by 300K-RNA, Extended Data Figure 8 shows that intron retention induces a frameshift or encodes a premature termination codon in all three frames for at least 97% introns in the Mendeliome and is therefore consistent with null outcomes in most instances.

It is important to acknowledge the low PPV of Top-4\* when used as a prediction of the nature of mis-splicing. However, we feel that prioritising sensitivity is of greatest importance, to avoid false negative predictions. RNA re-analysis of 4 cases with one or more

SpliceVault predicts the precise nature of variant-associated mis-splicing.

undetected Top-4\* events via our initial RNA Diagnostics testing revealed we had missed 6/9 of these rare events, due to experimental design and/or technical limitations. *A priori* knowledge of Top-4\* mis-splicing events has been transformative for our research-led clinical RNA Diagnostics program, facilitating both variant curation and strategic experimental design of RNA assays to specifically target probable mis-splicing events; expressly important for RT-PCR where primer design and extension times strongly influence which products may be amplified.

Our reinterrogation of early cases, showing we had missed several rare events, raises the possibility that Top-4\* PPV could be higher than we currently estimate. It also reinforces clinical benefits of being able to reliably predict probable mis-splicing events to improve completeness and accuracy of conclusions drawn from RNA diagnostics. Importantly, we cross-checked all other early cases (before 300K-RNA) to confirm that interpretation of likely pathogenicity would not be impacted by any undetected Top-4\* events that may have resulted (it was neither feasible nor economic to re-test all specimens).

Top-ranked unannotated splicing events are highly concordant between tissues and between GTEx and SRA (Figures 3A-B and S3A-B), in line with previous analyses showing that transcription rates rather than alternative splicing patterns underpin most tissue-specific variation in GTEx<sup>25</sup>. We suspect the predictive accuracy of our ranking method may be improved further by: 1) higher read depth RNA-Seq data across the breadth of manifesting tissues in rare disorders (e.g. there is no data currently available for cochlear), 2) incorporation of RNA-Seq data from human fetal samples to catalogue developmental alternative splicing, 3) use of cycloheximide to inhibit nonsense-mediate decay, 4) a better understanding of contexts that influence variant-associated double exon skipping, 5) an ability to rank likelihood of variant-activated intron retention, and 6) improved bioinformatic methods in sequencing read alignment. Consideration of Top-3\* events in specific manifesting tissues, if shown to maintain > 90% sensitivity, could substantially improve PPV and clinical utility for variant classification.

Long-read RNA may prospectively assist evaluation of new methods to improve fidelity of short-read alignment and reduce splice-junction artifacts in 300K-RNA, though has its own bioinformatic and technical challenges. Extended Data Figure 9 shows that long-read RNA-Seq (7M mean read depth, 740 nt average length) of 22 fibroblast specimens in GTEx v9<sup>26</sup>

SpliceVault predicts the precise nature of variant-associated mis-splicing.

identifies only 15% (180K) of the 1.16M total splice-junctions detected in 504 GTEx v8 fibroblast specimens (75 bp paired reads, Poly A enriched, 80M read depth). For comparison, in-house RNA-Seq data from 7 fibroblast specimens (150 bp paired reads, rRNA depleted total RNA, 100 – 200M read depth) subject to CHX treatment identifies 36% (420K/1.16M), substantially more than DMSO treated specimens (340K/1.16M) or in two randomised sets of 7 fibroblast specimens from GTEx v8 (245K/1.16M). This preliminary evidence indicates CHX treatment and high read depth of human cell lines may substantially enhance our detection of unannotated mis-splicing events through inhibition of NMD.

To our knowledge, 300K-RNA Top-4\* is the first evidence-based method for predicting the nature of variant-associated mis-splicing and will assist clinical laboratories in application of PVS1 or PP3/BP4/BP7, prioritisation of VUS for RNA analysis, as well as guide RNA-diagnostic testing to experimentally determine consequences for pre-mRNA splicing. Informed by the current investigation, Extended Data Figure 10 details our draft guidelines for possible use of 300K-RNA to assist application of the PVS1 criterion to essential splice-site variants, aligning closely with revised PVS1 guidelines<sup>10</sup>. Theorized consideration of intron retention and Top-4\* events by genetic pathology workforces provides a pragmatic, evidence-based method to reliably assess variant-associated exon-skipping and/or cryptic splice-site use within a larger distance window of 600 nt. Over 2022-23, the Australasian Consortium for RNA Diagnostics (SpliceACORD)<sup>3</sup> will rigorously evaluate the clinical accuracy and usefulness of Extended Data Figure 10 draft guidelines for cases prospectively recruited into RNA Diagnostic testing pipelines. Extended Data Figure 7E indicates these draft guidelines will allow application of PVS1 for ~50% of ES variants (IR and Top-4\*  $\leq$  2 in-frame events).

We provide SpliceVault, a web portal to access 300K-RNA (and 40K-RNA in hg19), which quantifies natural variation in splicing and potentially predicts the nature of variant-associated mis-splicing: (<https://kidsneuro.shinyapps.io/splicevault/>). Users require no bioinformatics expertise and can retrieve stochastic mis-splicing events for any splice-junction annotated in Ensembl or RefSeq. Default settings display 300K-RNA Top-4\* output according to the optimised parameters we describe herein, with the option to return all events, customise the number of events returned, distance scanned for cryptic splice-sites, maximum number of exons skipped, or list tissue-specific mis-splicing events. We hope SpliceVault will improve

SpliceVault predicts the precise nature of variant-associated mis-splicing.

the ability to classify and study splicing variants with accuracy and completeness, avoiding the non-actionable diagnostic endpoint of a variant of uncertain significance (VUS).

### **Acknowledgements**

We thank the families for their participation and invaluable contributions to this research. We also thank the clinicians and health care workers involved in their assessment and management.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx portal and dbGaP accession number phs000424.v8.p2 and phs000424.v9.

S.T.C. is supported by a National Health and Medical Research Council (NHMRC) of Australia Senior Research Fellowship APP1136197. This project received funding through NHMRC Ideas Grants APP1106084 and APP2002640 and a Medical Research Future Fund Rapid Applied Research Translation Program grant awarded to Sydney Health Partners. Part of this work was supported by Luminesce Alliance Innovation for Children's Health, a not for profit joint venture between the Sydney Children's Hospitals Network, the Children's Medical Research Institute and the Children's Cancer Institute, and from the Lenity Australia Foundation, a not-for-profit philanthropic organisation. R.D. and A.M.B. are supported by a University of Sydney Research Training Scholarship. S.J.B. is supported by a Muscular Dystrophy Association of New South Wales Sue Connor postgraduate training scholarship.

### **Author Contributions Statement**

R.D. conceived the project. R.D., A.M.B., S.J.B., R.M., H.J. curated datasets. R.D., R.M., H.J. performed computational analysis. A.M.B., S.J.B., S.B. performed experimental validation of splicing variants. R.D., A.M.B., S.T.C. wrote original draft manuscript with input and editing from all authors. R.D. and A.M.B. contributed equally. H.J. and S.T.C. jointly supervised this work.

## Competing Interests

Sandra T. Cooper is director and shareholder of Frontier Genomics Pty Ltd (Australia).

Sandra T. Cooper currently receives no consultancy fees or other remuneration for this role.

Himanshu Joshi offers Technology advice to Frontier Genomics Pty Ltd (Australia) and receives no remuneration for this role. The remaining authors declare no competing interests.

## References

1. Baralle, D. & Buratti, E. RNA splicing in human disease and in the clinic. *Clinical Science* **131**, 355–368 (2017).
2. López-Bigas, N., Audit, B., Ouzounis, C., Parra, G. & Guigó, R. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett* **579**, 1900–1903 (2005).
3. Bournazos, A. M. *et al.* Standardized practices for RNA diagnostics using clinically accessible specimens reclassifies 75% of putative splicing variants. *Genetics in Medicine* (2021) doi:10.1016/j.gim.2021.09.001.
4. Wai, H. A. *et al.* Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet Med* **22**, 1005–1014 (2020).
5. Murdock, D. R. *et al.* Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. *Journal of Clinical Investigation* (2020) doi:10.1172/JCI141500.
6. Maddirevula, S. *et al.* Analysis of transcript-deleterious variants in Mendelian disorders: implications for RNA-based diagnostics. *Genome Biology* **21**, 145 (2020).
7. Frésard, L. *et al.* Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nature Medicine* **25**, 911–919 (2019).
8. Lee, H. *et al.* Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genetics in Medicine* **22**, 490–499 (2020).
9. Richards, S. *et al.* 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**, 405–424 (2015).
10. Abou Tayoun, A. N. *et al.* Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum Mutat* **39**, 1517–1524 (2018).
11. Dawes, R., Joshi, H. & Cooper, S. T. Empirical prediction of variant-activated cryptic splice donors using population-based RNA-Seq data. *Nat Commun* **13**, 1655 (2022).
12. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
13. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
14. Brandão, R. D. *et al.* Targeted RNA-seq successfully identifies normal and pathogenic splicing events in breast/ovarian cancer susceptibility and Lynch syndrome genes. *International Journal of Cancer* **145**, 401–414 (2019).
15. Nellore, A. *et al.* Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biology* **17**, 266 (2016).
16. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Research* **49**, D884–D891 (2021).
17. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–745 (2016).
18. Leinonen, R., Sugawara, H. & Shumway, M. The Sequence Read Archive. *Nucleic Acids Res* **39**, D19–D21 (2011).
19. Wilks, C. *et al.* recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biology* **22**, 323 (2021).
20. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019).

SpliceVault predicts the precise nature of variant-associated mis-splicing.

21. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science Translational Medicine* **9**, (2017).
22. Akesson, L. S. *et al.* Rapid exome sequencing and adjunct RNA studies confirm the pathogenicity of a novel homozygous ASNS splicing variant in a critically ill neonate. *Human Mutation* **41**, 1884–1891 (2020).
23. Katiyar, D. *et al.* Two novel B9D1 variants causing Joubert syndrome: Utility of mRNA and splicing studies. *European Journal of Medical Genetics* **63**, 104000 (2020).
24. Jones, H. F. *et al.* Importance of muscle biopsy to establish pathogenicity of DMD missense and splice variants. *Neuromuscular Disorders* **29**, 913–919 (2019).
25. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
26. Glinos, D. A. *et al.* Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* 1–8 (2022) doi:10.1038/s41586-022-05035-y.
27. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).

