Genetics in Medicine

Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance

--Manuscript Draft--

Manuscript Number:	GIM-D-19-01085R2
Article Type:	Article
Section/Category:	Clinical Genetics and Genomics
Keywords:	RNA splicing variant interpretation genetic diagnosis genomic medicine RNA-seq
Corresponding Author:	Diana Baralle, MD MBBS University of Southampton Southampton, UNITED KINGDOM
First Author:	Htoo Wai
Order of Authors:	Htoo Wai
	Jenny Lord
	Matthew Lyon
	Adam Gunning
	Hugh Kelly
	Penelope Cibin
	Ellie Seaby
	Kerry Spiers-Fitzgerald
	Jed Lye
	Sian Ellard
	Simon Thomas
	David Bunyan
	Andrew Douglas
	Diana Baralle, MD MBBS
Manuscript Region of Origin:	UNITED KINGDOM
Abstract:	Purpose : Diagnosis of genetic disorders is hampered by large numbers of variants of uncertain significance (VUSs) identified through next-generation sequencing. Many such variants may disrupt normal RNA splicing. We examined effects on splicing of a large cohort of clinically identified variants and compared performance of bioinformatic splicing prediction tools commonly used in diagnostic laboratories. Methods : 257 variants (coding and non-coding) were referred for analysis across three laboratories. Blood RNA samples underwent targeted RT-PCR analysis with Sanger sequencing of PCR products and agarose gel electrophoresis. 17 samples also underwent transcriptome-wide RNA sequencing with targeted splicing analysis based on Sashimi plot visualisation. Bioinformatic splicing predictions were obtained using Alamut, HSF 3.1 and SpliceAI software. Results : 85 variants (33%) were associated with abnormal splicing. The most frequent abnormality was upstream exon skipping (39/85 variants), which was most often associated with splice donor region variants. SpliceAI had greatest accuracy in

	predicting splicing abnormalities (0.91) and outperformed other tools in sensitivity and specificity. Conclusion : Splicing analysis of blood RNA identifies diagnostically important splicing abnormalities and clarifies functional effects of a significant proportion of VUSs. Bioinformatic predictions are improving but still make significant errors. RNA analysis should therefore be routinely considered in genetic disease diagnostics.

Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance

Authors:
Htoo A. Wai PhD[1*], Jenny Lord PhD[1*], Matthew Lyon MSc[2], Adam Gunning BSc[3], Hugh Kelly BMedSc (Hons)[1], Penelope Cibin MSc[1], Eleanor G. Seaby BMBS[1, 4], Kerry Spiers-Fitzgerald BSc[1], Jed Lye BSc[1], Sian Ellard PhD[3], N. Simon Thomas PhD[1, 2], David J. Bunyan PhD[1, 2], Andrew G. L. Douglas MBChB DPhil[1, 5] [+]Diana Baralle MBBS MD[1, 5+]

Affiliations:
1. Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK
2. Wessex Regional Genetics Laboratory, Salisbury District Hospital, Salisbury, UK
3. Exeter Genomics Laboratory, Royal Devon & Exeter NHS Foundation Trust, Exeter, UK
4. Translational Genomics Unit, Broad Institute of MIT and Harvard, Cambridge, USA
5. Wessex Clinical Genetics Service, University Hospitals Southampton NHS Foundation Trust, Southampton, UK

Telephone and e-mail of corresponding author:
+44 (0)23 8120 6170
d.baralle@soton.ac.uk

[*]Joint first authors
[+]Joint senior authors

Conflict of Interest Notification Page

The authors declare no conflicts of interest.

