RNA sequencing uplifts diagnostic rate in undiagnosed rare disease patients

Carolina Jaramillo Oquendo,[1] Htoo A Wai,[1] Will Rich[1], David J. Bunyan,[2] N. Simon Thomas[2], David Hunt,[1,3], Jenny Lord,[1] Andrew G L Douglas,[1,4] Diana Baralle[1,3*]

1. Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK

2. Wessex Genomics Laboratory Service, Salisbury District Hospital, Salisbury, UK

3. Wessex Clinical Genetics Service, University Hospital Southampton NHS Foundation Trust, Southampton, UK

4. Oxford Centre for Genomic Medicine, Oxford University Hospitals NHS Foundation Trust, Oxford, UK

*Correspondence: d.baralle@soton.ac.uk

Abstract

RNA-sequencing is increasingly being used as a complementary tool to DNA sequencing in diagnostics where DNA analysis has been uninformative as it enables the identification of alternative splicing (AS), aberrant gene expression and allele specific expression. Using RNA from patient blood, we have used both RNA-sequencing and RT-PCR to detect splicing and gene expression outliers in a heterogenous cohort of 87 patients with suspected Mendelian disorders, 38% of which did not have a candidate sequence variant. Expression outliers were detected using OUTRIDER, and we compared the performance of multiple open-source alternative splicing tools (MAJIQ, rMATS-turbo, and LeafCutterMD) in identifying alternative splicing events. As well as clarifying the impact of variants of uncertain significance (VUSs), we trialled two novel approaches to identify new potential diagnoses in patients with no candidate variants. We were able to assess 85% of VUSs and validate splicing abnormalities in 18/48 patients with a VUS. Furthermore, we identified four new diagnoses by detecting

novel AS events in patients with no candidate sequence variants from prior genomic DNA testing (n=33) or those in which the candidate VUS did not affect splicing (n=23) and identified one additional diagnosis through detection of skewed X-inactivation. These results gave an overall uplift in diagnostic yield of 26%. In this work we demonstrate the utility of blood-based RNA analysis in improving diagnostic yields and highlight optimal approaches for such analysis.

Introduction

With the advancement of next generation sequencing, vast amounts of DNA sequencing data are continually generated to aid in the diagnosis and treatment of rare diseases. However, our ability to interpret genomic data has not grown at the same rate. The diagnostic yield of whole exome and genome sequencing alone remains relatively low leaving good scope for diagnostic rates to be improved[1–4]. Within the Genomics England 100,000 Genomes Project for example, the average diagnostic yield using whole genome sequencing is around 25%[3]. RNA sequencing (RNA-seq) is now being used as a complementary tool to DNA sequencing for diagnostic genetic testing in rare disease where DNA analysis alone has failed to identify a clear diagnosis[5–15]. While some studies have focused on specific disorder types, such as mitochondrial disease[6,12], muscle disorders[5], and neurodevelopmental disorders[13], others have looked at heterogeneous disease populations[7,11,14–16]. Unlike DNA sequencing, RNA-seq is both a qualitative and quantitative approach which allows identification of abnormal alternative splicing (AS), aberrant gene expression, and mono allelic expression, allowing improved interpretation of variants of uncertain significance (VUSs). An important benefit of the transcriptomic approach as compared to targeted reverse transcription PCR (RT-PCR) is that it is agnostic to the resulting abnormally spliced transcript, whereas RT-PCR must rely on targeted primer

designs that are intrinsically limited by factors such as known gene annotations and PCR amplicon lengths. RNA-seq therefore provides the opportunity not only to look at the splicing effects of known VUSs but also to scan the transcriptome for abnormal splicing events and expression abnormalities in other relevant genes that may be the cause of a patient's phenotype. This in turn allows the identification of molecular diagnoses in patients in which standard genomic DNA testing has not identified any candidate.

Whilst previous studies have illustrated a variety of ways in which RNA-seq can be used to detect clinically relevant splicing abnormalities, few have directly compared the results of RNA-seq to those obtained by RT-PCR. Such a comparison is of vital importance in terms of clarifying the clinical utility of RNA-seq in a diagnostic setting and in defining instances of when one or other technique may be of superior sensitivity.

In this study, we have used both RNA-seq and RT-PCR to 1) examine the splicing effects of clinically relevant VUSs in a heterogeneous cohort of patients with suspected Mendelian disorders; 2) assess the use of blood as the tissue of choice in the implementation of an RNA-seq clinical pipeline to improve diagnostic yield of patients with rare diseases; 3) compare open-source tools and in-house pipelines to detect AS and gene expression outliers in RNA-seq; and 4) investigate the feasibility of generating new diagnostic candidates using these tools in a subset of patients with no previously reported candidate VUSs in clinically relevant genes.

Material and Methods

Sample collection and preparation

Participants were enrolled into the University of Southampton's Splicing and Disease study with appropriate ethical approval (REC 11/SC/0269, IRAS 49685, ERGO 23056). This cohort of individuals comprises a combination of rare disease patients assessed by UK clinical

genetics services in whom a candidate VUS may or may not have been identified through conventional DNA-based testing (n=87). Within this cohort, 48 individuals had pre-existing candidate VUSs that had been previously clinically reported within genes of potential clinical relevance, six had previously confirmed known genetic diagnoses and 33 had unknown molecular diagnoses with no previously reported candidate VUSs in clinically relevant genes. Individuals without molecular diagnosis had a phenotype where a genetic cause was suspected. A number of samples (n=18) in this cohort have been previously reported[10]. RNA was extracted from 87 blood samples collected in PAXgene blood RNA tubes using the PAXgene blood RNA Kit (PreAnalytiX, Switzerland). Quantification was performed by NanoDrop spectrophotometer and Qubit fluorometer (Thermo, MA) and RNA integrity number assessed using an Agilent 2100 Bioanalyzer (Agilent, CA).

Data generation

Reverse transcription polymerase chain reaction (RT-PCR) analysis

Primers for variant of interest were designed to span at least three exons and where possible, up to seven exons. The cDNA was synthesised using High-Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific, USA). PCR was performed using GoTaq G2 DNA polymerase kit (Promega, USA). The PCR products were analysed in a 1% agarose gel prepared with Nancy-520 DNA gel stain (Sigma, USA). Subsequently, the PCR products were purified using the GeneJET PCR Purification Kit (Thermo Fisher Scientific, USA) and bidirectional Sanger sequencing was carried out by SourceBioscience. PCR experiments were repeated twice for reproducibility.

RNA sequencing

RNA samples were sequenced via Novogene (Hong Kong) in four separate batches (comprising 7, 16, 33 and 31 samples) using a total RNA-seq approach employing the

NEBNext rRNA Depletion Kit and the NEBNext Ultra Directional RNA Library Prep Kit (New England Biolabs, MA). Samples in batches 1, 2 and 4 also had NEBNext Globin Depletion Kit applied, whereas those in batch 3 did not. On average 76 million 150 base-pair paired-end reads were generated for each sample on a HiSeq 2000 instrument (Illumina, CA). FASTQ files underwent initial quality control filtering and adapter sequence removal by Novogene. Subsequent alignment was performed to the human genome reference (GRCh38) with annotations from GENCODE[17] release 38 using STAR aligner[18] v2.6.1c with optimised parameters via the University of Southampton's IRIDIS5 high-performance computing clusters. Scripts can be found in GitHub [https://github.com/carojoquendo/RNA_splicing_and_disease].

The MRSD web portal (https://mcgm-mrsd.github.io/) was used to predict the minimum number of sequencing reads required from RNA-seq experiments to confidently determine aberrant splicing events for a gene of interest[19]. Default values for confidence level (95%), read coverage (n=8) and splice junction proportion (75%) were used. The online SpliceAI server (https://spliceailookup.broadinstitute.org/) was used to predict splicing effect of all variants.

Evaluation of functional consequence

To determine the functional consequence at a transcript level for each VUS, RNA-seq data was loaded into the Integrative Genomics Viewer (IGV)[20] and each variant was visually inspected to search for splicing abnormalities. If there were no splicing abnormalities in the exons flanking the variant, it was determined that there were no splicing abnormalities resulting from the variant. Splicing abnormalities were classed as: exon skipping, inclusion of pseudoexon, intron retention, alternative 5′ (donor) splice site and alternative 3′ (acceptor)

splice site. The command-line tool ggsashimi[21] was used to create final sashimi plots to visualise junctions.

Expression outlier detection

Salmon[26] was used to quantify gene and transcript counts in mapping-based mode. Transcriptome indices for Salmon were generated using the GRCh38 genome and transcriptome reference from GENCODE release 38 (https://combine-lab.github.io/alevin-tutorial/2019/selective-alignment/). The R[27] package tximport[28] was used to collate and import raw read counts from all samples to be used as input into OUTRIDER[29]. The function filterExpression was used to remove genes that had low Fragments Per Kilobase of transcript per Million mapped reads (FPKM) expression values followed by the OUTRIDER function which ran the full OUTRIDER pipeline.

Detection of aberrant splicing

rMATS-turbo v4.1.2[30], MAJIQ v2.4[31] and LeafCutterMD[32] were used to detect aberrant splicing across all samples. For each tool, each sample was compared against other samples within the same batch with the exception of samples in batch 1 and 2 which were combined to increase power. rMATS-turbo was run with additional parameters --novelSSS to enable detection of novel splice sites, as well as --allow-clipping to allow alignments with soft or hard clipping to be used. MAJIQ modules build and DeltaPSI were run with default parameters using the GENCODE v38 annotation gff3 files. The DeltaPSI results were then input into the voila module which provides a tab-delimited text file to allow parsing of the MAJIQ result and filters out local splice variations (LSVs) with no junctions predicted to change over a certain value. Default parameters for the voila module were used. LeafCutterMD was also run with default parameters and results were annotated with gene symbols to extract genes and loci of interest for each sample.

Variant calling

GATK's Best Practices workflow for RNA-seq short variant discovery was followed to identify variants in the RNA-seq. First duplicate reads were marked by Picard's version 2.18.14 (http://broadinstitute.github.io/picard) MarkDuplicates function, followed by reformatting of the BAM files for HaplotypeCaller with GATK's (version 4.2.2)[33] SplitNCigarReads and Picard's AddOrReplaceReadGroups. The next step was Base Quality Recalibration, consisting of two tools: GATK's BaseRecalibrator and ApplyBQSR. Lastly GATK's HaplotypeCaller was used to call variants and write to VCF files.

To reduce spurious calls, VCF files were run through GATK's VariantFiltration tool, keeping calls with a minimum quality score of 50. bcftools[34] was used to further filter variants excluding any variants with a) less than eight reads covering the locus; b) calls with genotype quality lower than 16; c) calls with strand bias (FS metric) greater than 30; and d) variants with a quality normalised by depth of at least two[35]. After filtering, Ensembl's VEP[36] (version 103) was used to annotate the variants with additional information including but not limited to nearest gene, variant consequence (e.g. missense, splice_region) and minor allele frequency (MAF). The SpliceAI VEP plugin was used to produce a score per variant (delta score) based on the likelihood of the variant impacting splicing. SpliceAI scores range from 0-1 with scores closer to 1 being more likely to affect splicing. VCF files were further filtered to keep variants with a) population frequency less than 0.005; b) variants with a SpliceAI score ≥ 0.2; c) variants found in protein coding genes; and d) single nucleotide variants. Indels were not included as the majority had very poor quality and inclusion of indels introduced a significant number of false positives.