## Online Methods

### *Ethics declaration*

Consent for diagnostic genomic testing was supported by governance infrastructure of the relevant local ethics committees of the participating Australian Public Health Local Area Health Districts. Kids Neuroscience Centre’s biobanking and functional genomics human ethics protocol was approved by the Sydney Children’s Hospitals Network Human Research Ethics Committee (protocol 10/CHW/45 renewed with protocol 2019/ETH11736 (July 2019 – 2024)) with informed, written consent for all participants.

### *Creating 300K-RNA*

The 300K-RNA database collates splice-junctions collated from 335,663 publicly available RNA-Seq samples from GTEx<sup>12</sup> and SRA<sup>18</sup> using a unified Monorail pipeline by Wilks *et al.*<sup>19</sup>. Splice-junction read counts derived from 316,449 human RNA-Seq samples from SRA and 19,214 human RNA-Seq samples from GTEx were downloaded from the public resource recount3<sup>19</sup>. We then filtered split-reads to those which span at least one annotated splice-site, and for each splice-junction detected we tallied the number of samples it occurred in across the two data sources. For each splice-junction we filtered associated detected exon-skipping and cryptic-activation events according to the rules in Figure 2D-E.

Unannotated splicing events were ranked according to the number of samples in which the event was detected. This ranking process was completed with respect to each annotated splice-junction in ensembl transcripts (v104) and refseq transcripts (GRCh38, downloaded Aug 2021)

SpliceVault predicts the precise nature of variant-associated mis-splicing.

300K-RNA enables customized access to ranked splice-junction data from individual GTEx tissue sub-types. SRA metadata precludes breakdown into specimen subtypes. As SRA contains data from cancer specimens (genetically heterogeneous), maximum read-counts output for each splice junction are derived from GTEx data. Code used to create 300K-RNA is available at <https://github.com/kidsneuro-lab/300K-RNA>. The R package Snapcount<sup>28</sup> was used to retrieve information on individual samples for Figures 2E-F and S6B-F.

#### *SpliceAI $\Delta$ -score interpretive rules*

By default, SpliceAI outputs four delta ( $\Delta$ ) scores: acceptor loss, acceptor gain, donor loss and donor gain; for each the maximum  $\Delta$ -score within +/- 50 nt of the variant is reported. To adapt SpliceAI to the prediction of mis-splicing, we instead retrieved all  $\Delta$ -scores +/- 5000 nt of each variant, using a python script adapted from the SpliceAI GitHub (<https://github.com/Illumina/SpliceAI>). As input, we used the pre-mRNA sequence +/- 5000 nt of the variant. Two  $\Delta$ -scores returned at each base (variant nucleotide versus reference nucleotide) generated up to 20,002  $\Delta$ -scores per variant, of which we excluded all  $\Delta$ -scores  $\leq$  0.001 as neutral impact.

Across the 86 variants which could be scored by SpliceAI, 2,836  $\Delta$ -scores returned were above the 0.001 threshold. 86 of these were donor loss or acceptor loss  $\Delta$ -scores of the affected annotated splice-site, denoting a prediction of mis-splicing. Our custom interpretive rules (see explanation of Figure 4A-C in the manuscript text) applied to any  $\Delta$ -score  $>$  0.001 yielded predictions of erroneous use of: 161 cryptic acceptors, 340 cryptic donors, 215 exon skipping events and 49 intron retention events. Of the remaining 2,071 predictions; 1,637 were decreases in scores of unannotated splice-sites, 33 were increases in scores of annotated splice-sites, 315 were increases in the scores of unannotated splice-sites outside the bounds of the exon and intron flanking the variant splice-site or increases in the scores of unannotated donors for acceptor variants and vice versa – and deemed uninterpretable within our paradigm.

#### *Clinically Relevant Mendelian Disease Genes (Mendeliome)*

Clinically relevant genes were extracted from the Genomics England Panel App<sup>29</sup> (September 2021 release) and Online Mendelian Inheritance in Man database<sup>30</sup> (OMIM; February 2021

SpliceVault predicts the precise nature of variant-associated mis-splicing.

release). Genes were extracted from the Genomics England PanelApp using the Swagger PanelApp API (v1) ([panelapp.genomicsengland.co.uk/api/docs/](http://panelapp.genomicsengland.co.uk/api/docs/)), excluding disease susceptibility panels, and all genes below a confidence level 3 (green; diagnostic-grade). OMIM listed genes were excluded when their only phenotype associations were non-diseases, susceptibilities, provisional links, and somatic mutations.

#### *RNA re-analysis*

RNA extraction from whole blood, peripheral blood mononuclear cells or fibroblasts, and reverse transcription polymerase chain reaction and Sanger sequencing performed as described in Bournazos et al<sup>3</sup>. See Supplementary Table 4 for primers used for re-analysis of RNA by RT-PCR.

#### *Statistics and reproducibility*

No statistical method was used to predetermine sample size. All splicing variants were included in this study for which robust RNA assay data was available and met the following ascertainment criteria: *Inclusion Criteria*: Variants affecting an annotated splice-site and demonstrated to activate exon-skipping or cryptic splice-site use. *Exclusion Criteria*: Variants creating or modifying a cryptic splice-site. The experiments were not randomized. The Investigators were not blinded to allocation during experiments and outcome assessment.

#### **Data Availability**

Source data for 300K-RNA was downloaded from snaptron (<http://snaptron.cs.jhu.edu/data>). 300K-RNA can be easily accessed and queried through SpliceVault (<https://kidsneuro.shinyapps.io/splicevault/>). The data used for the analyses described in this manuscript were obtained from the GTEx portal and dbGaP accession number phs000424.v8.p2 and phs000424.v9.

#### **Code Availability**

All code required to perform analyses and generate Figures 1C, 2D-F, 3, 4E-G and Extended Data Figures 1-6, 8 and 9 is available at [https://github.com/kidsneuro-lab/SpliceVault\\_figures](https://github.com/kidsneuro-lab/SpliceVault_figures). All code required to create 300K-RNA is available at <https://github.com/kidsneuro-lab/300K-RNA>. Code used to create SpliceVault is available at <https://github.com/kidsneuro-lab/SpliceVault/>.

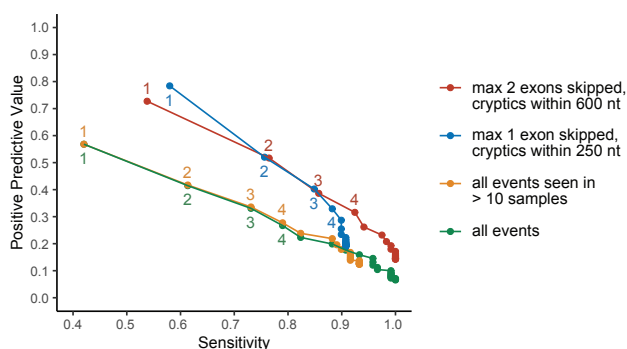


SpliceVault predicts the precise nature of variant-associated mis-splicing.

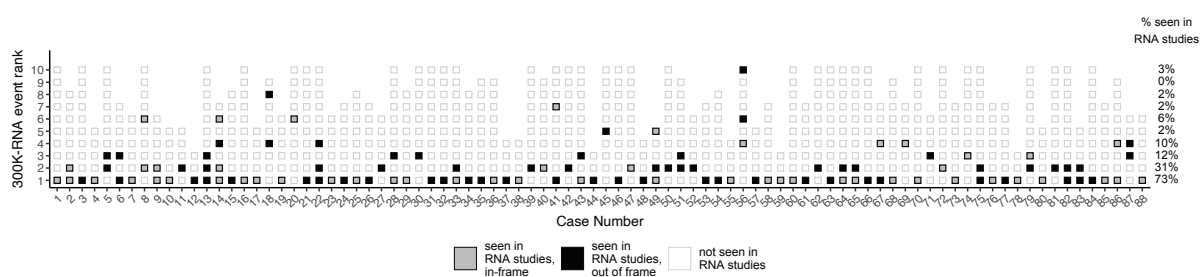
## Methods-only references

28. Wilks, C., Charles, R. & Langmead, B. snapcount: R/Bioconductor Package for interfacing with Snaptron for rapid querying of expression counts. (2019).  
Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).
- . Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science Translational Medicine* **9**, (2017).  
Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
29. Martin, A. R. *et al.* PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet* **51**, 1560–1565 (2019).
30. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). Online Mendelian Inheritance in Man, OMIM®. World Wide Web URL: <https://omim.org/> (2021).