Abstract

Purpose:
Diagnosis of genetic disorders is hampered by large numbers of variants of uncertain significance (VUSs) identified through next-generation sequencing. Many such variants may disrupt normal RNA splicing. We examined effects on splicing of a large cohort of clinically identified variants and compared performance of bioinformatic splicing prediction tools commonly used in diagnostic laboratories.
Methods:
257 variants (coding and non-coding) were referred for analysis across three laboratories. Blood RNA samples underwent targeted RT-PCR analysis with Sanger sequencing of PCR products and agarose gel electrophoresis. 17 samples also underwent transcriptome-wide RNA sequencing with targeted splicing analysis based on Sashimi plot visualisation. Bioinformatic splicing predictions were obtained using Alamut, HSF 3.1 and SpliceAI software.
Results:
85 variants (33%) were associated with abnormal splicing. The most frequent abnormality was upstream exon skipping (39/85 variants), which was most often associated with splice donor region variants. SpliceAI had greatest accuracy in predicting splicing abnormalities (0.91) and outperformed other tools in sensitivity and specificity.
Conclusion:
Splicing analysis of blood RNA identifies diagnostically important splicing abnormalities and clarifies functional effects of a significant proportion of VUSs. Bioinformatic predictions are improving but still make significant errors. RNA analysis should therefore be routinely considered in genetic disease diagnostics.

Keywords:
RNA splicing
variant interpretation
genetic diagnosis
genomic medicine
RNA-seq

INTRODUCTION

Use of next-generation sequencing (NGS) technologies in clinical practice has led to an unprecedented increase in the number of variants being identified in patients undergoing investigation for genetic disorders. Incomplete knowledge of the functional effects of variants and our limited understanding of genotype-phenotype correlations severely compromises attempts to definitively assign or refute pathogenicity for a large proportion of variants. Variant of uncertain significance (VUS) reporting rates vary over time and depending on local reporting policies but of all variants listed on ClinVar (as of 13 November 2019), 48% are asserted to be of uncertain significance (Figure S1).[1] In a clinical setting, this uncertainty has major implications for patients and their families, where having a clear genetic diagnosis can allow evidence-based management decisions to be taken and informed reproductive choices to be made.[2,3]

RNA splicing is thought to be disrupted by up to 62% of all pathogenic single nucleotide variants (SNVs).[4] Current bioinformatic filtering strategies and clinical interpretation guidelines tend to focus heavily on amino-acid-level effects in terms of both variant detection and assignment of pathogenicity.[5] This can lead to synonymous variants being filtered out at an early stage of analysis, even though such variants may affect splicing. Similarly, although deep intronic variant data are increasingly available via NGS approaches like genome sequencing, such non-coding variants are rarely considered owing to a lack of evidence on which to base interpretations. Where bioinformatic predictions suggest that a variant affects splicing, there can be scope for additional RNA-based investigations. However, such splicing prediction tools frequently produce conflicting results and their

accuracy and utility decreases outside of canonical splice sites and consensus splice regions.[6]

In this study, we looked for RNA splicing defects in a large cohort of VUSs identified in patients who had undergone diagnostic genetic testing. We compare *in silico* predictions of splicing with the results of blood RNA analysis and provide examples that illustrate the clinical utility of RNA-based testing in clinical diagnostics. These results support the routine use of RNA analysis in clinical diagnostic practice.

MATERIALS AND METHODS

Patients and variants

A cohort of patients with VUSs identified through routine diagnostic genetic testing was identified primarily through the Wessex Regional Genetics Laboratory, Salisbury (203 variants), with seven other patients identified through the Exeter Genomics Laboratory. Additional patients with 47 variants from across the UK were identified through the 'Splicing and Disease' research study at the University of Southampton, ethically approved by the Health Research Authority (IRAS Project ID 49685, REC 11/SC/0269) and by the University of Southampton (ERGO ID 23056). Informed consent for splicing studies was provided for all patients from whom samples were obtained

RNA extraction and cDNA preparation

Blood was collected in PAXgene Blood RNA tubes and RNA extracted using the PAXgene Blood RNA Kit (PreAnalytiX, Switzerland). cDNA was synthesised via reverse transcription

using random hexamer primers. For details of each laboratory's individual protocols see Supplementary Methods.