Filtering strategy for cases with no VUS

To find diagnoses in cases without a VUS or gene of interest, two separate strategies were used. The first strategy was a genotype to phenotype approach. Using the annotated and filtered VCF files a BED file was created with variant location, adding 25 base pairs up and downstream of the variant [chromosome start(-25bp) end(+25bp) gene]. Results from all three splicing tools were also converted into BED format. After sorting the BED files, the variant BED was overlapped with the rMATS-turbo, MAJIQ and LeafCutter results BED using bedtools (v2.30)[37] intersect keeping only overlapping features. For each sample there was a resulting overlap BED file corresponding to each splicing tool. Each genomic location in all the overlap files was then inspected in IGV.

The second strategy involved using genotype information available to filter results from splicing tools. To do this, appropriate panels from the UK Genomic Medicine Service (GMS) PanelApp[24] resource were applied to the splicing tools results and each AS event was also inspected in IGV.

Results

Summary of RNA-sequencing data outputs

The mean number of sequencing reads per sample was 76.6 million (61.3-130.2 million) and on average 80% of reads were uniquely mapping (**Figure S1**). However, three of the samples (SOT274, SOT283 and SOT150) were identified as outliers with less than 50% of reads mapping uniquely (**Figure S1**). This could indicate issues with sample preparation or sequencing (e.g. contamination, low concentration or quality of RNA) and could impact coverage for these samples. The mean number of splicing junctions identified across samples was 398,718 (303,637-621,161). Spearman's rank correlation between observed median TPM values and median TPM values found in the Genotype-Tissue Expression (GTEx)

portal[38] was 0.79 with a p-value < 0.001 (**Figure S1**) . When considering disease genes from the Online Mendelian Inheritance in Man (OMIM) database[39] and the UK Genomic Medicine Service's PanelApp resource, 67% (n=11,134) and 73% (n=2,574) of genes were expressed in blood respectively (TPM >1 in at least 4 samples).

As mentioned previously, globin depletion was not applied to one of the batches (batch 3). Analysis of the transcriptomic profiles showed this difference in targeting methodology as samples within batch 3 clustered together in principal component analysis as well as hierarchical clustering (**Figure S1**). Furthermore, median TPM values for the most abundant haemoglobin genes were in line with values reported in GTEx, which also did not utilise globin depletion. To avoid bias due to differences in sequencing methodology, samples were run in separate batches through the splicing tools and OUTRIDER was able to normalise the samples to eliminate this batch effect. TPM values across genes which had VUSs within our cohort were also assessed (**Figure S2**), which showed that gene coverage in genes of interest was not negatively affected by the lack of globin depletion and in fact in some cases, the TPM values were higher in batch 3. A possible explanation for the slight increase of reads in the non-depleted batch is that the globin depletion step could also be reducing the reads for some non-haemoglobin genes as well[40].

Splicing analysis in patients with a candidate VUS

48 of the 87 cases had VUSs in 36 different genes (**Figure S2**). Using default parameters, the MRSD tool predicted that we would only be able to assess 42% (n=15) of these genes in blood based on our mean number of sequencing reads (**Figure S2**). However, we found this tool to be overly conservative as we were able to assess 69% of genes (n=25) using RNA-seq and 83% of genes (n=30) using RT-PCR across cases with a VUS. Using the GTEx dataset as a reference[41], median TPM values across the 25 genes ranged from 0.89 - 73.24 with a mean

and median of 19.43 and 11.16 respectively. The experimental median TPM values in our sequencing data ranged from 2.142 – 75.896 with a mean and median of 15.718 and 7.637 respectively.

Visual inspection of the BAM files in IGV allowed the detection of alternative splicing in 29% (n=14) of samples with candidate VUSs (**Table 1**). Of these splice-altering VUSs, 13/14 were predicted to affect splicing according to SpliceAI (Δ score ≥ 0.2) and all 14 were validated via RT-PCR. The gene with the lowest median TPM value in GTEx and for which we were able to detect alternative splicing was *KAT6B* with a GTEx TPM of 0.890 and an experimental median TPM value of 9.79. For those genes that did not have enough coverage to be assessed properly, the range of TPM values according to GTEx were 0.00 – 4.91 with a mean and median of 0.49 and 0.08 respectively. Out of the 38 VUSs where alternative splicing events were not detected, 15 variants could not be assessed due to insufficient gene coverage. However, RT-PCR was able to validate aberrant splicing in 4 additional cases out of the 15 with low coverage in the RNA-seq data. Two of the four VUSs fell within the *TERT* gene, one fell within the *PRG4* gene, and the fourth variant fell in the *TAOK1* gene. All three genes, *TERT, PRG4,* and *TAOK1* have low expression in blood with a median GTEx TPM value of 0.045, 0.196, and 4.905 respectively. Details of all assessed variants can be found in **Table S1**. Cases (n=23) which showed no detectable splicing abnormalities linked to the VUS were analysed subsequently as unknown cases.

***Table 1**. Variants of unknown significance (VUSs) for which alternative splicing (AS) was observed in IGV. SpliceAI Lookup scores are indicated stating if donor loss (DL), donor gain (DG), acceptor loss (AL) or acceptor gain (AG) is predicted. Events that were identified by the splicing tools but had an adjusted p-value > 0.05 are denoted with an asterisk (*).*

Gene	Variant of unknown significance (VUS)	SpliceAI prediction (type|Δ score|pre-mRNA pos)	Observed splicing abnormality	Tools that identified aberrant splicing event	Median TPM blood (GTEx)
SF3B4	NM_005850.5:c.417C>T Previously reported[10]	DG|0.37|2; AG|0.18|-124	Novel splice donor and acceptor sites in exon 3. r.416_540del, p.(Asp140LeufsTer3)	None	46.190
MED13L	NM_015335.4:c.2570-4_2574del Previously reported[10]	AL|0.99|5; AG|0.98|-4	Alternative splice acceptor site in exon 15. r.2570_2578del, p.(Thr857_Asp860delinsAsn)	rMATS, MAJIQ and LeafCutterMD	5.887
DKC1	NM_001363.5:c.915+10 G>A Previously reported[10]	DG|0.87|1; DL|0.02|-10	Novel splice donor site in intron 9, r.915_916ins915+1_915+11, p.(Asn307SerfsTer3)	rMATS	7.531
NF1	NM_001042492.2:c.1168_1179del12 Previously reported[10]	DL|0.04|18	Skipping of exon 10 r.1063_1185del, p.(Asp355_Lys39	rMATS	1.673

Gene	Variant of unknown significance (VUS)	SpliceAI prediction (type|Δ score|pre-mRNA pos)	Observed splicing abnormality	Tools that identified aberrant splicing event	Median TPM blood (GTEx)
			5del)		
NF1	NM_001042492.2:c.7895A>G Previously reported[10]	DG|0.43|-1; AL|0.12|-25; AG|0.01|-83	Skipping of exon 54. r.7870_7970del, p.(Thr2625Ter)	rMATS	1.673
P3H1	NM_022356.4:c.1224-80G>A Previously reported[10]	DG|0.62|3	Novel splice acceptor site created within intron 7, r.1223_1224ins1223+1_1223+92, p.(Ser409Ter), r.1223_1224ins1223+1_1223+92, r.1224_1228del, p.(Ser409Ter), r.1223_1224ins1223+1_1223+92, r.1224_1240del, p.(Ser409Ter)	rMATS	9.066
TSC2	NM_000548.5:c.4492A>C Previously reported[10]	DL|0.41|1; DG|0.23|-253	Activation of cryptic splice donor site within exon 34, r.4240_4493del, p.(Val1414PhefsTer24)	rMATS and MAJIQ	14.030
UBR4	NM_020765.3:c.8488+3A>G	DL|0.26|3; DG|0.13|-117; AG|0.03|430	Retention of intron 57, r.8488_8489ins8488+1_8489-1,	none	10.800

Gene	Variant of unknown significance (VUS)	SpliceAI prediction (type|Δ score|pre-mRNA pos)	Observed splicing abnormality	Tools that identified aberrant splicing event	Median TPM blood (GTEx)
			p.(Ser2831ArgfsTer23)		
SMARCE1	NM_003079.5:c.8-4A>G	AL|0.22|-4; AG|0.48|-1	Alternative 3' splice acceptor site within intron 2, r.7_8ins8-3_8-1 p.(Lys3delinsThrGlu)	none	7.107
EFTUD2	NM_004247.4:c.702+5G>A	DL|0.93|5; DG|0.13|-73	Skipping of exon 9, r.620_702del, p.(His208AspfsTer26)	rMATS	17.770
ARID1A	NM_006015.6:c.3198G>A	DG|0.25|-65; DL|0.03|0	Skipping of exon 11, r.2989_3198del, p.(Lys997_Gln1066del)	rMATS	11.160
KAT6B	NM_012330.4:c.2629+5G>A	DL|0.98|-5; DG|0.02|-37	Skipping of exon 13 r.2536_2629del, p.(Glu846AlafsTer71)	rMATS*	0.891
PHF8	NM_015107.3:c.784-2A>G	AL|0.99|-2; AG|0.27|-11	Skipping of exon 8 and skipping of exons 7and 8; r.784_946del, p.(Glu263GlyfsTer6); r.597_946del, p.(Leu200ValfsTe	rMATS, MAJIQ and LeafCutterMD	8.11

Gene	Variant of unknown significance (VUS)	SpliceAI prediction (type|Δ score|pre-mRNA pos)	Observed splicing abnormality	Tools that identified aberrant splicing event	Median TPM blood (GTEx)
			r23)		
WDR26	NM_001379403.1:c.823-10A>G	AG|1.00|-1; AL|0.87|-10	Alternative 3' splice acceptor site in intron 2 (in-frame insertion of three amino acids) r.822_823ins823-9_823-1 p.(Lys274_Ala275insPheLeuGln)	rMATS, MAJIQ and LeafCutterMD	38.32

One variant, in *KAT6B* which was only picked up by rMATS-turbo and did not pass the significance threshold, is predicted to cause a splice donor loss leading to exon skipping. In IGV we only observed 4 reads showing the exon skipping event, which is likely why this event was not detected by the tools, however, this AS event has been validated via RT-PCR, showing it is a genuine splicing event. It is likely that the low number of reads is due to the low expression of *KAT6B* in blood.

Illustrative cases of variants affecting splicing

Case 1 – SOT226 (*PHF8*): confirming effects of a canonical splice site variant. rMTAS-turbo, MAJIQ and LeafCutterMD all identified skipping of exon 8 as well as skipping of both exons

7-8 linked to a canonical splice acceptor variant (chrX:54011286T>C, NM_015107.3:c.784-2A>G) within intron 7 in gene *PHF8* which codes for a histone demethylase shown to have roles in gene regulation, growth, and development[42–45], and known to cause intellectual developmental disorder, X-linked syndrome, Siderious type[46]. *In-silico* analysis supported a deleterious effect of VUS on splicing (SpliceAI Δ score = 0.99). Furthermore, this sample showed no normal junctions mapping to exon 8 and therefore it appears that this patient had no normal transcripts of this gene (**Figure 1**). The two resulting transcripts are predicted to be out of frame, leading to a deletion of 163 and 350 nucleotides. As this is an X-linked recessive gene in a male patient, the results of the RNA sequencing are in agreement with the pattern of inheritance. RT-PCR confirmation identified both exon 8 skipping as well as the double exon skipping of exons 7-8. Exon 8 and 7 both fall within the protein's Jumonji C (JmjC) domain, which has catalytic function responsible for preferential demethylation of H3K9me2[47–49].