## Extended Data

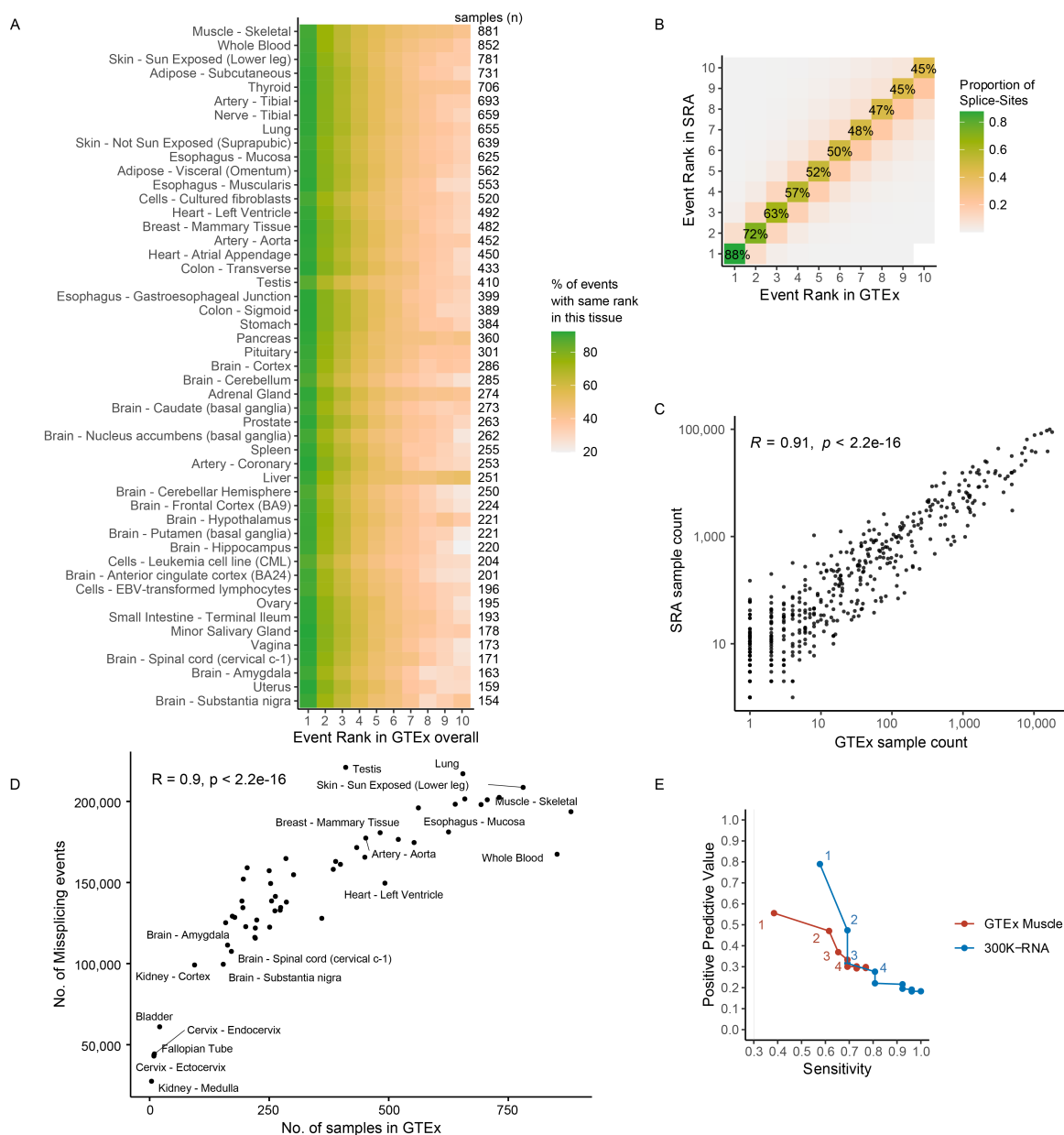


**Extended Data Figure 1. Sensitivity and PPV of 300K-RNA using different filtering criteria for ranking mis-splicing events.** The asterisk i.e ‘Top-4\*’ is used denote application of a filter to limit ranked events to those involving skipping of one or two exons and cryptic activation within 600 nt of the annotated splice-site.



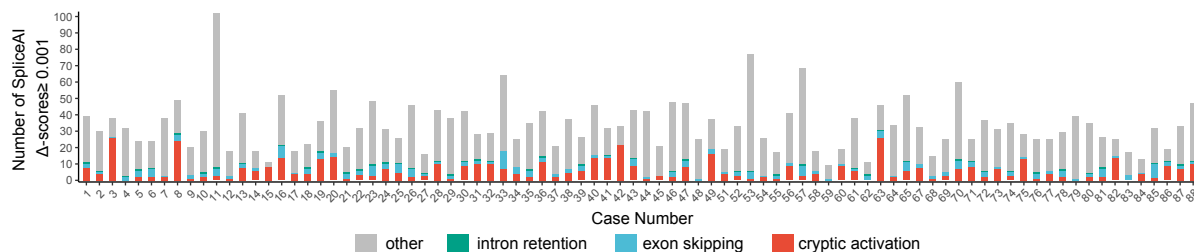
**Extended Data Figure 2. Top-10\* mis-splicing events seen in 300K-RNA\* for our cohort of 88 variants, filled if they were seen in RNA studies.** *Grey fill*: multiple of 3 (maintains frame). *Black fill*: not a multiple of 3 (disrupts frame). *No fill*: Event not seen in RNA studies. \* = skipping one or two exons and cryptic activation within 600 nt of the annotated splice-site.

# SpliceVault predicts the precise nature of variant-associated mis-splicing.

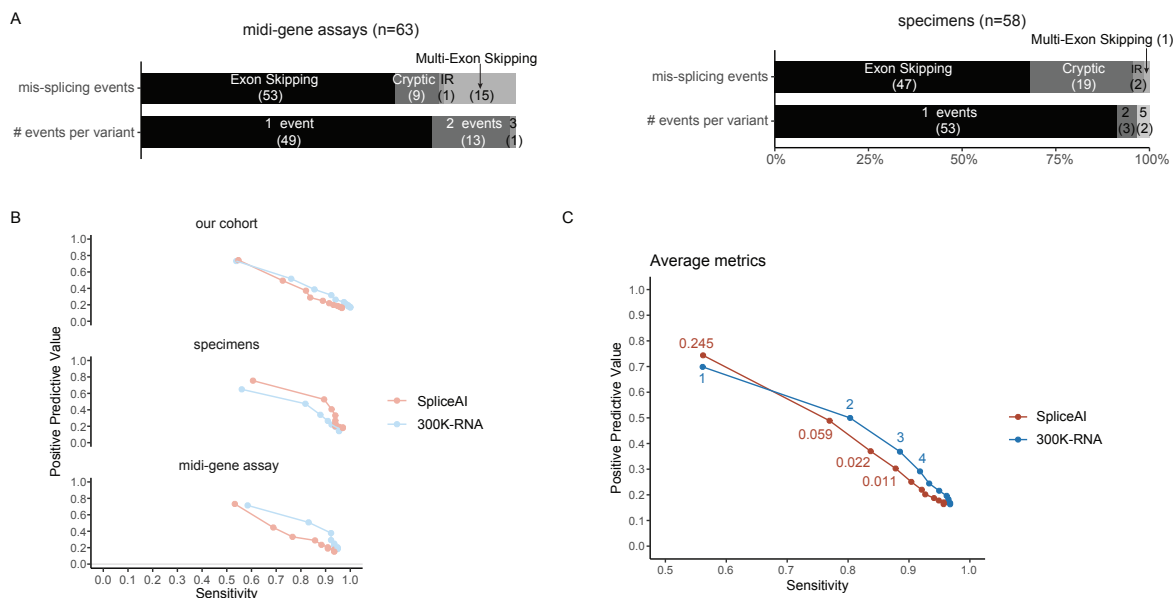


**Extended Data Figure 3. 300K-RNA event rankings across tissues and data-sources.** **A)** Heatmap showing the proportion of mis-splicing events\* with the same event rank in each GTEx tissue subtype, as compared to all GTEx tissue subtypes combined, for 98,810 annotated splice-sites in the Mendeliome (see methods). The top-1\* event is > 86% concordant across all tissues with > 100 samples in GTEx. **B)** Concordance of top-ranked mis-splicing events\* in GTEx versus SRA. The top-1\* event in GTEx is the top-1\* event in SRA for 88% all splice-sites in the Mendeliome. **C)** Sample counts are highly correlated between GTEx and SRA for all unannotated mis-splicing events\* in 300K-RNA at the splice-sites affected by our cohort of 88 variants (Spearman correlation,  $R = 0.91$ ,  $p < 2.2e^{-16}$ ). **D)** Sequencing breadth in GTEx samples increases sensitivity i.e. the more specimens the greater the proportion of total mis-splicing events detected, with testis being a notable outlier (Spearman correlation,  $R = 0.91$ ,  $p < 2.2e^{-16}$ ). **E)** Sensitivity and PPV of 300K-RNA Top-4\* versus GTEx Muscle Top-4\* derived from 881 samples in GTEx, for 19 variants associated with muscle disorders and where RNA testing was performed using muscle RNA. \* = skipping one or two exons and cryptic activation within 600 nt of the annotated splice-site

SpliceVault predicts the precise nature of variant-associated mis-splicing.

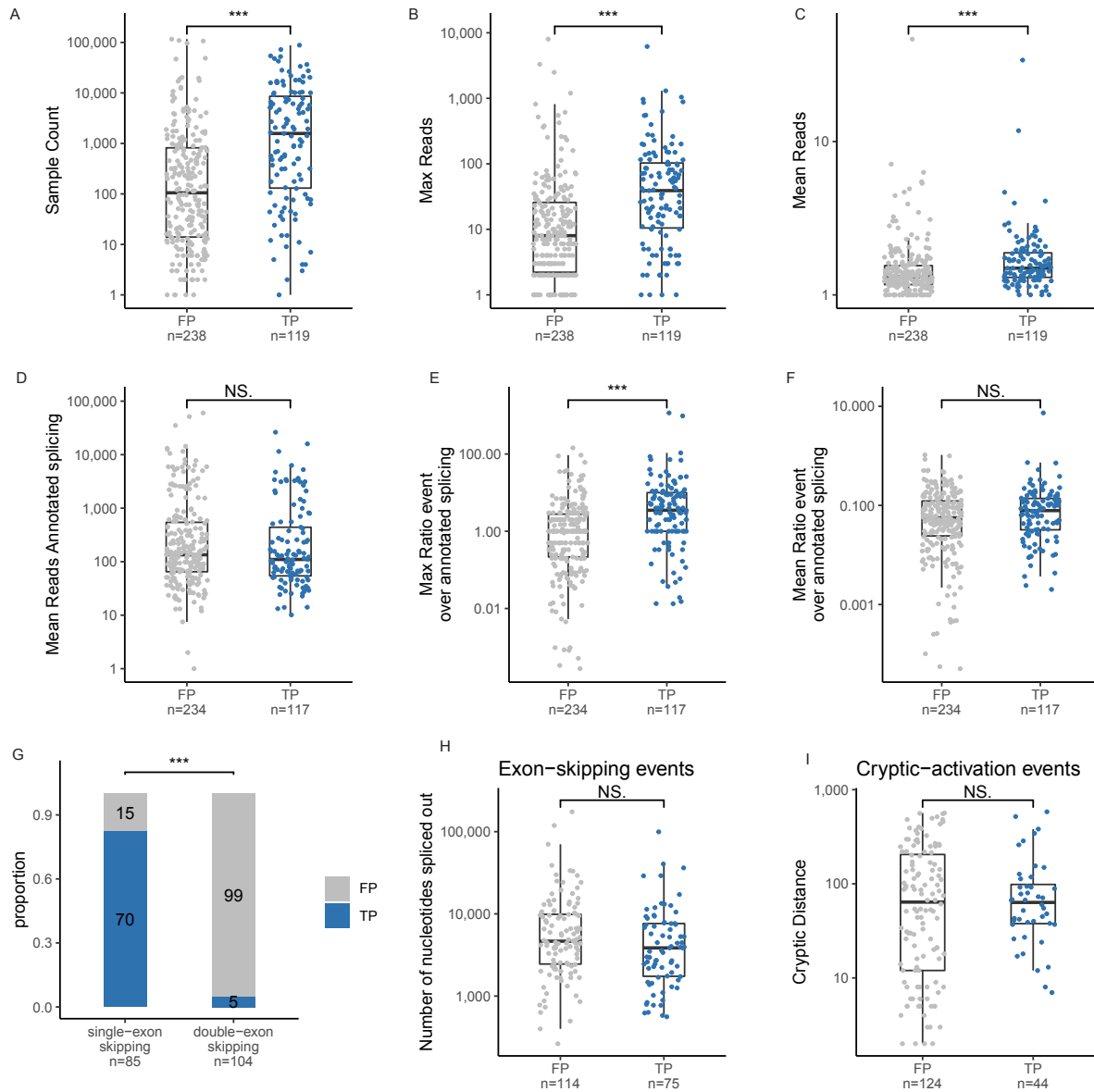


**Extended Data Figure 4. SpliceAI  $\Delta$ -scores above 0.001 for 86/88 variants scored by SpliceAI.** Predictions are coloured according to the event type assigned by our interpretive rules.