RT-PCR analysis

Primers were designed to amplify the region around each variant (sequences available on request). Wherever possible, primer sequences were positioned at least two exons up- and downstream of the target variant. PCR products were evaluated by agarose gel electrophoresis against control samples and purified PCR products were analysed by direct Sanger sequencing. In a number of cases, PCR products separated by gel electrophoresis were purified and sequenced or cloned into *E. coli* using a TA-cloning vector. Plasmids recovered from single-clone bacterial cultures were analysed by Sanger sequencing. Please see Supplementary Methods for laboratory-specific PCR, Sanger sequencing and bacterial cloning conditions.

RNA-seq analysis

For full information, see Supplementary Methods. In brief, selected RNA samples underwent RNA-seq via Novogene (Hong Kong) using the NEBNext Globin and rRNA Depletion Kit and NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs, MA). At least 70M 150bp paired-end reads (21Gb raw data) per sample were generated on a HiSeq 2000 instrument (Illumina, CA). Raw data were filtered for quality and had adapter sequences removed by Novogene. Reads were aligned to the human genome (GRCh38) using STAR (v2.6.1c)[7] on the University of Southampton's IRIDIS 4 high performance computing cluster and the splicing effects of specific variants was assessed visually using the Integrative Genomics Viewer[8] (Broad Institute, MA) and its inbuilt Sashimi plot function.[9] A threshold of three or more reads was required to call an abnormal splice

event and use of the novel junction had to reach at least 5% read support compared to the alternative canonical junction. Where appropriate, percent spliced in (PSI) values were calculated for abnormal splicing events.[10]

In silico splicing predictions

All variants were assessed bioinformatically for predicted splicing effects using Alamut Visual version 2.11 (Interactive Biosoftware, Rouen, France), which incorporates predictions from MaxEntScan (MES), NNSplice and Splice Site Finder (SSF).[11–13] Individual tools were deemed to predict altered splicing where the change in splice site score was ≥10% (MES) or ≥5% (NNSplice and SSF).[14,15] An overall prediction of altered splicing was called where two out of three Alamut programs agreed. Additional splice prediction information was obtained using Human Splicing Finder (HSF) version 3.1 (threshold ≥0.2) and from publicly available SpliceAI scores (v1.3) for variants (threshold ≥0.2).[16,17] Missing score rate, sensitivity, specificity, overall accuracy, positive and negative predictive values were calculated for each tool individually and for the combined Alamut 2/3 assessment (equations in supplementary methods). The package pROC (v1.15.3)[18] was used in R (v3.5.1)[19] in RStudio[20] to plot receiver operating characteristic (ROC) curves (ggplot2, v3.2.1)[21] for the overlapping set of variants scored by all tools and calculate the area under the curve (AUC) for each tool and for the combined Alamut 2/3 assessment.

RESULTS

Variants affecting splicing

A total of 257 different variants were assessed for their effect on splicing (Table S1). 243 variants were single-nucleotide substitutions, while 14 variants spanned multiple

nucleotides (10 deletions, 1 insertion, 2 deletion-insertions and 1 deletion with an *in cis* SNV). Variants were located across 62 genes in total, with particularly high numbers of variants in *BRCA1* (42), *BRCA2* (42) and *FBN1* (87). In all, 85 variants (33%) were found to be associated with abnormal splicing. 44/57 single-nucleotide substitution variants (77%) located within the donor splice site or splice region (defined by sequence ontology as extending from the third last base of the exon up to the eight base of the intron) and 13/19 single-nucleotide substitution variants (68%) located within the acceptor splice site or splice region (from the eighth last base of the intron up to the third base of the exon) were found to alter splicing (Fig. 1).[22] 175 variants in total did not involve annotated splice regions and of these, 23 (13%) affected splicing (21/167 single-nucleotide substitutions).

39 variants were associated with skipping of the upstream exon (as defined by the location of a variant lying closer to that exon's donor splice site than to an acceptor splice site), which was the most frequent splicing abnormality identified. Only 15 variants were associated with downstream exon skipping, however the analysed variant cohort contained relatively fewer acceptor region variants. These exon skipping figures include three cases in which both upstream and downstream exons were skipped and one case of double upstream exon skipping. 23 variants led to use of an alternative splice donor site and 16 the use of an alternative splice acceptor site, while intron retention was associated with only three variants. For four variants there were multiple splicing abnormalities identified.