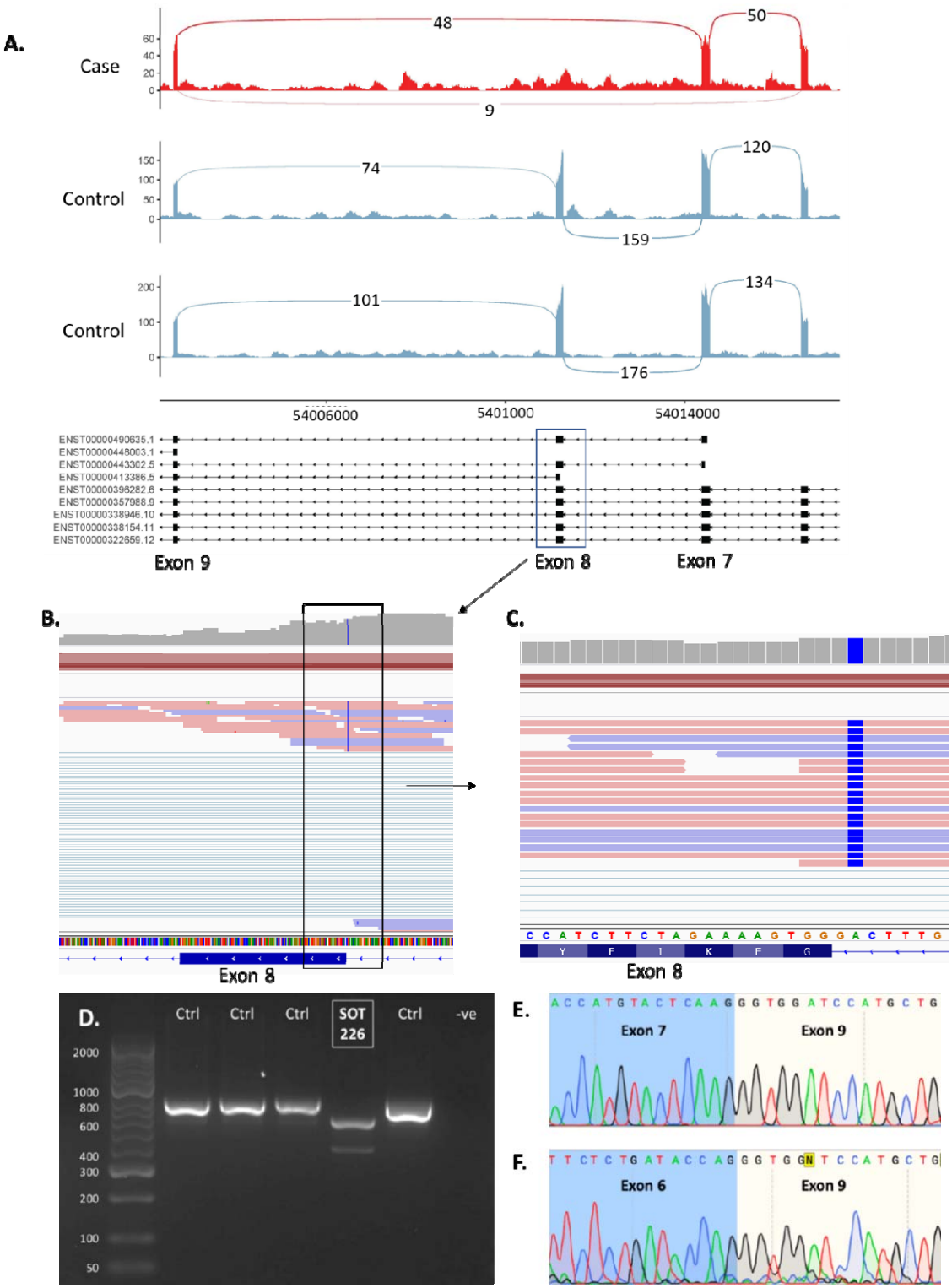


Figure 1. *Exon skipping caused by intronic variant in PHF8 gene.* ***A)*** *Sashimi plot of the proband and two controls of the exon skipping region in PHF8. For the proband only (red*

track), we observed exon 8 skipping (exon downstream of variant). ***B)*** *IGV screenshot of coverage across exon 8. Figure shows there are no reads mapping back to normal exon 8 splicing junctions.* ***C)*** *Close up of PHF8(NM_015107.3):c.784-2A>G variant which falls within a canonical splice site. D) RT-PCR demonstrates two abnormally spliced products both of shorter length than the predominant band in controls. E) Sanger sequencing of the larger band confirms skipping of exon 8. F) Sanger sequencing of the smaller band reveals skipping of exons 7-8.*

Case 2 – SOT241 (*WDR26*): resolving the effects of an intronic variant 10bp upstream of the splice acceptor. rMATS-turbo, MAJIQ and LeafCutterMD all identified an alternative splice acceptor site being created within the second intron of the Skraban-Deardorff syndrome associated[50] *WDR26* gene, in close proximity to an intronic variant (chr1:224431591T>C, NM_001379403.1:c.823-10A>G). This variant is found in ClinVar[51] (rs1674395504, Variation ID: 962945) with three submissions, one as pathogenic, likely pathogenic and a third as a VUS, leading to a classification of conflicting interpretations of pathogenicity overall. *In-silico* analysis of this variant also supported a deleterious effect on splicing (SpliceAI Δ score = 1). In this case 58% of junctions include the new acceptor site (**Figure S3**). The RNA sequencing therefore suggests that about half of the transcripts are alternatively spliced, which is consistent with the heterozygous state of this variant and the in-frame insertion of three amino acids that is predicted to result: r.822_823ins823-9_823-1 p.(Lys274_Ala275insPheLeuGln).

Splicing analysis in patients without candidate VUSs

The unknown patient cohort comprised 33 individuals without a candidate VUS and 23 further cases where the original candidate VUS had not been found to alter splicing. To identify new candidate events in these cases, we took a systematic approach to narrow

down the results obtained from the splicing tools to a manageable number so these could be inspected manually in IGV. LeafCutterMD identified an average of 1,144 (593-3,345) significant events per individual (raw p-values ≤ 0.05) with an effect size greater than 0.2 or less than -0.2. The MAJIQ voila module DeltaPSI identified an average 1,372 (644-3,699) significant events per sample. While rMATS-turbo identified an average of 3,578 (2,370-115,522) significant events per sample with an inclusion level greater than 0.2 or less than -0.2 (**Figure S4**).

Our first approach using filtered VCF files to extract AS events within 25 base pairs of a variant reduced the mean number of events per proband ~300 fold (rMATS-turbo: 12 (0-69); MAJIQ: 9 (0-40), LeafCutterMD: 3 (0-14), (**Figure S4**)). Inspection of all events in IGV led to the identification of two new candidate variants/event.

Case 3 – SOT260 (*NARS1*): identification of a splice-altering variant in a child with undiagnosed global developmental delay. rMATS-turbo, MAJIQ and LeafCutterMD all identified two AS events within the *NARS1* gene. The first was an alternative donor site within exon 13 and the second was retention of intron 13 (**Figure 2**). These events were proximal to a heterozygous missense variant within exon 13 (*NARS1* c.1460C>T) predicted to affect splicing (SpliceAI Δ score = 0.93) by creating a new donor site. Deleterious variants in *NARS1* are associated with Neurodevelopmental disorder with microcephaly, impaired language, and gait abnormalities, which would be consistent with the patient's phenotype. *NARS1* pathogenicity is generally associated with biallelic deleterious variants, however a recent study by Manole and colleagues has shown that *de novo* variants, including a recurrent nonsense mutation at the end of the protein can have a gain-of-function effect that alters normal protein function by interfering with the ATP-binding domain, crucial for enzymatic function[52]. In this case the intron retention is predicted to lead to an out-of-

frame transcript, while the new donor site is predicted to lead to an in-frame deletion of 19 amino acids, both affecting the ATP-binding domain.

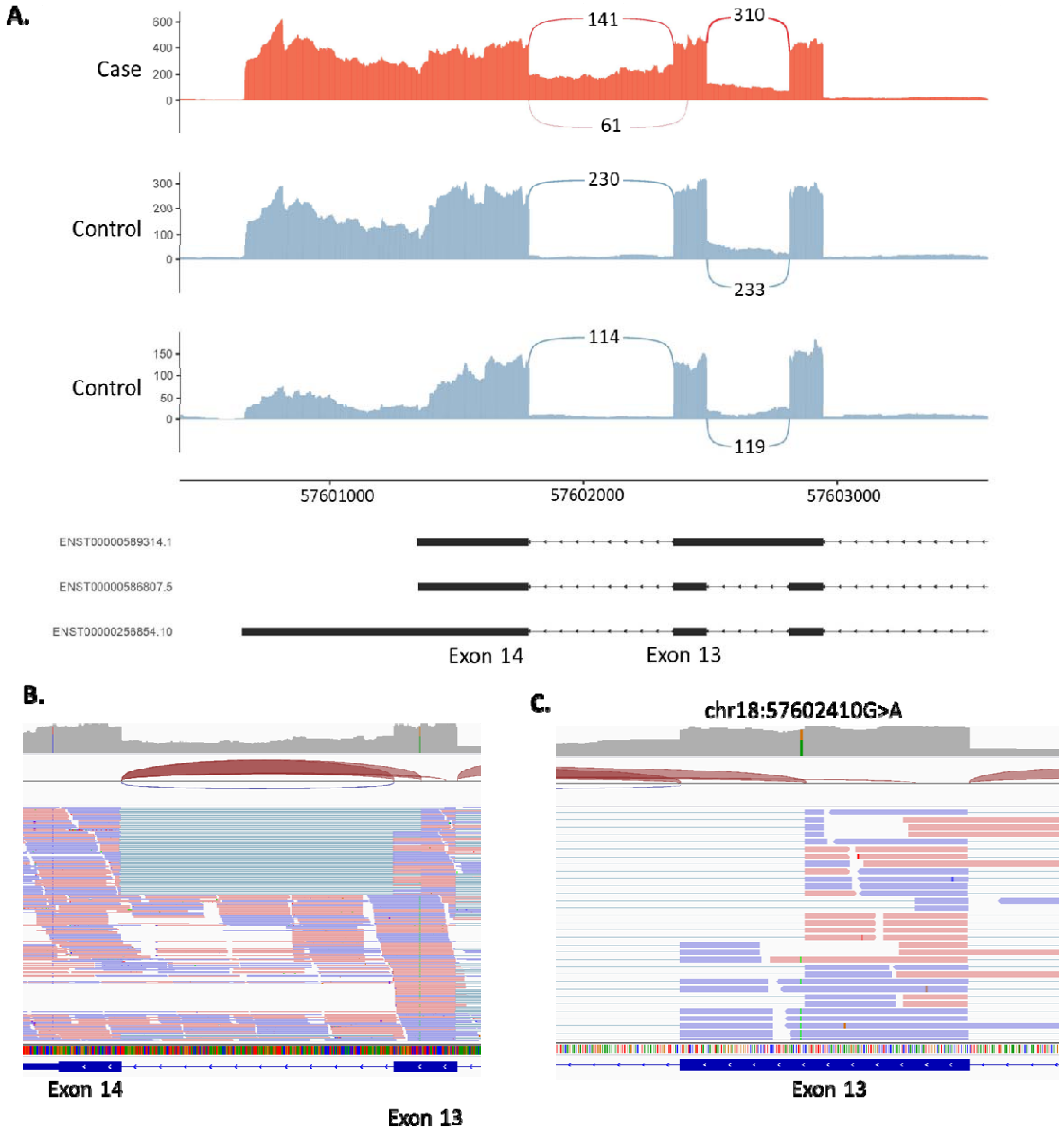


Figure 2. *Alternative donor site and intron retention in NARS1 gene.* ***A)*** *Sashimi plot of the proband and two controls of the alternative donor and intron retention region in NARS1. For the proband only (red track), we observed an alternative donor site in exon 13 as well as intron 13 retention.* ***B)*** *IGV screenshot of coverage across exons 13 and 14.* ***C)*** *Close up of*

NARS1:c.1460C>T variant, a deep exonic variant predicted to affect splicing by creating a new donor site within exon 13.