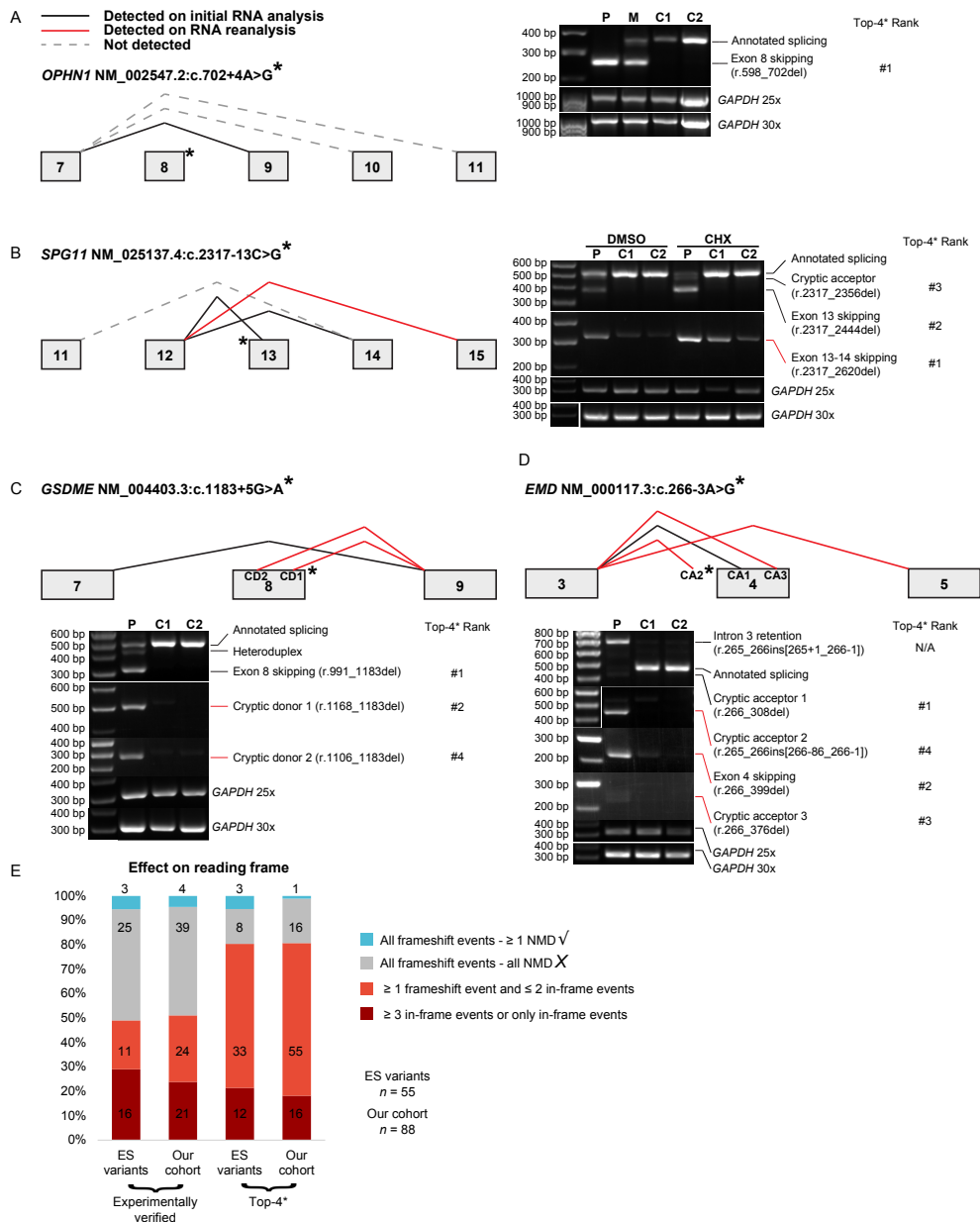
**Extended Data Figure 5. Comparison of 300K-RNA Top-4\* with SpliceAI using two additional variant cohorts.** **A)** Mis-splicing events induced by variants in two cohorts curated from literature for additional validation of the Top-4\* approach: 58 variants studied in patient specimens and 63 variants studied through midi-gene assays (scrutinized to ensure technical design permitted detection of multi-exon skipping events, see Supplementary Tables 2,3 for references). **B)** Sensitivity and PPV of 300K-RNA and SpliceAI for exon-skipping and cryptic activation predictions at different thresholds. Points on the curves correspond to thresholds used in Figure 2B. **C)** The average sensitivity and PPV of 300K-RNA and SpliceAI across three variant cohorts shown in (B). Points on the curves correspond to thresholds used in Figure 4B. \* = skipping one or two exons and cryptic activation within 600 nt of the annotated splice-site.

SpliceVault predicts the precise nature of variant-associated mis-splicing.



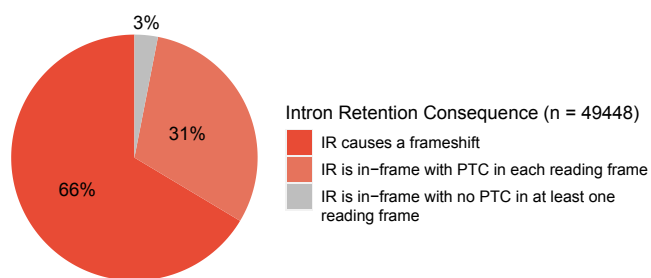
**Extended Data Figure 6. Investigating features of events seen/not seen in RNA studies for our 88 variants.** **A)** True positive (TP) events seen in RNA studies tend to be seen in higher sample numbers than false positives (FP) in the Top-4 (two-sided Wilcoxon rank sum test;  $W = 20,222$ ,  $p = 4e^{-11}$ ,  $n = 238$  Top-4\* FP and 119 TP biologically independent events). **B-C)** TP tend to be seen in a higher number of max (**B**) and mean (**C**) reads than FP (two-sided Wilcoxon rank sum test;  $W = 20,144$ ,  $p = 7e^{-11}$  and  $W = 18,284$ ,  $p = 7e^{-6}$  respectively,  $n = 238$  Top-4\* FP and 119 TP biologically independent events). **D)** There is no significant difference in the mean read-depth for the annotated splice-junction around which the unannotated event is detected, between TP and FP (two-sided Wilcoxon rank sum test;  $W = 12,848$ ,  $p = 0.3$ ,  $n = 234$  Top-4\* FP and 117 TP biologically independent events). The mean read depth is taken only across samples where the event is detected. 4/238 FP and 2/119 TP were detected only in samples where no annotated splicing was detected, so are excluded from D-F. **E)** The maximum ratio of the reads representing the unannotated event and the annotated event in any one sample tends to be higher in TP than FP (two-sided Wilcoxon rank sum test;  $W = 19,126$ ,  $p = 1e^{-9}$ ,  $n = 234$  Top-4\* FP and 117 TP biologically independent events), however there is no significant difference in the mean ratio (two-sided Wilcoxon rank sum test;  $W = 15,181$ ,  $p = 0.1$ ,  $n = 234$  Top-4\* FP and 117 TP biologically independent events) (**F**). **G)** Single-exon skipping events\* are significantly more likely to be seen in RNA studies than double-exon skipping events\* (Chi-squared test;  $\chi^2 = 114.29$ ,  $p = 1.1e^{-26}$ ). **H)** Total length (nt) of the fragment excised from the pre-mRNA from single and double exon skipping was not statistically different between events seen/not seen in RNA studies (two-sided Wilcoxon rank sum test;  $W = 2,554$ ,  $p = 0.0502$ ,  $n = 114$  Top-4\* FP and 75 TP biologically independent exon-skipping events). **I)** Relative distance between the annotated splice-site and activated cryptic splice-site was not statistically different between TP and FP (two-sided Wilcoxon rank sum test;  $W = 2,834$ ,  $p = 0.70$ ,  $n = 124$  Top-4\* FP and 44 TP biologically independent cryptic-activation events). A-E and H-I are box-whisker plots, with internal lines denoting the median value, and the lower and upper limits of the boxes representing 25<sup>th</sup> and 75<sup>th</sup> percentiles. Whiskers extend to the largest and smallest values at most 1.5IQR. \* = skipping of one or two exons or cryptic activation within 600 nt of the annotated splice-site.