Illustrative examples

Several examples from this cohort are pertinent in illustrating the variability in splicing effect seen across different variants (see Fig. 2).

RNA-seq detects a splice variant missed by Sanger sequencing:

This hemizygous *DKC1* c.915+10G>A variant, identified in a male patient with dyskeratosis congenita, produced normal results from direct Sanger sequencing of RT-PCR products (Fig. 2). Similarly, gel electrophoresis did not suggest the presence of more than one RT-PCR product. However, RNA-seq revealed creation of a novel intronic donor splice site, resulting in an insertion of 11 extra nucleotides, which was subsequently confirmed by isoform-specific RT-PCR and sequencing of cloned amplicons. The novel junction had a PSI value of 20%, calculated as the number of length-normalised inclusion reads divided by the total number of length-normalised inclusion and exclusion reads.[10]

A complex deep intronic variant affects splicing:

This heterozygous *P3H1 (LEPRE1)* c.1224-80G>A variant was identified in a patient with osteogenesis imperfecta. RT-PCR revealed a variety of differently sized bands on electrophoresis and PCR product cloning identified at least four alternative splicing events, including intron retention, creation of two novel intronic splice donor sites (inserting 68 or 92 nt), with some additional use of an alternative exonic splice acceptor site (inserting 92 nt intronic sequence but deleting the first 17 nt of exon 8). RNA-seq analysis was only able to confidently identify use of one of the two intronic splice donor sites. Interestingly, the amino acid sequence of any intron retention event (including those utilising the subsequent novel intronic donor site) is predicted to result in introduction of a premature termination codon immediately beyond the end of exon 7.

An apparent canonical splice site variant has no consequence:

A heterozygous canonical splice donor site *DCTN1* c.414+1G>A variant in intron 5 was predicted to disrupt splicing based on NM_004082.4. However, the variant was found to be

present at a relatively high minor allele frequency (MAF) of 3.0×10^{-4} in the Latino population and 6.4×10^{-5} in the gnomAD database (rs576198476). RT-PCR analysis identified that *DCTN1* exons 5-7 are in fact constitutively skipped in both this patient and in controls, negating any potential splicing effects caused by the variant.

A deep exonic cryptic splice site:

This heterozygous *BRCA1* c.4868C>G transversion 119 nt upstream from the donor splice site of *BRCA1* exon 15 is predicted to introduce a conservative alanine to glycine substitution at amino acid 1623. However, RNA analysis shows that this variant in fact creates an exonic cryptic splice donor site at this position, leading to a 119 nt deletion and frameshift of the transcript.

A "likely benign" intronic variant causes pathogenic exon skipping:

A heterozygous non-coding *BRCA1* c.5153-26A>G transition 26 nt upstream from the start of exon 18 is annotated as "likely benign" on ClinVar (rs80358109). However, RNA analysis confirms that this variant induces skipping of the downstream exon 18, resulting in an out-of-frame transcript. Interestingly, although there is no predicted effect on the native splice acceptor site, several prediction tools incorrectly suggest creation of a novel cryptic acceptor site.

A deep intra-exonic splice effect:

A heterozygous *SF3B4* c.417C>T synonymous variant located 254 nt into exon 3 was predicted to lead to an enhanced alternative splice site. RT-PCR analysis confirmed the creation of an alternative deep exonic splice donor site. However, use of this novel donor site was found to be coupled to use of a novel splice acceptor site also within *SF3B4* exon 3,

leading to an intra-exonic deletion of 125 nt. This effect has previously been reported for this variant using a minigene assay.[23]

RNA-seq coverage

17 samples also underwent RNA-seq analysis. In four cases, RNA-seq was able to detect a splicing abnormality consistent with initial RT-PCR results. In one case (*DKC1*), RNA-seq identified a splicing abnormality not initially detected by RT-PCR. In another case (*SF3B4*), the splicing abnormality seen by RT-PCR was only seen in two RNA-seq reads and so fell below the reporting threshold. In 11 other cases, no reportable splicing abnormality was detected. Of note, splice junction depth of coverage varied considerably across assayed genes and also within genes, which in several cases limited the ability of RNA-seq to detect low-level splice junction usage.