Case 4 -SOT247 (*ARFGEF1*): inclusion of a cryptic exon in a child with undiagnosed developmental delay. rMATS-turbo and MAJIQ identified an AS event within the *ARFGEF1* gene associated with a deep intronic variant (chr8:67274262A>T, NM_006421.5:c.1337+1713T>G). This particular case was originally referred for analysis of a VUS (*SON* c.1160C>T), which after assessment in IGV was not observed to cause aberrant splicing. The *ARFGEF1* variant was predicted to affect splicing (SpliceAI Δ score = 0.67). The sequencing data shows the creation of a new acceptor and donor site within intron 9 suggesting the inclusion of a cryptic exon (**Figure 3**), which would result in an out-of-frame insertion of 186 nucleotides.

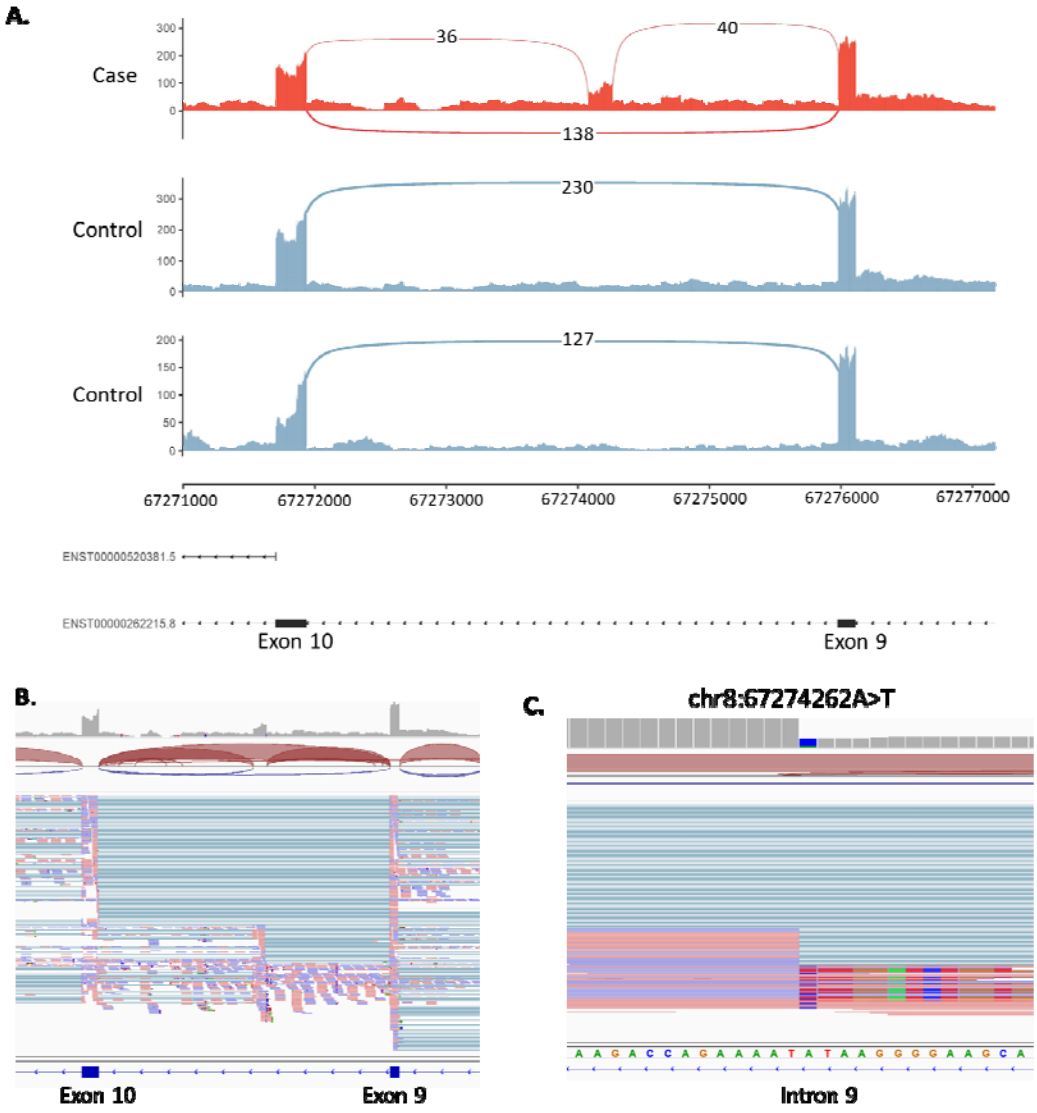


Figure 3. *Activation of pseudoexon caused by intronic variant in ARFGEF1 gene.* ***A)*** *Sashimi plot of the proband and two controls of the ARFGEF1 region of interest. For the proband only (red track), two novel splice junctions can be seen suggesting the activation of a pseudoexon in intron 9.* ***B)*** *IGV screenshot of coverage across region of interest.* ***C)*** *Close up of chr8:67274262A>T variant.*

Our second filtering strategy was a phenotype to genotype approach. Using the phenotype information available, results from the splicing tools were filtered using the appropriate Genomic Medicine Service (GMS) gene panels. As we were limited by available phenotype

information, only twenty unknown cases were assessed this way. This strategy led to the identification of one new candidate variant/event.

Case 5 – SOT262 (*AP4E1*): identification of pseudoexon inclusion and a second frameshift variant in a child with undiagnosed hypotonia. The hypotonic infant GMS panel (v18.1) was applied to rMATS-turbo, MAJIQ and LeafCutterMD results, all of which identified activation of a pseudoexon within intron 1 of *AP4E1* involving use of one alternative splice acceptor site and two alternative donor sites (**Figure 4**). The two resulting transcripts are predicted to be out of frame, leading to an insertion of 142 and 38 nucleotides. These events were associated with an intronic variant (chr15:50911536G>A) weakly predicted to affect splicing. SpliceAI delta scores were 0.11 and 0.09 for acceptor gain (-32bp) and donor gain (5bp) respectively, but these were just below the 0.2 cut-off. However, analysis of the mutated sequence using ESEfinder predicts that the G>A base transition identified at this position may act as an exonic splicing enhancer through creation of a binding site for splicing factors SC35 (SRSF2) and/or SRp40 (SRSF5)[53]. Furthermore, a heterozygous single-nucleotide deletion was observed in exon 6, NM_007347.5:c.567del, p.(Leu190TrpfsTer43), predicted to lead to an out-of-frame transcript (**Figure 4C**). This event was not picked up by the first method, as the variant was filtered out due to stringent quality thresholds [genotype quality (GQ) < 16; variant had a GQ of 6] required to manage noise when calling variants in RNA-seq data. Biallelic variants in *AP4E1* are associated with spastic paraplegia type 51, which is consistent with the phenotype information we have available for the proband.

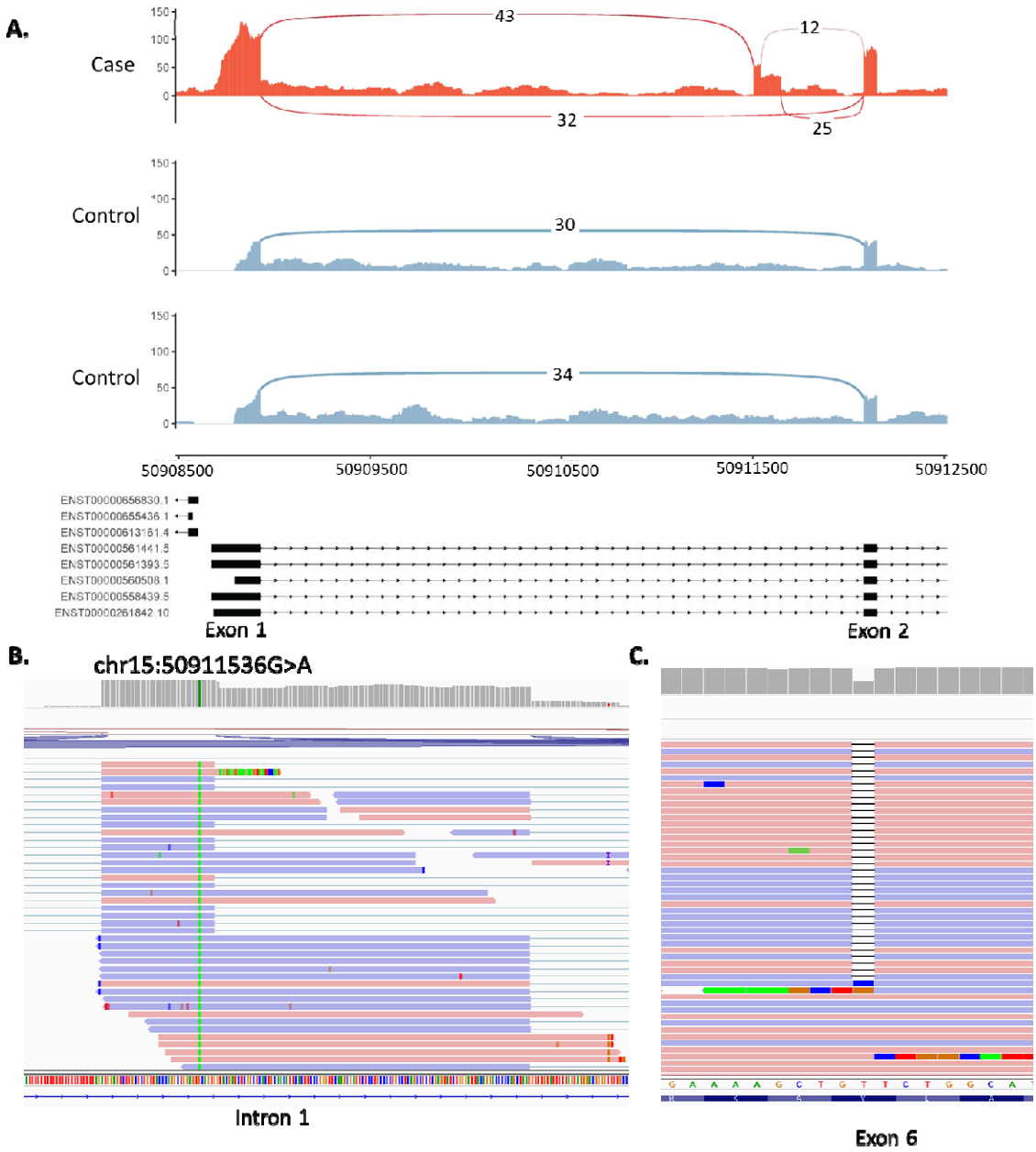


Figure 4. *Activation of pseudoexon caused by intronic variant in the AP4E1 gene.* ***A)*** *Sashimi plot of the proband and two controls of the AP4E1 region of interest. For the proband only (red track), three novel splice junctions can be seen suggesting the activation of a pseudoexon in intron 1.* ***B)*** *Close up of NM_007347.5:c.151-542G>A variant in IGV.* ***C)*** *Heterozygous single nucleotide deletion observed in exon 6 (chr15:50929032delT).*