# SpliceVault predicts the precise nature of variant-associated mis-splicing.

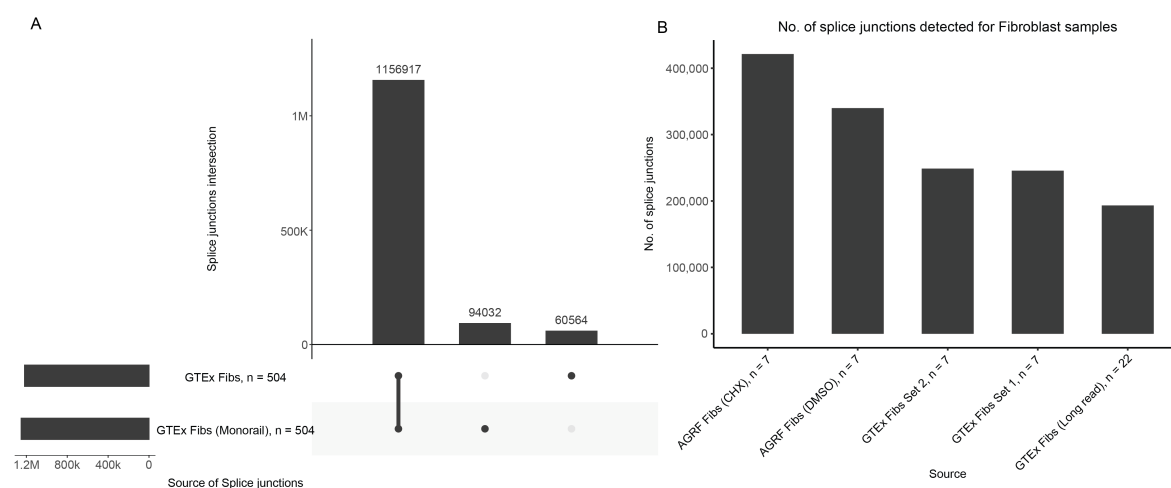


**Extended Data Figure 7. RNA re-analysis to check for undetected 300K-RNA Top-4 mis-splicing events.** *Black lines:* mis-splicing identified during initial RNA analysis<sup>3</sup>. *Red lines:* Top-4\* events detected upon re-analysis. *Grey lines:* Top-4\* events undetected upon re-analysis. **A)** No additional Top-4 events were identified for *OPHN1* c.702+4A>G. **B)** For *SPG11* c.2317-13C>G, Top-1\* exon 13 + 14 skipping was detected on re-analysis (missed initially by RT-PCR due to primer positioning and by RNA-seq due to low read depth and NMD), but Top-4\* event, exon 12+13 skipping, was not detected. **C)** RT-PCR using primers specific for two exonic cryptic donors shows their variant-associated increased use for *GSDME* c.1183+5G>A. These rare events were missed during initial RNA analysis due to PCR biases and challenges resolving Sanger sequencing chromatograms due to heteroduplex formation. **D)** RT-PCR identifies rare use of two cryptic acceptors and exon 4 skipping associated with *EMD* c.266-3A>G missed during initial RNA analysis due to PCR biases and heteroduplex formation. **E)** 49% (27/55) essential splice-site variants (ES) across the three variant datasets induce  $\geq 1$  in-frame events, with a similar proportion of 51% (45/88) in our overall cohort. Use of Intron Retention (IR) and Top-4\* as proxy for a prediction of variant-induced mis-splicing increases the relative number of ES variants with  $\geq 1$  in-frame event to 80% (45/55) and to 81% (71/88) for our overall cohort. P=proband, C1=control 1, C2=control 2, DMSO=dimethyl sulfoxide, CHX=cycloheximide, CD1=cryptic donor 1, CD2=cryptic donor 2, CA1=cryptic acceptor 1, CA2=cryptic acceptor 2, CA3=cryptic acceptor 3, ES=essential splice-site, NMD=nonsense mediated decay.

SpliceVault predicts the precise nature of variant-associated mis-splicing.



**Extended Data Figure 8. The consequence of intron retention upon the open reading frame for 49,448 canonical introns in the Mendeliome** (see methods). Variant-activated intron retention elicits a frameshift for 66% introns or encodes a premature termination codon (PTC) in all reading frames for 31% introns. In summary, IR will induce a frameshift or encode a PTC for at least 97% cases and is therefore consistent with null outcomes in most instances. Intron coordinates were extracted from Ensembl (104) via the hg38 genome assembly. IR: intron retention; PTC: premature termination codon.



**Extended Data Figure 9. A)** Upset plot showing splice-junctions concordantly and uniquely detected by GTEx and Monorail processing pipelines for 504 fibroblast specimens. **B)** Upset plot showing GTEx V9 long-read RNA-Seq (7M mean read depth, 740 nt average length) for 22 fibroblast specimens identifies 15% (180K/1.16M) of all fibroblast SJ detected in 504 GTEx v8 fibroblast samples (75 bp paired reads, 80M depth, Poly A enrichment). In-house RNA-Seq data from 7 fibroblast specimens (150 bp paired reads, 100 – 200M depth, rRNA depleted total RNA) subject to CHX treatment identifies 36% (420K/1.16M) of all SJ, substantially more than DMSO treated specimens (340K/1.16M) or in two randomised sets of 7 fibroblast specimens from GTEx v8 (245K/1.16M). Splice junctions present in GTEx v9 long read RNA-Seq data was reverse-engineered from the transcript count information generated using FLAIR<sup>26</sup>

SpliceVault predicts the precise nature of variant-associated mis-splicing.

	PVS1_Very strong	PVS1_Strong	PVS1_Moderate
Empirical evidence	1. Intron retention (IR) and 300K-RNA Top-4* all result in a frameshift or encode a premature termination codon (PTC);	1. Intron retention (IR) and 300K-RNA Top-4* all result in a frameshift or encode a premature termination codon (PTC); <b>OR</b> 2. Intron retention (IR) and 300K-RNA Top-4* include one in-frame event.	1. Intron retention (IR) and 300K-RNA Top-4* all result in a frameshift or encode a premature termination codon (PTC); <b>OR</b> 2. Intron retention (IR) and 300K-RNA Top-4* include one or two in-frame events.
Transcript disease association	3. IR and 300K-RNA Top-4 events affect splicing of one or more <b>constitutive exon(s) present in the clinically-relevant isoform(s)</b> expressed by the manifesting tissue(s); <b>AND/OR</b> 4. IR and 300K-RNA Top-4 events activate inclusion of <b>ectopic/intronic sequences into the clinically-relevant isoform(s)</b> expressed by the manifesting tissue(s);		
Pathogenetic mechanism	5. Loss-of-function variants are a known causal basis for disease;	5. Loss-of-function variants are a known causal basis for disease; <b>AND/OR</b> 6. Truncating variants, missense variants and/or in-frame indels are a known causal basis for disease	
	6. <b>Transcripts with encoded PTCs</b> are predicted to activate nonsense mediated decay	6. <b>Transcripts with encoded PTCs</b> are predicted to activate nonsense mediated decay; <b>AND/OR</b> 7. <b>Transcripts with encoded PTCs</b> are in a genetic context that may evade NMD. However, there are one or more analogous truncating variants in the same region of the gene classified likely/pathogenic, <b>OR</b> , the truncated region is critical to function of the gene product. 8. <b>Transcript with an in-frame event</b> altering length of the gene product disrupt a region of the gene with: i) evident clinical importance as shown by presence of one or more causal missense variants or in-frame indels; <b>AND/OR</b> ii) evident functional importance via disruption of a known domain critical to function of the gene product.	6. <b>Transcripts with encoded PTCs</b> are predicted to activate nonsense mediated decay; <b>AND/OR</b> 7. <b>Transcripts with encoded PTCs</b> are in a genetic context that may evade NMD. However, there are one or more analogous truncating variants in the same region of the gene classified likely/pathogenic, <b>OR</b> , the truncated region is critical to function of the gene product. 8. <b>Transcripts with an in-frame event</b> altering length of the gene product disrupt a region of the gene with: i) evident clinical importance as shown by presence of one or more causal missense variants or in-frame indels; <b>AND/OR</b> ii) evident functional importance via disruption of a known domain critical to function of the gene product; <b>AND/OR</b> iii) inferred functional importance via disruption of an evolutionarily conserved region intolerant to genetic variation.

**Extended Data Figure 10. Draft Guidelines for potential use of empirical evidence from 300K-RNA to assist application of the PVS1 criterion for essential splice-site variants based on probable mis-splicing outcomes** (being assessed in a clinical evaluation trial by the Australian Consortium for RNA Diagnostics (SpliceACORD)). We recommend pathology consideration of Intron Retention (IR) and 300K RNA Top-4\* for all disease relevant transcript(s). PVS1 levels of evidence are influenced by the collective nature of probable induced mis-splicing, relative to evidence supporting null outcomes for the encoded gene product. For use of PVS1 at a Very Strong evidence level, IR and Top-4\* events should all be consistent with null outcomes. PVS1 applied at Strong or Moderate should be considered when IR and Top-4\* events include one or two in-frame events, adjusting the evidence weighting according to; the number and nature of in-frame events, known clinical relevance of the affected region of disease relevant transcript(s), and the established pathogenetic mechanism(s) associated with a given gene and disorder. We favour weighting of known clinical relevance, biological function or evolutionary conservation of the disrupted gene region, over relative length of the in-frame disruption. We recommend additional consideration of abnormal or alternative transcription initiation or termination for first intron and last intron variants, respectively. We recommend use of the PM4 criterion for essential splice-site variants when IR and Top-4\* events include three or more in-frame events.

## Discussion

### 4.1 Developing a method to predict the mis-splicing outcomes of splicing variants

#### 4.1.1 Current splicing algorithms are not geared towards aiding pathology interpretation.

*Pathology interpretation of splicing variants requires consideration of likely outcomes for the mRNA transcript.*

Whether a splicing variant is causally linked to a rare disease (i.e. pathogenic) depends critically on its impact on the mRNA transcript(s) and downstream dysfunction of the encoded gene product, making pathology interpretation extremely challenging. The activation (or upregulation) of exon skipping, cryptic splicing, and intron retention events at a genetic locus can generate transcripts degraded by NMD, or transcripts producing aberrant protein, at different quantities and in different cell-types. The specific mis-splicing outcomes induced by a variant and the amount of normally spliced transcript still produced can have a critical impact on pathology interpretation, such as occurs for the differential diagnosis of severe-lethal Duchenne, or milder Becker muscular dystrophy phenotypes based on either in-frame or out-of-frame consequences elicited by variants affecting *DMD*.

*The focus of algorithms so far has been variant prioritisation / triage to functional studies.*

In the past decade, new machine-learning and deep-learning algorithms that can predict with increased reliability whether a DNA variant is likely to disrupt splicing have emerged<sup>43,115,116,119,121,122,124,125,133</sup>. However, they have been largely geared towards triaging high-likelihood splice-altering variants for RNA functional studies that can ascertain the specific splicing defect(s) induced and inform clinical relevance. Most algorithms therefore predict simply whether mis-splicing will occur in some capacity or not, not the specific mis-splicing outcome induced. While SPANR<sup>115</sup> and MMSplice<sup>116</sup> offer predictions of the likelihood of exon skipping, these predictions have not been validated in a clinical setting. Chapter 2 critically evaluates the performance of SpliceAI<sup>124</sup> to predict variant-associated



activation of cryptic donors with 75% sensitivity and 99% specificity. Chapter 3 shows SpliceAI output can also be adapted to predict exon skipping (85% sensitivity, 52% Positive Predictive Value). While current splicing algorithms are invaluable triage tools, my thesis responds to the need for evidence-based, clinically-validated tools for pathology interpretation of splicing variants.

### **4.1.2 Past splicing behaviour is a potent predictor of future behaviour**

*Variant-associated mis-splicing is upregulated 'leaky' mis-splicing.*

In chapter 2 and 3, we develop a novel method to predict the probable nature of variant-induced mis-splicing of the pre-mRNA, based on empirical data. Split-reads representing the use of a cryptic splice-site or exon-skipping event seen in an RNA-seq sample constitute empirical proof that a given splicing event is biophysically executable: proof made even stronger by the recurrence of this split-read across hundreds or thousands of independent RNA-seq samples, across different tissues and sample sources. The key finding of my thesis is that in the event of a deleterious splicing variant, the spliceosome will revert to whatever common 'mistakes' are made usually at that splice-site, making predicting variant-associated cryptic activation and exon skipping events as simple as cataloguing past behaviour at that splice-site, using public databases of RNA-sequencing.