Bioinformatic splicing predictions

We scored all variants with Alamut Visual (v2.11), including MES, NNSplice and SSF, and also with HSF and SpliceAI. Thresholds for change were selected above which a variant was deemed to be predicted to be splice affecting based on previous literature.[14,15,17] A combined Alamut score was also calculated, where a variant was deemed to be predicted as splice affecting if two out of three individual tools within Alamut passed the defined threshold. The overall sensitivity, specificity, accuracy, positive and negative predictive values for each tool and the combined Alamut assessment are given in Table 1, based on all variants that were scored by each method. Figure 3 shows ROC curves with AUC values based on the overlapping set of variants scored by all tools. SpliceAI performed the best in predictions of splicing disruption of all the tools/approaches across all of the considered metrics, with overall accuracy exceeding 90% (see Table 1).

DISCUSSION

VUS clarification and clinical impact through splicing analysis

This study has helped to clarify the effects on splicing of over 250 VUSs in clinically important disease genes. 34% of these VUSs were found to affect splicing. However, while this overall figure is certainly within the range of previous estimates for the proportion of variants affecting splicing, it should be noted that this cohort of variants was specifically selected for splicing studies. As such, there will have been some intrinsic bias in selection, since we expect variants were more likely to be referred for RNA studies if they fell within a splice region or if clinical diagnostic laboratories had already highlighted a potential predicted effect on splicing. Furthermore, the prior probability of these patients having a pathogenic variant in the genes tested is likely to be increased, since UK diagnostic genetic testing generally requires that a patient's phenotype potentially fits with the genes being tested. Nevertheless, this cohort does represent a true-to-life set of clinically identified VUSs for which clarification of pathogenicity was sought by referring clinicians.

The results of this study show that RNA splicing analysis, using RT-PCR or transcriptomics, has the ability to produce clear results that help clarify variant interpretation. Where abnormal splicing is detected, this analysis constitutes a functional assay that provides supporting evidence of pathogenicity.[5] In many such cases, these results therefore have direct clinical utility by allowing a genetic diagnosis to be made. Indeed, the results of at least one of these cases (*AARS*) has been used to inform prenatal testing in a subsequent pregnancy.

Only 30% of the variants in this study fell within annotated splice regions, while 13% of non-splice region variants still affected splicing. This highlights the need to consider possible splicing effects whenever deep exonic or deep intronic variants are identified. With increasing use of genome sequencing, increasing numbers of intronic variants will be identified through clinical diagnostic testing. Interpretation of such variants beyond the splice region remains largely uncertain. However, through RNA analysis, potential splicing effects of such variants can be detected.

Furthermore, a number of these results illustrate the danger of assuming the effects of splice site variants. The *DCTN1* c.414+1G>A example is a case in point of a benign canonical splice site variant and our cohort also includes two normal *BRCA2* canonical splice site variants (*BRCA2* c.6937+1G>T and *BRCA2* c.8331+2T>C) that do not appear to cause abnormal splicing (with the caveat that splicing effects in blood may potentially differ from those in other tissues). In addition, the *SF3B4* example shows how difficult it can be to predict splice junction usage, since even if one correctly identifies creation of a cryptic donor site, one may not necessarily predict the acceptor site it will use. This particular variant appears to create a type of non-canonical splicing event known as an "exitron", where a novel intron is defined entirely within a large exon.[24]