Gene expression outlier analysis with OUTRIDER

OUTRIDER identified 175 gene expression outliers across 39 samples. Of the 39 samples that had expression outliers 16 were cases with a VUS, 18 were cases without a VUS and 5 were cases with known molecular diagnosis. Ten cases within our cohort had known chromosome microdeletions previously identified through microarray analysis (3 cases with known diagnosis and 7 with unknown diagnosis). In 5/10 of these cases, OUTRIDER identified genes with significantly lower expression which overlapped the deleted regions previously identified (**Figure 5**). For the 16 cases which had a VUS, none of the outliers identified matched the gene in which the VUS was found. A deeper analysis of the results did show that there were two cases (SOT070 and SOT069) for which the expression rank of the gene in which the VUS was found was 1 [lowest expression in the whole cohort] but was not significant after correction for multiple testing (**Figure S5**).

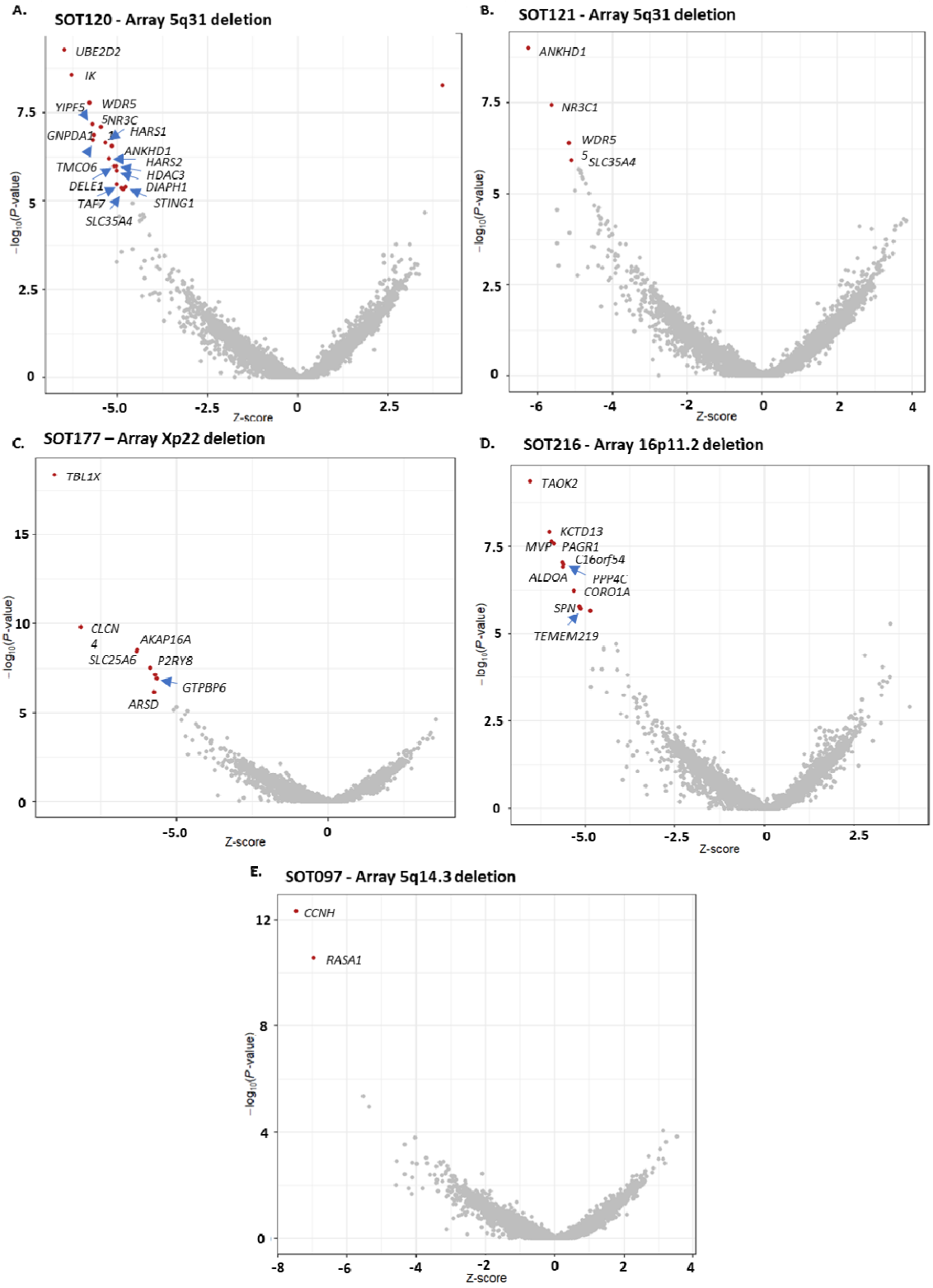


Figure 5. *RNA aberrant expression detection with OUTRIDER.* ***A)*** *Expression outliers for sample SOT120, proband with an array 5q31 deletion (n=18, 17 within deleted region).* ***B)***

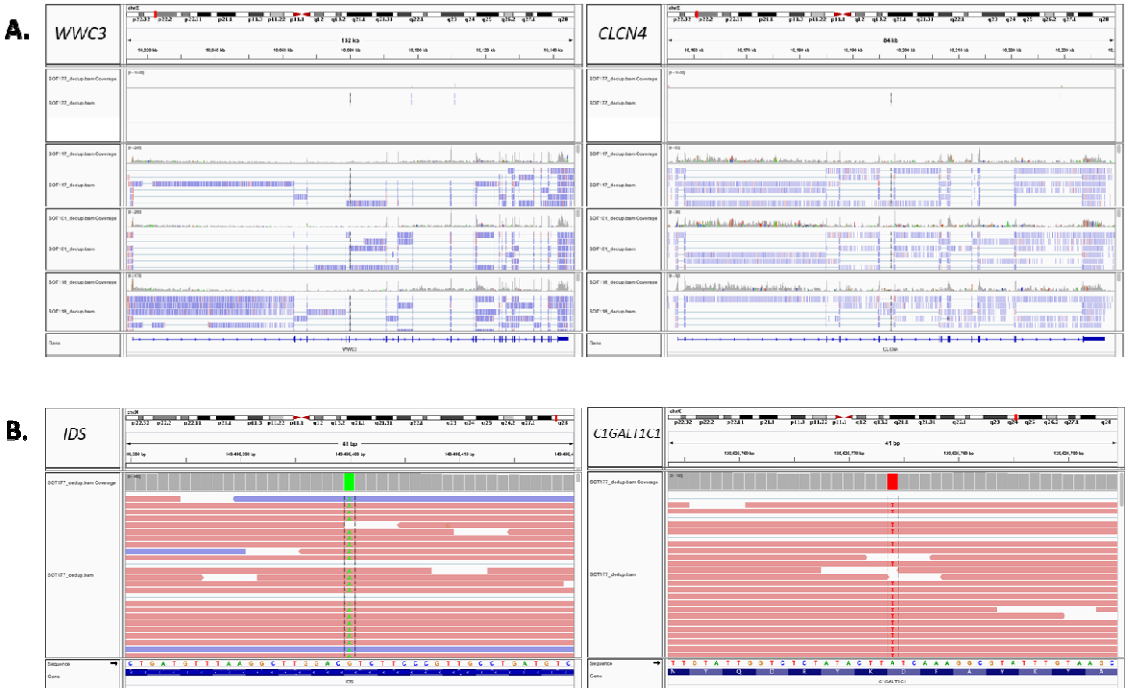
Expression outliers for sample SOT121, proband with an array 5q31 deletion (n=4, all within deleted region). ***C)*** *Expression outliers for sample SOT177, proband with array Xp22 deletion (n=8, 7 within deleted region).* ***D)*** *Expression outliers for sample SOT216, proband with 16p11.2 deletion (n=11, 10 within deleted region).* ***E)*** *Expression outliers for sample SOT097, proband with 5q14.3 deletion (n=2, both within deleted region). For all plots, red points represent those that fell within the deleted regions.*

Detection of skewed X-inactivation

In one of the analysed cases (SOT177), RNA-seq was able to confirm skewed X-inactivation. This individual was a female child with developmental delay and dysmorphic features. Chromosome microarray analysis had identified a *de novo* 10.2 Mb deletion of Xp22.33p22.2. However, this copy number variant was classified as a VUS owing to the child being female and the assumption that the X chromosome carrying the deletion would be preferentially inactivated. Standard DNA-based X-inactivation testing proved uninformative in this case, but further primer sets showed unilateral inactivation. Trio whole-genome sequencing was subsequently undertaken in order to further seek a potential cause for the patient's condition. No candidate variant was identified. However, it was possible to use parental SNP data to determine that the Xp deletion had occurred on the paternal X chromosome (**Figure S6**).

Analysis of 9 additional heterozygous expressed SNPs in the patient's RNA-seq data from loci across both arms of the X chromosome also revealed monoallelic paternal expression of X-linked genes (**Figure 6**). This therefore confirms complete skewing of X-inactivation towards the paternally inherited X-chromosome carrying the 10.2 Mb deletion. The cause of this extreme skewing currently remains unknown, as no candidates were found on the maternal X. However, the deletion is now thought to be causative for the patient's

presenting phenotype, resulting in a functional nullisomy for all genes in the deletion region that are subject to X inactivation.



C.