*Empirical evidence presents advantages for incorporation into clinical diagnostics*

While SpliceAI can be adapted to predict cryptic activation and exon skipping with success (Chapter 3), SpliceVault's basis in the reporting of empirical data has inherent advantages for its incorporation into clinical diagnostics. Using empirical proof that a splice-site is useable by the spliceosome for predictive purposes negates the need for *a priori* knowledge of the constellation of short and long-range sequence features regulating any given splicing decision. While SpliceAI requires 5,000 nt of sequence context on either side of the intron-exon border to provide a prediction of what may happen, SpliceVault simply reports what does in fact happen, even if we can't yet understand exactly how or why.

This confers the advantage of providing empirical data that can be interpreted and weighted by human physicians to inform genetic diagnoses, which underpin prognosis, treatment, and family planning for patients and families. A molecular diagnosis has profound implications for clinical care and must be supported by robust evidence. ACMG/AMP guidelines are the current standard for clinical variant interpretation, and these rules are built around a framework of layering evidence from multiple independent sources and with different weighted ‘evidence levels’, to classify a variant. SpliceVault was developed with ACMG/AMP guidelines in mind, intended to provide evidence that can be weighted to directly inform diagnosis.

In contrast, deep learning algorithms are essentially ‘black-boxes’, meaning that predictions lack transparency and human-understanding. Importantly, unlike other clinical applications of AI coming into use – for example the interpretation of radiology imaging, SpliceAI is not using deep learning to mimic something a human can already do (and a human physician can validate independently). Instead SpliceAI is using deep learning to do a task beyond human comprehension – predict the likelihood of bonafide exons and their flanking splice-sites among 10,000 nt of scanned pre-mRNA sequence constitutes a splice-site in the human genome. It’s therefore hard to imagine deep learning predictions underpinning clinical decision making in this context and at this point in time, without confirmation from an orthogonal method.

### **4.1.3 Implications for variant interpretation and RNA functional studies.**

*Variant-associated mis-splicing vs ‘leaky’ mis-splicing.*

We have shown that most, if not all, mis-splicing events caused by variants that deleteriously impact authentic splice-sites (as opposed to variants creating novel cryptic splice-sites), already happen *sans* variant as rare, ‘leaky’ mis-splicing events across the population. This has important general implications for the interpretation of splicing variants and additionally for the development of tools to detect aberrant splicing in patient samples.

### *RT-PCR methods*

SpliceVault can assist and augment RT-PCR based RNA-diagnostics by allowing informed design of primers to amplify expected mis-splicing events. In Chapter 3 we found that many 300K-RNA Top-4 events not found in initial RNA studies were found upon reanalysis using primers designed specifically to amplify SpliceVault predictions. SpliceVault predictions will have great utility in guiding RT-PCR primer design, which strongly influence which products may be amplified and so detected.

### *RNA-seq methods*

There has been recent interest in developing methods to detect ‘outlier’ splicing events in patient RNA-seq samples, by comparison to splicing events seen in either public RNA-seq databases (often GTEx) or else RNA-seq samples from disease controls<sup>94,134–139</sup>. However, in chapter 3, we show that 100% of mis-splicing events seen across our 88 splicing variants are present in at least 1 sample in GTEx or SRA, and 92% are among the 4 most ubiquitous mis-splicing events at that site across samples in 300K-RNA. As we’ve shown that mis-splicing events induced by splicing variants represent abnormal levels of events which occur naturally, the level of mis-splicing that is within ‘normal range’ requires precise quantification to properly identify outliers. Importantly, approaches which aim to identify outliers by filtering mis-splicing events seen in disease-samples to those absent or seen in less than two<sup>138</sup> or five<sup>139</sup> or some other arbitrary number of control samples will need to be revisited.

### *Implications for the application of PVS1*

According to revised PVS1 guidelines from 2018, the application of PVS1 for essential splice-site variants requires theoretical consideration of consequences from exon-skipping, intron retention and use of any cryptic splice-site within 20 nucleotides (nt)<sup>83</sup>- a task which currently amounts to guess work, or the use of *in silico* tools which have not been clinically validated for this purpose. Chapter 3 details draft clinical guidelines for the application of PVS1 using an evidence-based method to theorise likely outcomes for the application of PVS1, rather than current guesswork. We recommend pathology consideration of Intron Retention and 300K RNA Top-4\* for all disease relevant transcript(s).

Using evidence from emergent RNA Diagnostic testing pipelines, these draft guidelines developed in Chapter 3 for the use of SpliceVault in the application of PVS1 will undergo testing and revision by the Australasian Consortium for RNA Diagnostics (SpliceACORD) that comprises a group of >150 genetic experts<sup>89</sup>.

## **4.2 Future Directions- improving and expanding SpliceVault**

*Higher quality datasets will give us a better picture of ‘normal’ mis-splicing*

In Chapter 3, the bulk of RNA-seq samples that constitute 300K-RNA are derived from SRA- a heterogeneous database including for example single-cell RNA-seq samples and cancer samples. While using SRA data was useful to substantially increase the breadth of data used in SpliceVault (imperative when looking for very rare splicing events), the missing metadata and wide range of RNA-seq approaches, single-cell data, and cancer samples, means that 300K-RNA likely does not yet faithfully represent ‘normal’ mis-splicing.

Considering this, the high concordance of top-ranked events seen between SRA and GTEx in Chapter 3 (Figure 3B) is remarkable and encouraging for the robustness of our method. However, SpliceVault will likely be enhanced moving forward by increasing the homogeneity of its component RNA-seq samples and including samples from tissue-types not currently represented in GTEx, such as eye or cochlear (important for Mendelian disorders involving blindness or deafness). Value could also be added by the addition of developmental RNA-seq samples, which could provide key insights on how ‘normal’ mis-splicing changes over time, in different developmental contexts. As RNA-seq is more and more commonly deployed in clinical settings<sup>94,134–139</sup> it’s likely we will not have long to wait for larger and/or deeper RNA-seq control datasets.

*Increasing sequencing depth may be the best way to increase tissue-specific sensitivity*

We show in Chapter 3 that using top-ranked events in GTEx Muscle to predict the mis-splicing events associated with 19 variants associated with muscle disorders (and subject to RNA Diagnostics on muscle samples), showed lower sensitivity than using 300K-RNA Top-

4\* (Extended Data Figure 3E). This, along with the fact that the number of detected mis-splicing events increases with the breadth of RNA-seq samples representing that tissue (Extended Data Figure 3D), indicates that current tissue-specific RNA-seq reference databases are likely insufficiently broad and/or deep to dig into tissue-specific mis-splicing differences. SpliceVault works by using a large breadth of RNA-seq data to improve detection of stochastic leaky events. However, in the future it's likely more feasible and economical to go 'deeper' with higher read-depth on a per-sample basis, to detect these stochastic leaky events more reliably in individual samples.

*Cycloheximide treated specimens may improve sensitivity and alter top-ranked events.*

The sensitivity of SpliceVault for rare out-of-frame events which are degraded by NMD may also be increased by the integration of cycloheximide-treated specimens with NMD inhibited. In addition, cycloheximide-treated specimens may alter Top-4 ranking: In Chapter 3 we found evidence that in-frame events may be enriched in 300K-RNA Top-4, suggesting that out-of-frame events degraded by NMD may be incorrectly ranked lower than their true frequency.

*Interrogation of the alignment of individual splicing events may increase accuracy*

While our cut-off of predicting the Top-4 events in 300K-RNA showed high sensitivity (92%), the low positive predictive value (32%) is an area in need of improvement (Chapter 3, Figure 4E). One possible avenue that could improve SpliceVault's PPV is interrogation of the alignment of individual splicing events.

Short-read RNA-sequencing requires the alignment of short reads (usually ~150 nt) to a reference transcript. Split-reads are especially vulnerable to either multi-mapping (mapped to multiple sites in the genome) or mis-mapping (mapped to the wrong spot in the genome) as they must be aligned to either side of an exon-intron junction which may or may not be annotated in transcript databases. It is therefore entirely possible that certain highly ranked 300K-RNA events are in fact spurious and incorrectly aligned, no matter the number of samples that event is seen in (the RNA-seq aligner could easily and is in fact likely to make the same mistake across different samples, as it runs the same alignment algorithm independently each time). SpliceVault rankings could be further refined by inclusion of

quality metrics and precise read support of individual mis-splicing events, and there's every possibility that both sensitivity and positive predictive values could be improved by this effort.

Of course, this is just one possibility. There are likely many genuine, correctly aligned Top-4 mis-splicing events that are nevertheless not upregulated in the event of a variant at that splice-site. Understanding why some Top-4 mis-splicing events are upregulated in the event of a variant and others are not (notably many double-exon skipping events, Chapter 3 Extended Data Figure 6G) will require further investigation to exclude the caveat of mis-alignment and elucidate a mechanistic explanation such that we can incorporate new evidence-based insight to improve the ranking method.

### *Long-read sequencing may allow better detection of pseudo-exon and intron retention events*

In addition to improving predictions for the mis-splicing events currently encompassed by SpliceVault, future iterations could focus on intron retention and pseudo-exon events which aren't currently represented as they cannot be directly quantified by split-reads in short-read RNA-seq data.

Long-read sequencing, with reads up to 10 kb in length, may be able to encompass retention of long introns and pseudo-exons in a single split-read, and allow integration into our ranking. However there remain many caveats and difficulties with bioinformatic alignment of long-read sequencing data, which will have to be worked through before its full potential as a technology is known. Furthermore, detected intron retention is ambiguous as it can reflect pre-mRNA (a true 'mis-splicing' event) or splicing intermediates (not yet spliced).

Long-read data may also allow us to better differentiate between pseudo-exon events and distal intronic cryptics. In Chapter Three we filter out 'cryptics' further than 600 nt away from authentic splice-sites to improve sensitivity and PPV of 300K-RNA Top-4 (Extended Data Figure 1). In fact, these split-reads may also represent pseudo-exon events, if they occur proximally to another 'cryptic' event associated with the other end of the intron. Long-read RNA-seq may be able to encompass an entire pseudo-exon in a single read and so avoid this ambiguity.