Locus	Gene	Heterozygous SNPs (GRCh38)	Parental origin	RNA-seq: REF	RNA-seq: ALT	Skewed expression
Xp22.2	*MOSPD2*	X:14911253:G>A	Maternal	66 reads G	0 reads A	REF 100% - Paternal
Xp22.13	*PHKA2*	X:18907981:C>T	Maternal	76 reads C	0 reads T	REF 100% - Paternal
Xp11.23	*TBC1D25*	X:48559322:C>T	Maternal	45 reads C	0 reads T	REF 100% - Paternal
Xp11.22	*HUWE1*	X:53625883:A>G	Maternal	77 reads A	1 read C (0 reads G)	REF 99% - Paternal
Xq11.2	*MTMR8*	X:64268644:A>T	Paternal	0 reads A	14 reads T	ALT 100% - Paternal
Xq13.1	*PHKA1*	X:72695733:G>A	Maternal	6 reads G	1 read A	REF 86% - Paternal
Xq24	*DOCK11*	X:118605272:A>G	Maternal	201 reads A	0 reads G	REF 100% - Paternal
Xq24	*C1GALT1C1*	X:120626774:A>T	Paternal	0 reads A	114 reads T	ALT 100% - Paternal
Xq28	*IDS*	X:149496400:G>A	Paternal	0 reads G	460 reads A	ALT 100% - Paternal

Figure 6. *RNA-seq confirms skewed X-inactivation.* ***A.*** *IGV screenshots of RNA-seq data from WWC3 and CLCN4, which lie within the Xp22.33p22.2 deletion. The patient's sample (top track) shows no RNA-seq coverage compared to controls.* ***B.*** *IGV screenshots of RNA-seq data for heterozygous SNPs illustrating lack of maternal allele expression.* ***C.*** *Table of expression of selected heterozygous X-linked SNPs from across the X chromosome confirming skewing towards the paternal X.*

Comparison of open-source tools to detect alternative splicing

A direct comparison of the splicing tools rMATS-turbo, MAJIQ, LeafCutterMD was not possible as they produce very different outputs. MAJIQ and LeafCutter collate AS events into clusters, while rMATS-turbo does not. It is relevant to note that rMATS-turbo and MAJIQ are designed to be run with balanced groups i.e., similar number of controls and cases, while LeafCutterMD was specially designed for cohorts with no replicates. To compare sensitivity of the tools, we focused on cases where patients had a VUS and aberrant splicing was observed in IGV (n=14). The splicing tools rMATS-turbo, MAJIQ, and LeafCutterMD each identified 11, 4 and 2 of the AS events respectively. rMATS-turbo had by far the best sensitivity identifying 79% of the AS events, where 7 were events identified solely by this tool.

The splicing effects of three variants were consistently missed by the three splicing tools, *SF3B4* c.417C>T, *UBR4* c.8488+3A>G and *SMARCE1* c.8-4A>G. The *SF3B4* variant is predicted to affect splicing (SpliceAI Δ score = 0.37). This gene has a high GC content and low mappability in large regions of its exons. Only two reads mapped to the new junction, however, this event has been validated via RT-PCR [results previously reported[10]] and has also been characterised using a β-globin hybrid minigene assay[54]. The *UBR4* variant is predicted to cause a donor loss leading to intron retention (SpliceAI Δ score = 0.26). In IGV 46 reads covering the intronic region were observed, which included the variant and loss of the donor site, but this was not detected by any of the tools (**Figure S7**). Lastly, the *SMARCE1* variant is predicted to cause an acceptor loss (SpliceAI Δ score = 0.22). Like the *SF3B4* variant, it also has few (n=9) reads mapping to the new junction. In this case the ratio of reads covering the abnormal junctions to normal junction is 9:81. All three events were validated via RT-PCR and in the case of the *UBR4* intron retention with additional qPCR.

Discussion

In this work we have systematically assessed patients with no diagnosis by DNA testing and patients with a VUS. Importantly, we show that it is possible to make diagnoses using just RNA-seq in patients without a candidate VUS as well as classify VUSs using blood-based RNA-seq and RT-PCR to uplift diagnostic yield in rare disease patients. This work displays the variety of events that can be picked up using RNA-seq (i.e. deep intronic and exonic variants, complex splicing abnormalities, deletions, skewed x-inactivation) highlighting the wide range of applications this technology can have in the clinical setting. Furthermore, we compared different open-source tools and their ability to detect AS events highlighting their strengths and limitations.

Splicing analysis in patients with VUSs

In this study, high-throughput blood-based RNA sequencing allowed us to evaluate the effect on splicing of 37/52 VUSs across 48 patients in clinically important genes. 38% of assessed VUSs caused aberrant splicing detectable by RNA-seq, helping to clarify variant interpretation and provide supporting evidence of pathogenicity[55]. For the 15/52 VUSs in which splicing could not be assessed using RNA-seq, identification of splicing and expression abnormalities were limited by gene expression in blood, ultimately affecting gene coverage. In comparison to RNA-seq, RT-PCR proved to be more sensitive allowing us to assess 41 VUSs and identified a further four likely pathogenic AS events, meaning that in total 35% of VUSs in this cohort (44% of those that could be adequately assayed) were found to affect splicing (**see Table S1**). These figures are in concordance with SpliceAI predictions which had a sensitivity and specificity of 94% and 91% respectively. The increase in sensitivity of RT-PCR can be attributed to the targeted approach allowing amplification of AS events in lowly expressed genes[56], as well as amplification of AS events with low inclusion levels (in some

cases accounting for NMD). For the nine cases where neither RNA-seq nor RT-PCR was able to resolve the VUS, such variants would need to be assessed with other tissue types or via alternative methods such as minigene analysis or potentially using animal models should the collection of an appropriate or adequately representative tissue not be feasible[56].

The MRSD tool was used to predict the minimum required sequencing depth for genes of interest and was found to be very conservative. From our empirical data, genes with GTEx whole-blood TPM values of 5 or above are likely to be assessable for splicing analysis using the RNA-seq parameters employed in this study, while genes with TPM values down to 0.9 may be assessable by RT-PCR. A TPM threshold of 5 would correspond to 1104/3113 (35%) of genes listed in the UK Genomic Medicine Service's PanelApp[24] list of disease genes, while a threshold of 0.9 would include 1866/3113 (60%), see **Figure S8**. Based on our analysis, we recommend that RT-PCR should be the first-choice test to assess VUSs in genes with low expression in blood such as *BRCA1, BRCA2* and *FBN1.* In some instances, informative RT-PCR results can be obtained even in genes reported to have a TPM value of zero in GTEx[10]. However, in most other cases RNA-seq is likely to prove more advantageous as a first line test. RNA-seq can identify splicing events with more granularity, particularly when new AS events entail only one or a few nucleotides. Furthermore, with RNA-seq we can quantify splice isoforms, identify expression outliers, and most importantly we can look at alternative splicing events without prior expectation of what the causal variants may be. Case 4 (SOT247) highlights just how useful having transcriptome wide data is, as this patient was referred with a VUS that did not cause aberrant splicing, but the ability to look at the entire transcriptome led to the identification of a likely disruptive splicing event in a different gene. While the VUSs in this cohort were enriched for variants affecting splicing, these were clinically identified VUSs for which clarification of pathogenicity was sought by clinicians,

highlighting the need for this type of analysis to be integrated into clinical practice. Overall, we were able to assess 83% of VUSs (RNA-seq and RT-PCR combined), establishing blood as a suitable tissue for validating aberrant splicing in rare disease patients.

Splicing analysis in patients without VUSs

Out of a total of 56 cases without a previously identified VUS or with a VUS but with no aberrant splicing observed, our RNA-seq analysis identified four cases with relevant splicing alterations and one case with skewed X-inactivation, suggesting a potential diagnostic uplift rate of 9%. This is an important untapped group of variants with few established high-throughput methods of analysis in these types of cohorts[57]. We thereby demonstrate that it is possible to identify new candidate diagnoses and splicing events in patients with no prior candidate sequence variants, although with a much lower yield than if a VUS were previously identified. Interestingly, all four splicing events found in patients with no candidate variants were caused by deep intronic variants, regions of the genome often overlooked in genomic investigations and where it is difficult to predict functional effects. If prediction algorithms are to be used for prioritisation of variants, these may need to be tailored by genomic region, such as having a more permissive SpliceAI score threshold for deep intronic variants. As demonstrated by the activation of a pseudoexon caused by a deep intronic variant in the *ARFGEF1* gene whose SpliceAI delta score was just below the widely used 0.2 cut-off.

The low diagnostic yield in patients without a candidate VUS could be attributed to a number of factors: 1) The patient could have a variant affecting splicing in a gene that is not expressed in blood or there is a tissue specific impact that is not present in blood; 2) The molecular cause of the disease does not affect splicing; 3) The tools are not able to identify these events with high confidence (e.g. aberrant isoform is undergoing nonsense mediated

decay or difficult to align) ; 4) Performance of the tools is variable, some events are picked up better than other (e.g. exon skipping compared to intron retention); and 5) The variant could have been filtered out.

RNA sequencing may not ideal for variant calling as it generates high numbers of false positive calls compared to DNA sequencing due to both biological and technical differences. During the transcription process inside the cells RNA is edited (polyadenylation and splicing of introns), adding an additional layer of complexity when aligning RNA compared to DNA. Furthermore, expression is highly variable both in time and space meaning coverage across genes is inconsistent and dependent on tissue type. The alignment process itself can introduce technical errors and even before sequencing random errors can be introduced during the reverse transcription from RNA to cDNA[58]. We were limited to using RNA for variant calling as there was no matching DNA sequencing available for most cases in this study. Consequently, the high number of false positive variant calls led to strict filtering criteria where only SNVs were inspected and thus events caused by indels will have been missed. Nonetheless, the use of gene panels to restrict results from the three splicing tools did recover a variant that had been excluded due to harsh filters. This approach does limit the analysis by restricting to known disease genes and was only applicable for 37% of our cohort for which we had phenotypic data available to enable identification of appropriate gene panels. If matched whole genome sequencing data and detailed phenotypic information were available, integration of this data would likely increase events identified in this patient subgroup and potentially increase diagnostic yield.

Even with the limitations mentioned, RNA variant calling did identify variants that would be found in genomic DNA and was able to identify the underlying causes of aberrant splicing as well as the aberrant splicing event. Furthermore, while a diagnostic uplift of 9% might

seems small, when we apply this to cohorts such as the Genomics England 100,000 Genomes project, this translates into hundreds of new diagnoses that will impact to patients' lives and end a long diagnostic odyssey.

Identification of splicing events linked to VUSs by different splicing tools

While we did not set out to benchmark a comprehensive selection of splice junction detection tools, we did however want to establish if widely used tools could be used to detect alternative splicing in rare disease patients as datasets used in previous benchmarking studies were not comparable to ours[59–61]. Using the 14 cases which had aberrant splicing linked to known VUSs, rMATS-turbo had the highest sensitivity followed by MAJIQ and then LeafcutterMD. While we expected all tools to perform well, LeafCutterMD was the only tool developed for outlier splicing detection and therefore it was unexpected that it had the lowest sensitivity. There were three variants which were consistently missed by the splicing tools (*SF3B4* c.417C>T, *UBR4* c.8488+3A>G and *SMARCE1* c.8-4A>G) and is likely that the low number of reads covering the variants within *SF3B4* and *SMARCE1* is the reason the splicing tools are not picking up these AS events. However, for the intron retention in *UBR4*, it is unclear why the tools are not identifying this event as there is strong supporting evidence (**Figure S7**).

In patients with no prior VUS, the sheer number of significant events resulting from the three splicing tools created a challenge in terms of identifying new potential alternative splicing events that could be linked to the patients' conditions. However, we were able to filter these down to a manageable number and identify new likely disruptive alternative splicing events albeit with strict filtering criteria, rendering it likely genuine events were missed. The low concordance between tools and high number of false positives highlights

the need for new tools and methods that are sensitive but are also able to reduce the number of spurious calls.

Gene expression analysis

OUTIRDER was able to detect half of the known microdeletions in the cohort, but it did not detect significant alterations in gene expression in genes which did show aberrant splicing. There were only two instances where the VUS gene was ranked 1 (lowest expression for the whole cohort) and although neither passed the significance threshold, this suggest that there is a likely decrease of normal transcripts, but this method is not sensitive enough to detect it. This finding also indicates that abnormal splicing is not necessarily associated with a significant reduction in gene expression, at least in blood, and over-reliance on such expression changes for the identification of splicing abnormalities is unlikely to have reliable sensitivity. This is particularly interesting as we would expect a large number of the splicing abnormalities to shift the reading frame and therefore undergo nonsense mediated decay (NMD) significantly decreasing the abundance of the transcript. This lack of change in expression of genes with aberrant splicing could be due multiple factors such as limited sensitivity of the tools as mentioned previously; the impact of NMD is not very pronounced in blood-based RNA-seq and tissue specific RNA-seq is required; and/or the targeting methodologies bias the type of transcripts and number of transcripts we observe. Furthermore, whole blood has been shown to have high variability in gene expression profiles particularly when compared to skin fibroblasts[16]. Some studies suggest that fibroblast RNA enables the investigation of a more comprehensive set of genes than whole blood and that this is likely the better tissue for detecting clinically relevant differences in gene expression[12, 14, 19]. While blood-based RNA analysis may not be optimal, it does offer a number of benefits over fibroblasts. It is more routinely sampled, less invasive to obtain and

does not require cell culture prior to testing, meaning it is cheaper, faster and has lower requirements in terms of specialised knowledge and facilities. This study demonstrates that despite its limitations, it is valuable in this context.

Therapeutic applications

Accurate diagnoses facilitate appropriate clinical management, accurate genetic counselling and informed reproductive decision making, but in some cases, there would be the potential for bespoke RNA-targeted therapies to be designed to correct a given splicing abnormality and slow or halt the progression of an individual's disease. Cases such as that of the *AP4E1* pseudoexon inclusion variant highlighted in this study may be especially suitable targets in this regard, on account of the gradual neurodegenerative nature of the associated condition and the known efficacy of other antisense oligonucleotide therapies delivered to the central nervous system such as nusinersen[62,63]. Notwithstanding the substantial challenges and barriers facing development of such bespoke therapeutics, precedent does exist for n=1 oligonucleotide therapies[64]. The utility of RNA-seq in being able to identify these types of mutations means that an effective personalised medicine healthcare system will benefit from having access to RNA transcriptomics within the diagnostic clinical setting.

Conclusion

RNA-seq should be considered as a complementary tool in genetic testing to uplift diagnostic yield in cohorts of patients with rare disorders. When considering analysis of RNA, RT-PCR should be the first-choice test to assess VUSs in genes with low expression, but high throughput RNA sequencing is more advantageous as a first line test. Our analyses suggest at least one third of patients with rare disorders could benefit from the increased diagnostic yield offered by this kind of analysis. Overall, we were able to validate splicing abnormalities in patients with a VUS and identified five new diagnoses by detecting novel AS events in

patients with no candidate sequence variants, giving an overall uplift in diagnostic yield of 26%.

Acknowledgements

This work was funded by a National Institute for Health Research (NIHR) Research Professorship grant (RP-2016-07-011) awarded to DB. The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. The authors thank all patients and families taking part in this research.

Author contributions

Conceptualization, D.B, A.G.L.D, J.L; Data Curation C.J.O, J.L; Formal Analysis C.J.O, A.G.L.D, J.L, H.A.W; Funding Acquisition D.B; Investigation C.J.O, A.G.L.D, J.L, H.A.W, W.R, D.J.B, D.H, N.S.T; Methodology C.J.O, A.G.L.D, J.L, H.A.W; Project administration, D.B; Resources, D.B, A.G.L.D; Software, C.J.O, J.L, W.R; Supervision D.B; Validation, H.A.W, A.G.L.D, D.H;D.J.B; Visualization C.J.O; Writing – Original Draft, C.J.O, A.G.L.D, Writing – Review & Editing; C.J.O, A.G.L.D, J.L, D.J.B, D.H, N.S.T, D.B

Declaration of interest

The authors declare no competing interests.

Web resources

GATKs Best Practices Workflow, https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels-

Picard, http://broadinstitute.github.io/picard.

fastqc, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

ggsashimi, https://github.com/guigolab/ggsashimi

R, https://www.r-project.org/

RStudio, http://www.rstudio.com/

spliceAI lookup, https://spliceailookup.broadinstitute.org/

rMATS-turbo, https://github.com/Xinglab/rmats-turbo/blob/v4.1.2/README.md

LeafCutter, https://davidaknowles.github.io/leafcutter/

MAJIQ, https://majiq.biociphers.org/

Data and code availability

The code generated during this study can be found in GitHub

[https://github.com/carojoquendo/RNA_splicing_and_disease].

References

1. Ewans, L.J., Minoche, A.E., Schofield, D., Shrestha, R., Puttick, C., Zhu, Y., Drew, A., Gayevskiy, V., Elakis, G., Walsh, C., et al. (2022). Whole exome and genome sequencing in mendelian disorders: a diagnostic and health economic analysis. Eur. J. Hum. Genet. *30*, 1121–1131. 10.1038/s41431-022-01162-2.

2. Nurchis, M.C., Altamura, G., Riccardi, M.T., Radio, F.C., Chillemi, G., Bertini, E.S., Garlasco, J., Tartaglia, M., Dallapiccola, B., and Damiani, G. (2023). Whole genome sequencing diagnostic yield for paediatric patients with suspected genetic disorders: systematic review, meta-analysis, and GRADE assessment. Arch. Public Health *81*, 93. 10.1186/s13690-023-01112-4.

3. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report (2021). N. Engl. J. Med. *385*, 1868–1880. 10.1056/NEJMoa2035790.

4. Álvarez-Mora, M.I., Sánchez, A., Rodríguez-Revenga, L., Corominas, J., Rabionet, R., Puig, S., and Madrigal, I. (2022). Diagnostic yield of next-generation sequencing in 87 families with neurodevelopmental disorders. Orphanet J. Rare Dis. *17*, 60. 10.1186/s13023-022-02213-z.

5. Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L., et al. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. *9*. 10.1126/scitranslmed.aal5209.

6. Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat. Commun. *8*, 15824. 10.1038/ncomms15824.

7. Frésard, L., Smail, C., Ferraro, N.M., Teran, N.A., Li, X., Smith, K.S., Bonner, D., Kernohan, K.D., Marwaha, S., Zappala, Z., et al. (2019). Identification of rare-disease genes using

blood transcriptome sequencing and large control cohorts. Nat. Med. *25*, 911–919. 10.1038/s41591-019-0457-8.

8. Gonorazky, H.D., Naumenko, S., Ramani, A.K., Nelakuditi, V., Mashouri, P., Wang, P., Kao, D., Ohri, K., Viththiyapaskaran, S., Tarnopolsky, M.A., et al. (2019). Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. Am. J. Hum. Genet. *104*, 1007. 10.1016/j.ajhg.2019.04.004.

9. Hamanaka, K., Miyatake, S., Koshimizu, E., Tsurusaki, Y., Mitsuhashi, S., Iwama, K., Alkanaq, A.N., Fujita, A., Imagawa, E., Uchiyama, Y., et al. (2019). RNA sequencing solved the most common but unrecognized NEB pathogenic variant in Japanese nemaline myopathy. Genet. Med. Off. J. Am. Coll. Med. Genet. *21*, 1629–1638. 10.1038/s41436-018-0360-6.

10. Wai, H.A., Lord, J., Lyon, M., Gunning, A., Kelly, H., Cibin, P., Seaby, E.G., Spiers-Fitzgerald, K., Lye, J., Ellard, S., et al. (2020). Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. Genet. Med. Off. J. Am. Coll. Med. Genet. *22*, 1005–1014. 10.1038/s41436-020-0766-9.

11. Colin, E., Duffourd, Y., Chevarin, M., Tisserant, E., Verdez, S., Paccaud, J., Bruel, A.-L., Tran Mau-Them, F., Denommé-Pichon, A.-S., Thevenon, J., et al. (2023). Stepwise use of genomics and transcriptomics technologies increases diagnostic yield in Mendelian disorders. Front. Cell Dev. Biol. *11*. 10.3389/fcell.2023.1021920.

12. Yépez, V.A., Gusic, M., Kopajtich, R., Mertes, C., Smith, N.H., Alston, C.L., Ban, R., Beblo, S., Berutti, R., Blessing, H., et al. (2022). Clinical implementation of RNA sequencing for Mendelian disease diagnostics. Genome Med. *14*. 10.1186/s13073-022-01019-9.

13. Dekker, J., Schot, R., Bongaerts, M., Valk, W.G. de, Veghel-Plandsoen, M.M. van, Monfils, K., Douben, H., Elfferich, P., Kasteleijn, E., Unen, L.M.A. van, et al. (2023). Web-accessible application for identifying pathogenic transcripts with RNA-seq: Increased sensitivity in diagnosis of neurodevelopmental disorders. Am. J. Hum. Genet. *110*, 251–272. https://doi.org/10.1016/j.ajhg.2022.12.015.

14. Lee, H., Huang, A.Y., Wang, L.-K., Yoon, A.J., Renteria, G., Eskin, A., Signer, R.H., Dorrani, N., Nieves-Rodriguez, S., Wan, J., et al. (2020). Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. Genet. Med. Off. J. Am. Coll. Med. Genet. *22*, 490–499. 10.1038/s41436-019-0672-1.

15. Maddirevula, S., Kuwahara, H., Ewida, N., Shamseldin, H.E., Patel, N., Alzahrani, F., AlSheddi, T., AlObeid, E., Alenazi, M., Alsaif, H.S., et al. (2020). Analysis of transcript-deleterious variants in Mendelian disorders: implications for RNA-based diagnostics. Genome Biol. *21*, 145. 10.1186/s13059-020-02053-9.

16. Murdock, D.R., Dai, H., Burrage, L.C., Rosenfeld, J.A., Ketkar, S., Müller, M.F., Yépez, V.A., Gagneur, J., Liu, P., Chen, S., et al. (2021). Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. J. Clin. Invest. *131*. 10.1172/JCI141500.

17. Frankish, A., Carbonell-Sala, S., Diekhans, M., Jungreis, I., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Arnan, C., Barnes, I., et al. (2023). GENCODE: reference annotation for the human and mouse genomes in 2023. Nucleic Acids Res. *51*, D942–D949. 10.1093/nar/gkac1071.

18. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinforma. Oxf. Engl. *29*, 15–21. 10.1093/bioinformatics/bts635.

19. Rowlands, C.F., Taylor, A., Rice, G., Whiffin, N., Hall, H.N., Newman, W.G., Black, G.C.M., O'Keefe, R.T., Hubbard, S., Douglas, A.G.L., et al. (2022). MRSD: A quantitative approach for assessing suitability of RNA-seq in the investigation of mis-splicing in Mendelian disease. Am. J. Hum. Genet. *109*, 210–222. https://doi.org/10.1016/j.ajhg.2021.12.014.

20. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26. 10.1038/nbt.1754.