This would also allow the expansion of SpliceVault to the cataloguing of natural ‘leaky’ expression of pseudo-exons. It’s already been shown that pseudo-exon events upregulated in the event of a variant are commonly detected at lower levels in control samples<sup>140</sup>. An expansion of SpliceVault to include a genome-wide catalogue and per-intron ranking of pseudo-exons detected in control samples would open the door for interpretation of many currently intractable deep-intronic variants.

### **4.3 Conclusions**

Until now, it has been impossible to predict the probable nature of variant-induced mis-splicing of the pre-mRNA: a huge challenge for variant interpretation in diagnostic genomics. We’ve shown that simply ranking the most common ‘mistakes’ made by the spliceosome local to each exon-intron junction accurately predicts what happens when a genetic variant disrupts the splice-site at that exon-intron junction: Past behaviour is the best predictor of future behaviour.

A priori knowledge of 300K-RNA Top-4 mis-splicing events has been transformative for our research-led clinical RNA Diagnostics program, facilitating both variant curation and strategic experimental design of RNA assays to specifically target probable mis-splicing events. SpliceVault is a critical advance for diagnostic genomic pathology and is likely to be used extensively by the clinical genetics community as RNA Diagnostics enters routine clinical practice. 300K-RNA will impact variant classification and improve accuracy and completeness of RNA diagnostic testing - to avoid the non-actionable diagnostic endpoint of a splicing variant of uncertain significance.

## References

1. Miko, I. Gregor Mendel and the Principles of Inheritance. *Nature Education* **1**, 134 (2008).
2. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). Online Mendelian Inheritance in Man, OMIM®. World Wide Web URL: <https://omim.org/> (2021).
3. Boycott, K. M. *et al.* International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *Am J Hum Genet* **100**, 695–705 (2017).
4. Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet* **28**, 165–173 (2020).
5. Farnaes, L. *et al.* Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. *npj Genomic Med* **3**, 1–8 (2018).
6. McCandless, S. E., Brunger, J. W. & Cassidy, S. B. The Burden of Genetic Disease on Inpatient Care in a Children’s Hospital. *The American Journal of Human Genetics* **74**, 121–127 (2004).
7. Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic medicine -- progress, pitfalls, and promise. *Cell* **177**, 45–57 (2019).
8. Haendel, M. *et al.* How many rare diseases are there? *Nature Reviews Drug Discovery* **19**, 77–78 (2020).
9. Stranneheim, H. & Wedell, A. Exome and genome sequencing: a revolution for the discovery and diagnosis of monogenic disorders. *Journal of Internal Medicine* **279**, 3–15 (2016).
10. Marshall, C. R. *et al.* The Medical Genome Initiative: moving whole-genome sequencing for rare disease diagnosis to the clinic. *Genome Medicine* **12**, 48 (2020).
11. Scocchia, A. *et al.* Clinical whole genome sequencing as a first-tier test at a resource-limited dysmorphology clinic in Mexico. *NPJ Genom Med* **4**, 5 (2019).
12. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
13. The Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68 (2015).
14. Zurynski, Y. *et al.* Australian children living with rare diseases: experiences of diagnosis and perceived consequences of diagnostic delays. *Orphanet Journal of Rare Diseases* **12**, 68 (2017).
15. Marshall, D. A. *et al.* The value of diagnostic testing for parents of children with rare genetic diseases. *Genetics in Medicine* **21**, 2798–2806 (2019).
16. Wright, C. F., FitzPatrick, D. R. & Firth, H. V. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet* **19**, 253–268 (2018).
17. Liu, Z., Zhu, L., Roberts, R. & Tong, W. Toward Clinical Implementation of Next-Generation Sequencing-Based Genetic Testing in Rare Diseases: Where Are We? *Trends in Genetics* **35**, 852–867 (2019).
18. Rehm, H. L. & Fowler, D. M. Keeping up with the genomes: scaling genomic variant interpretation. *Genome Medicine* **12**, 5 (2019).
19. Clift, K. *et al.* Patients’ views on variants of uncertain significance across indications. *J Community Genet* **11**, 139–145 (2020).
20. Wai, H. A. *et al.* Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet Med* **22**, 1005–1014 (2020).



21. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062–D1067 (2018).
22. Mattick, J. S., Dinger, M., Schonrock, N. & Cowley, M. Whole genome sequencing provides better diagnostic yield and future value than whole exome sequencing. *Med J Aust* **209**, 197–199 (2018).
23. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. *New England Journal of Medicine* **385**, 1868–1880 (2021).
24. Shieh, J. T. *et al.* Application of full-genome analysis to diagnose rare monogenic disorders. *npj Genom. Med.* **6**, 1–10 (2021).
25. He, W.-B. *et al.* RNA splicing analysis contributes to reclassifying variants of uncertain significance and improves the diagnosis of monogenic disorders. *J Med Genet* jmedgenet-2021-108013 (2022) doi:10.1136/jmedgenet-2021-108013.
26. Lord, J. & Baralle, D. Splicing in the Diagnosis of Rare Disease: Advances and Challenges. *Front Genet* **12**, 689892 (2021).
27. Wilkinson, M. E., Charenton, C. & Nagai, K. RNA Splicing by the Spliceosome. *Annu. Rev. Biochem.* **89**, annurev-biochem-091719-064225 (2020).
28. Clancy, S. & Brown, W. Translation: DNA to mRNA to Protein. *Nature Education* **1**, 101 (2008).
29. Gooding, C. *et al.* [No title found]. *Genome Biol* **7**, R1 (2006).
30. Wimmer, K. *et al.* AG-exclusion zone revisited: Lessons to learn from 91 intronic NF1 3' splice site mutations outside the canonical AG-dinucleotides. *Human Mutation* humu.24005 (2020) doi:10.1002/humu.24005.
31. Corvelo, A., Hallegger, M., Smith, C. W. J. & Eyras, E. Genome-Wide Association between Branch Point Properties and Alternative Splicing. *PLoS Comput Biol* **6**, (2010).
32. Fu, X.-D. & Ares, M. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet* **15**, 689–701 (2014).
33. Jolma, A. *et al.* Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. *Genome Res.* **30**, 962–973 (2020).
34. Dominguez, D. *et al.* Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell* **70**, 854-867.e9 (2018).
35. Giudice, G., Sánchez-Cabo, F., Torroja, C. & Lara-Pezzi, E. ATtRACT—a database of RNA-binding proteins and associated motifs. *Database (Oxford)* **2016**, (2016).
36. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
37. Cook, K. B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Research* **39**, D301–D308 (2011).
38. Piva, F., Giulietti, M., Burini, A. B. & Principato, G. SpliceAid 2: A database of human splicing factors expression data and RNA target motifs. *Human Mutation* **33**, 81–85 (2012).
39. Baeza-Centurion, P., Miñana, B., Schmiedel, J. M., Valcárcel, J. & Lehner, B. Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell* **176**, 549-563.e23 (2019).
40. Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J. & Lehner, B. The complete local genotype–phenotype landscape for the alternative splicing of a human exon. *Nature Communications* **7**, 11558 (2016).
41. Baeza-Centurion, P., Miñana, B., Valcárcel, J. & Lehner, B. Mutations primarily alter the inclusion of alternatively spliced exons. *Elife* **9**, e59959 (2020).
42. Ke, S. *et al.* Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Res* **28**, 11–24 (2018).

## References

43. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell* **163**, 698–711 (2015).
44. Mueller, W. F., Larsen, L. S. Z., Garibaldi, A., Hatfield, G. W. & Hertel, K. J. The Silent Sway of Splicing by Synonymous Substitutions. *J Biol Chem* **290**, 27700–27711 (2015).
45. Wang, Y., Ma, M., Xiao, X. & Wang, Z. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat Struct Mol Biol* **19**, 1044–1052 (2012).
46. Wang, Y. *et al.* A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nat Struct Mol Biol* **20**, 36–45 (2013).
47. Ke, S. *et al.* Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res* **21**, 1360–1374 (2011).
48. Culler, S. J., Hoff, K. G., Voelker, R. B., Berglund, J. A. & Smolke, C. D. Functional selection and systematic analysis of intronic splicing elements identify active sequence motifs and associated splicing factors. *Nucleic Acids Res* **38**, 5152–5165 (2010).
49. Wang, Z. *et al.* Systematic Identification and Analysis of Exonic Splicing Silencers. *Cell* **119**, 831–845 (2004).
50. Erkelenz, S. *et al.* Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. *Nucleic Acids Res* **42**, 10681–10697 (2014).
51. Lim, K. H., Ferraris, L., Filloux, M. E., Raphael, B. J. & Fairbrother, W. G. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proceedings of the National Academy of Sciences* **108**, 11093–11098 (2011).
52. Zhang, C., Li, W.-H., Krainer, A. R. & Zhang, M. Q. RNA landscape of evolution for optimal exon and intron discrimination. *Proceedings of the National Academy of Sciences* **105**, 5797–5802 (2008).
53. Voelker, R. B. & Berglund, J. A. A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. *Genome Res* **17**, 1023–1033 (2007).
54. Goren, A. *et al.* Comparative Analysis Identifies Exonic Splicing Regulatory Sequences—The Complex Definition of Enhancers and Silencers. *Molecular Cell* **22**, 769–781 (2006).
55. Zhang, X. H. F., Leslie, C. S. & Chasin, L. A. Dichotomous splicing signals in exon flanks. *Genome Res* **15**, 768–779 (2005).
56. Fairbrother, W. G. Predictive Identification of Exonic Splicing Enhancers in Human Genes. *Science* **297**, 1007–1013 (2002).
57. Wang, Y. *et al.* Mechanism of alternative splicing and its regulation. *Biomedical Reports* **3**, 152–158 (2015).
58. Kornblihtt, A. R. *et al.* Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol* **14**, 153–165 (2013).
59. Wang, Z. & Burge, C. B. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802–813 (2008).
60. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *The American Journal of Human Genetics* **102**, 11–26 (2018).
61. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
62. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).