21. Garrido-Martín, D., Palumbo, E., Guigó, R., and Breschi, A. (2018). ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. PLOS Comput. Biol. *14*, 1–6. 10.1371/journal.pcbi.1006360.

22. Piovesan, A., Caracausi, M., Ricci, M., Strippoli, P., Vitale, L., and Pelleri, M.C. (2015). Identification of minimal eukaryotic introns through GeneBase, a user-friendly tool for parsing the NCBI Gene databank. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes *22*, 495–503. 10.1093/dnares/dsv028.

23. Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012). BEDOPS: high-performance genomic feature operations. Bioinforma. Oxf. Engl. *28*, 1919–1920. 10.1093/bioinformatics/bts277.

24. Martin, A.R., Williams, E., Foulger, R.E., Leigh, S., Daugherty, L.C., Niblock, O., Leong, I.U.S., Smith, K.R., Gerasimenko, O., Haraldsdottir, E., et al. (2019). PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat. Genet. *51*, 1560–1565. 10.1038/s41588-019-0528-2.

25. Putri, G.H., Anders, S., Pyl, P.T., Pimanda, J.E., and Zanini, F. (2022). Analysing high-throughput sequencing data in Python with HTSeq 2.0. Bioinformatics *38*, 2943–2945. 10.1093/bioinformatics/btac166.

26. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419. 10.1038/nmeth.4197.

27. R Core Team (2021). R: A language and environment for statistical computing.

28. Soneson, C., Love, M.I., and Robinson, M.D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research *4*, 1521. 10.12688/f1000research.7563.2.

29. Brechtmann, F., Mertes, C., Matusevičiūtė, A., Yépez, V.A., Avsec, Ž., Herzog, M., Bader, D.M., Prokisch, H., and Gagneur, J. (2018). OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. Am. J. Hum. Genet. *103*, 907–917. 10.1016/j.ajhg.2018.10.025.

30. Shen, S., Park, J.W., Lu, Z., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U. S. A. *111*, E5593-5601. 10.1073/pnas.1419161111.

31. Vaquero-Garcia, J., Barrera, A., Gazzara, M.R., González-Vallinas, J., Lahens, N.F., Hogenesch, J.B., Lynch, K.W., and Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. eLife *5*, e11752. 10.7554/eLife.11752.

32. Jenkinson, G., Li, Y.I., Basu, S., Cousin, M.A., Oliver, G.R., and Klee, E.W. (2020). LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. Bioinforma. Oxf. Engl. *36*, 4609–4615. 10.1093/bioinformatics/btaa259.

33. Van der Auwera, G.A., and O'Connor, B.D. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (O'Reilly Media).

34. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*. 10.1093/gigascience/giab008.

35. Jehl, F., Degalez, F., Bernard, M., Lecerf, F., Lagoutte, L., Désert, C., Coulée, M., Bouchez, O., Leroux, S., Abasht, B., et al. (2021). RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species. Front. Genet. *12*, 655707. 10.3389/fgene.2021.655707.

36. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122. 10.1186/s13059-016-0974-4.

37. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma. Oxf. Engl. *26*, 841–842. 10.1093/bioinformatics/btq033.

38. The Genotype-Tissue Expression (GTEx) project. (2013). Nat. Genet. *45*, 580–585. 10.1038/ng.2653.

39. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. *43*, D789-798. 10.1093/nar/gku1205.

40. Harrington, C.A., Fei, S.S., Minnier, J., Carbone, L., Searles, R., Davis, B.A., Ogle, K., Planck, S.R., Rosenbaum, J.T., and Choi, D. (2020). RNA-Seq of human whole blood: Evaluation of globin RNA depletion on Ribo-Zero library method. Sci. Rep. *10*, 6271. 10.1038/s41598-020-62801-6.

41. Carithers, L.J., Ardlie, K., Barcus, M., Branton, P.A., Britton, A., Buia, S.A., Compton, C.C., DeLuca, D.S., Peter-Demchok, J., Gelfand, E.T., et al. (2015). A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreservation Biobanking *13*, 311–319. 10.1089/bio.2015.0032.

42. Loenarz, C., Ge, W., Coleman, M.L., Rose, N.R., Cooper, C.D.O., Klose, R.J., Ratcliffe, P.J., and Schofield, C.J. (2010). PHF8, a gene associated with cleft lip/palate and mental retardation, encodes for an Nε-dimethyl lysine demethylase. Hum. Mol. Genet. *19*, 217–222. 10.1093/hmg/ddp480.

43. Qi, H.H., Sarkissian, M., Hu, G.-Q., Wang, Z., Bhattacharjee, A., Gordon, D.B., Gonzales, M., Lan, F., Ongusaha, P.P., Huarte, M., et al. (2010). Histone H4K20/H3K9 demethylase PHF8 regulates zebrafish brain and craniofacial development. Nature *466*, 503–507. 10.1038/nature09261.

44. Sobering, A.K., Bryant, L.M., Li, D., McGaughran, J., Maystadt, I., Moortgat, S., Graham, J.M.J., van Haeringen, A., Ruivenkamp, C., Cuperus, R., et al. (2022). Variants in PHF8 cause a spectrum of X-linked neurodevelopmental disorders and facial dysmorphology. HGG Adv. *3*, 100102. 10.1016/j.xhgg.2022.100102.

45. Yu, L., Wang, Y., Huang, S., Wang, J., Deng, Z., Zhang, Q., Wu, W., Zhang, X., Liu, Z., Gong, W., et al. (2010). Structural insights into a novel histone demethylase PHF8. Cell Res. *20*, 166–173. 10.1038/cr.2010.8.

46. Laumonnier, F., Holbert, S., Ronce, N., Faravelli, F., Lenzner, S., Schwartz, C.E., Lespinasse, J., Van Esch, H., Lacombe, D., Goizet, C., et al. (2005). Mutations in PHF8 are associated with X linked mental retardation and cleft lip/cleft palate. J. Med. Genet. *42*, 780–786. 10.1136/jmg.2004.029439.

47. Fortschegger, K., de Graaf, P., Outchkourov, N.S., van Schaik, F.M.A., Timmers, H.T.M., and Shiekhattar, R. (2010). PHF8 targets histone methylation and RNA polymerase II to activate transcription. Mol. Cell. Biol. *30*, 3286–3298. 10.1128/MCB.01520-09.

48. Horton, J.R., Upadhyay, A.K., Qi, H.H., Zhang, X., Shi, Y., and Cheng, X. (2010). Enzymatic and structural insights for substrate specificity of a family of jumonji histone lysine demethylases. Nat. Struct. Mol. Biol. *17*, 38–43. 10.1038/nsmb.1753.

49. Feng, W., Yonezawa, M., Ye, J., Jenuwein, T., and Grummt, I. (2010). PHF8 activates transcription of rRNA genes through H3K4me3 binding and H3K9me1/2 demethylation. Nat. Struct. Mol. Biol. *17*, 445–450. 10.1038/nsmb.1778.

50. Skraban, C.M., Wells, C.F., Markose, P., Cho, M.T., Nesbitt, A.I., Au, P.Y.B., Begtrup, A., Bernat, J.A., Bird, L.M., Cao, K., et al. (2017). WDR26 Haploinsufficiency Causes a Recognizable Syndrome of Intellectual Disability, Seizures, Abnormal Gait, and

Distinctive Facial Features. Am. J. Hum. Genet. *101*, 139–148. 10.1016/j.ajhg.2017.06.002.

51. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067. 10.1093/nar/gkx1153.

52. Manole, A., Efthymiou, S., O'Connor, E., Mendes, M.I., Jennings, M., Maroofian, R., Davagnanam, I., Mankad, K., Lopez, M.R., Salpietro, V., et al. (2020). De Novo and Bi-allelic Pathogenic Variants in NARS1 Cause Neurodevelopmental Delay Due to Toxic Gain-of-Function and Partial Loss-of-Function Effects. Am. J. Hum. Genet. *107*, 311–324. 10.1016/j.ajhg.2020.06.016.

53. Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q., and Krainer, A.R. (2003). ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res. *31*, 3568–3571. 10.1093/nar/gkg616.

54. Cassina, M., Cerqua, C., Rossi, S., Salviati, L., Martini, A., Clementi, M., and Trevisson, E. (2017). A synonymous splicing mutation in the SF3B4 gene segregates in a family with highly variable Nager syndrome. Eur. J. Hum. Genet. EJHG *25*, 371–375. 10.1038/ejhg.2016.176.

55. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–423. 10.1038/gim.2015.30.

56. Wai, H.A., Constable, M., Drewes, C., Davies, I.C., Svobodova, E., Dempsey, E., Saggar, A., Homfray, T., Mansour, S., Douzgou, S., et al. (2022). Short amplicon reverse transcription-polymerase chain reaction detects aberrant splicing in genes with low expression in blood missed by ribonucleic acid sequencing analysis for clinical diagnosis. Hum. Mutat. *43*, 963–970. 10.1002/humu.24378.

57. Marwaha, S., Knowles, J.W., and Ashley, E.A. (2022). A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. Genome Med. *14*, 23. 10.1186/s13073-022-01026-w.

58. Guo, Y., Zhao, S., Sheng, Q., Samuels, D.C., and Shyr, Y. (2017). The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data. BMC Genomics *18*, 690. 10.1186/s12864-017-4022-x.

59. Ding, L., Rath, E., and Bai, Y. (2017). Comparison of Alternative Splicing Junction Detection Tools Using RNA-Seq Data. Curr. Genomics *18*, 268–277. 10.2174/1389202918666170215125048.

60. Mehmood, A., Laiho, A., Venäläinen, M.S., McGlinchey, A.J., Wang, N., and Elo, L.L. (2020). Systematic evaluation of differential splicing tools for RNA-seq studies. Brief. Bioinform. *21*, 2052–2065. 10.1093/bib/bbz126.

61. Jiang, M., Zhang, S., Yin, H., Zhuo, Z., and Meng, G. (2023). A comprehensive benchmarking of differential splicing tools for RNA-seq analysis at the event level. Brief. Bioinform., bbad121. 10.1093/bib/bbad121.

62. Ebrahimi-Fakhari, D., Teinert, J., Behne, R., Wimmer, M., D'Amore, A., Eberhardt, K., Brechmann, B., Ziegler, M., Jensen, D.M., Nagabhyrava, P., et al. (2020). Defining the clinical, molecular and imaging spectrum of adaptor protein complex 4-associated hereditary spastic paraplegia. Brain *143*, 2929–2944. 10.1093/brain/awz307.

63. Pane, M., Coratti, G., Sansone, V.A., Messina, S., Catteruccia, M., Bruno, C., Sframeli, M., Albamonte, E., Pedemonte, M., Brolatti, N., et al. (2023). Type I spinal muscular atrophy patients treated with nusinersen: 4-year follow-up of motor, respiratory and bulbar function. Eur. J. Neurol. 10.1111/ene.15768.

64. Kim, J., Hu, C., Moufawad El Achkar, C., Black, L.E., Douville, J., Larson, A., Pendergast, M.K., Goldkind, S.F., Lee, E.A., Kuniholm, A., et al. (2019). Patient-Customized Oligonucleotide Therapy for a Rare Genetic Disease. N. Engl. J. Med. *381*, 1644–1652. 10.1056/NEJMoa1813279.