63. Vaquero-Garcia, J. *et al.* A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife* **5**, e11752 (2016).
64. Nellore, A. *et al.* Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biology* **17**, 266 (2016).
65. Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *PLoS Genet* **6**, e1001236 (2010).
66. Kurosaki, T., Popp, M. W. & Maquat, L. E. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. *Nat Rev Mol Cell Biol* **20**, 406–420 (2019).
67. Nickless, A., Bailis, J. M. & You, Z. Control of gene expression through the nonsense-mediated RNA decay pathway. *Cell & Bioscience* **7**, 26 (2017).
68. Linde, L., Boelz, S., Neu-Yilik, G., Kulozik, A. E. & Kerem, B. The efficiency of nonsense-mediated mRNA decay is an inherent character and varies among different cells. *Eur J Hum Genet* **15**, 1156–1162 (2007).
69. Sato, H. & Singer, R. H. Cellular variability of nonsense-mediated mRNA decay. *Nat Commun* **12**, 7203 (2021).
70. Rivas, M. A. *et al.* Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**, 666–669 (2015).
71. Khajavi, M., Inoue, K. & Lupski, J. R. Nonsense-mediated mRNA decay modulates clinical outcome of genetic disease. *Eur J Hum Genet* **14**, 1074–1081 (2006).
72. Coban-Akdemir, Z. *et al.* Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. *Am J Hum Genet* **103**, 171–187 (2018).
73. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Research* **49**, D884–D891 (2021).
74. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–745 (2016).
75. Wai, H., Douglas, A. G. L. & Baralle, D. RNA splicing analysis in genomic medicine. *The International Journal of Biochemistry & Cell Biology* **108**, 61–71 (2019).
76. Riolo, G., Cantara, S. & Ricci, C. What’s Wrong in a Jump? Prediction and Validation of Splice Site Variants. *Methods and Protocols* **4**, 62 (2021).
77. Cartegni, L. & Krainer, A. R. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat Genet* **30**, 377–384 (2002).
78. Niksic, M., Romano, M., Buratti, E., Pagani, F. & Baralle, F. E. Functional analysis of cis-acting elements regulating the alternative splicing of human CFTR exon 9. *Hum Mol Genet* **8**, 2339–2349 (1999).
79. Sylvester, B. *et al.* A Synonymous Exonic Splice Silencer Variant in IRF6 as a Novel and Cryptic Cause of Non-Syndromic Cleft Lip and Palate. *Genes (Basel)* **11**, (2020).
80. Truty, R. *et al.* Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation. *The American Journal of Human Genetics* **108**, 696–708 (2021).
81. Sanders, S. J., Schwartz, G. B. & Farh, K. K.-H. Clinical impact of splicing in neurodevelopmental disorders. *Genome Med* **12**, (2020).
82. Richards, S. *et al.* 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**, 405–424 (2015).
83. Abou Tayoun, A. N. *et al.* Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum Mutat* **39**, 1517–1524 (2018).

84. Broomfield, J., Hill, M., Guglieri, M., Crowther, M. & Abrams, K. Life Expectancy in Duchenne Muscular Dystrophy: Reproduced Individual Patient Data Meta-analysis. *Neurology* **97**, e2304–e2314 (2021).
85. Jefferies, J. L. *et al.* Genetic Predictors and Remodeling of Dilated Cardiomyopathy in Muscular Dystrophy. *Circulation* **112**, 2799–2804 (2005).
86. Yang, Y. *et al.* Comprehensive genetic diagnosis of patients with Duchenne/Becker muscular dystrophy (DMD/BMD) and pathogenicity analysis of splice site variants in the DMD gene. *J Zhejiang Univ Sci B* **20**, 753–765 (2019).
87. Zhang, X. *et al.* Functional analysis of variants in DMD exon/intron 10 predicted to affect splicing. *J Hum Genet* (2022) doi:10.1038/s10038-022-01035-y.
88. Tuffery-Giraud, S. *et al.* Genotype-phenotype analysis in 2,405 patients with a dystrophinopathy using the UMD-DMD database: a model of nationwide knowledgebase. *Hum Mutat* **30**, 934–945 (2009).
89. Bournazos, A. M. *et al.* Standardized practices for RNA diagnostics using clinically accessible specimens reclassifies 75% of putative splicing variants. *Genetics in Medicine* (2021) doi:10.1016/j.gim.2021.09.001.
90. Arrabal, L. *et al.* Genotype–phenotype correlations in sepiapterin reductase deficiency. A splicing defect accounts for a new phenotypic variant. *Neurogenetics* **12**, 183–191 (2011).
91. Lord, J. *et al.* Pathogenicity and selective constraint on variation near splice sites. *Genome Res.* **29**, 159–170 (2019).
92. Rowlands, C. *et al.* Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. *Sci Rep* **11**, 20607 (2021).
93. Katz, Y. *et al.* Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics* **31**, 2400–2402 (2015).
94. Yepez, V. A. *et al.* Clinical implementation of RNA sequencing for Mendelian disease diagnostics. *medRxiv* 2021.04.01.21254633 (2021) doi:10.1101/2021.04.01.21254633.
95. Zeng, W. & Mortazavi, A. Technical considerations for functional sequencing assays. *Nat Immunol* **13**, 802–807 (2012).
96. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).
97. Douglas, A. G. L. & Baralle, D. Translating RNA splicing analysis into diagnosis and therapy. *OBM Genetics* **5**, (2021).
98. Chhangawala, S., Rudy, G., Mason, C. E. & Rosenfeld, J. A. The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biology* **16**, 131 (2015).
99. Mapleson, D., Venturini, L., Kaithakottil, G. & Swarbreck, D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *Gigascience* **7**, (2018).
100. Ballouz, S., Dobin, A., Gingeras, T. R. & Gillis, J. The fractured landscape of RNA-seq alignment: the default in our STARS. *Nucleic Acids Research* **46**, 5125–5138 (2018).
101. Dobin, A. & Gingeras, T. R. Mapping RNA-seq Reads with STAR. *Curr Protoc Bioinformatics* **51**, 11.14.1–11.14.19 (2015).
102. Whitley, S. K., Horne, W. T. & Kolls, J. K. Research Techniques Made Simple: Methodology and Clinical Applications of RNA Sequencing. *Journal of Investigative Dermatology* **136**, e77–e82 (2016).
103. Nonis, A., De Nardi, B. & Nonis, A. Choosing between RT-qPCR and RNA-seq: a back-of-the-envelope estimate towards the definition of the break-even-point. *Anal Bioanal Chem* **406**, 3533–3536 (2014).

## References

104. Rowlands, C. F., Baralle, D. & Ellingford, J. M. Machine Learning Approaches for the Prioritization of Genomic Variants Impacting Pre-mRNA Splicing. *Cells* **8**, 1513 (2019).
105. Shapiro, M. B. & Senapathy, P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res* **15**, 7155–7174 (1987).
106. Reese, M. G., Eeckman, F. H., Kulp, D. & Haussler, D. Improved Splice Site Detection in Genie. *Journal of Computational Biology* **4**, 311–323 (1997).
107. Pertea, M. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Research* **29**, 1185–1190 (2001).
108. Yeo, G. & Burge, C. B. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. **11**, 377–394 (2004).
109. Verma, B., Akinyi, M. V., Norppa, A. J. & Frilander, M. J. Minor spliceosome and disease. *Seminars in Cell & Developmental Biology* **79**, 103–112 (2018).
110. Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Research* **37**, e67–e67 (2009).
111. Leman, R. *et al.* Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: an international collaborative effort. *Nucleic Acids Res* **46**, 7913–7923 (2018).
112. Aalberts, D. P., Daub, E. G. & Dill, J. W. Quantifying optimal accuracy of local primary sequence bioinformatics methods. *Bioinformatics* **21**, 3347–3351 (2005).
113. Lear, A. L., Eperon, L. P., Wheatley, I. M. & Eperon, I. C. Hierarchy for 5' splice site preference determined in vivo. *J Mol Biol* **211**, 103–115 (1990).
114. Conboy, J. G. Unannotated splicing regulatory elements in deep intron space. *Wiley Interdiscip Rev RNA* **12**, e1656 (2021).
115. Xiong, H. Y. *et al.* The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
116. Cheng, J. *et al.* MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol* **20**, 48 (2019).
117. Adamson, S. I., Zhan, L. & Graveley, B. R. Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biol* **19**, (2018).
118. Cheng, J., Celik, M. H., Kundaje, A. & Gagneur, J. MTSplice predicts effects of genetic variants on tissue-specific splicing. (2020) doi:10.1101/2020.06.07.138453.
119. Mort, M. *et al.* MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol* **15**, R19 (2014).
120. Gelfman, S. *et al.* Annotating pathogenic non-coding variants in genic regions. *Nat Commun* **8**, 236 (2017).
121. Jagadeesh, K. A. *et al.* S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nat Genet* **51**, 755–763 (2019).
122. Danis, D. *et al.* Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *bioRxiv* 2021.01.28.428499 (2021) doi:10.1101/2021.01.28.428499.
123. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21**, 577–581 (2003).
124. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).
125. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using Pangolin. 23 (2021).

126. Vihinen, M. Problems in variation interpretation guidelines and in their implementation in computational tools. *Mol Genet Genomic Med* (2020) doi:10.1002/mgg3.1206.
127. Riepe, T. V., Khan, M., Roosing, S., Cremers, F. P. M. & 't Hoen, P. A. C. Benchmarking deep learning splice prediction tools using functional splice assays. *Human Mutation* **42**, 799–810 (2021).
128. Rehm, H. L. *et al.* ClinGen — The Clinical Genome Resource. *N Engl J Med* **372**, 2235–2242 (2015).
129. Pejaver, V. *et al.* Evidence-based calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for clinical use of PP3/BP4 criteria. <http://biorxiv.org/lookup/doi/10.1101/2022.03.17.484479> (2022) doi:10.1101/2022.03.17.484479.
130. Finkel, R. S. *et al.* Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy. *New England Journal of Medicine* **377**, 1723–1732 (2017).
131. Rossor, A. M., Reilly, M. M. & Sleight, J. N. Antisense oligonucleotides and other genetic therapies made simple. *Practical Neurology* **18**, 126–131 (2018).
132. Rinaldi, C. & Wood, M. J. A. Antisense oligonucleotides: the next frontier for treatment of neurological disorders. *Nat Rev Neurol* **14**, 9–21 (2018).
133. Gelfman, S., Cohen, N., Yearim, A. & Ast, G. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. *Genome Res* **23**, 789–799 (2013).
134. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science Translational Medicine* **9**, (2017).
135. Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat Commun* **8**, (2017).
136. Frésard, L. *et al.* Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nature Medicine* **25**, 911–919 (2019).
137. Lee, H. *et al.* Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genetics in Medicine* **22**, 490–499 (2020).
138. Murdock, D. R. *et al.* Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. *Journal of Clinical Investigation* (2020) doi:10.1172/JCI141500.
139. Gonorazky, H. D. *et al.* Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am J Hum Genet* **104**, 466–483 (2019).
140. Petersen, U. S. S., Doktor, T. K. & Andresen, B. S. Pseudoexon activation in disease by non-splice site deep intronic sequence variation - wild type pseudoexons constitute high-risk sites in the human genome. *Hum Mutat* (2021) doi:10.1002/humu.24306.