Targeted testing and transcriptome-wide analysis

Our analysis helps provide some insight into the comparative use of RT-PCR and RNA-seq to look at splicing. Compared to transcriptome-wide RNA-seq, RT-PCR should generally prove more sensitive for detecting substantial splicing abnormalities such as exon skipping, since targeted amplification allows a very low limit of detection. However, endpoint RT-PCR and Sanger sequencing are not truly quantitative methods and suffer from biases such as

preferential amplification of shorter products. Whole transcriptome RNA-seq, conversely, may provide more reliable quantification of splice isoforms through calculated read-based metrics such as PSI values.[10] On the other hand, transcriptome-wide RNA-seq is intrinsically limited in its depth of coverage by the number of reads obtained per sample, particularly where a gene is poorly expressed. A number of RNA-seq samples in this cohort did indeed have relatively poor coverage across the target region for the variants in question. However, where a splice abnormality results in a small-scale change, for example insertion of a few nucleotides as seen with *DKC1* c.915+10G>A, RNA-seq may succeed in identifying this where Sanger sequencing of PCR products fails. Small-scale splicing changes are easily missed on gel electrophoresis and coupled with the poor sensitivity of Sanger sequencing to detect low-level sample heterogeneity, this is an instance where RNA-seq can outperform RT-PCR. Another potential approach to raise coverage depth could be to perform a targeted RNA-seq library prep focussed on the gene region of interest. However, this would be at the expense of RNA-seq's other great advantage; the ability to look for alternative pathogenic splicing events or even alternative pathogenic sequence variants in the same or in other genes.

Bioinformatic tool comparison

The ability to accurately predict the affect a given sequence variant will have on splicing is highly desirable in prioritising variants for functional validation, or even as a diagnostic assessment in its own right. However, despite a multitude of different prediction methods being available, there is little consensus on the best tools or the optimal usage and score thresholds to use. A common approach is to score a variant with several (three-five) tools and take a consensus approach – if the majority of tools predict an effect, the variant is

predicted to be splice affecting. In our assessment, we found little benefit of this consensus approach over the use of individual tools. Across all scored variants the Alamut 2/3 consensus gave similar sensitivity and specificity to component tools MES and SSF, and gave a comparable AUC in the analysis considering the overlapping variant set that were scored by all tools. The newer, machine learning based approach, SpliceAI, outperformed the other tools across metrics, classifying over 90% of variants consistently with the experimental data. Our data suggest this method could assist in clinical interpretation of variants potentially affecting splicing, and offer benefits over existing approaches that are currently in use diagnostically.

Despite questions over the accuracy and applicability of *in silico* splice prediction tools, in this cohort, a high proportion of variants were correctly predicted to alter normal splicing, particularly given the high proportion of variants outside of the immediate splicing area.[14,15] However, this is likely to be at least partially explained by the bias in case selection, since we expect variants were more likely to be referred for splicing analysis where diagnostic genetic test reports had predicted a possible splicing effect.

Limitations of testing and using blood as a proxy tissue

In analysing blood RNA, there are intrinsic limitations. Most obviously, only genes that are transcribed in blood can be detected. Genes that are highly tissue-specific in their expression can therefore prove problematic to analyse. Alternative cell types may be available in some cases from skin or muscle biopsies and RNA from such sources has been successfully used for splicing analysis.[25,26] However, even in the absence of such samples, low-level basal transcription of the genome is known to take place and some 80% of all human coding sequences have been identified in blood.[27] In this study, reference was made

to GTEx transcript per million (TPM) values (Table S1).[28,29] Interestingly, informative RT-PCR results were obtainable for a number of genes reported to have TPM values of zero (*FBN2*, *COL3A1*, *COL4A1*, *COL5A1*), although this is not reliably the case for all such genes.

A further important consideration is the tissue-specificity of splicing. Use of blood as a proxy tissue assumes that similar splicing events are taking place in clinically relevant tissues, which is not necessarily the case. Another limitation in detection may occur if nonsense-mediated decay (NMD) is efficient enough to remove all abnormally spliced transcripts from a sample. Indeed, variability in NMD contributes to uncertainty in quantifying the relative usage of aberrant splice events.[30,31] This means that simple quantification metrics of splice site usage are unlikely to be directly informative of pathogenicity and need to be considered in comparison to control samples.

Mechanistic insights into splicing

A notable finding in this study is that splice-altering variants located close to the donor splice site tend to cause skipping of the upstream exon. In considering the splicing reaction, where the donor splice site is first cleaved and ligated to the intronic branch point to form a lariat, one might expect a disrupted donor splice site to cause intron retention. However, retention of introns appears to be a relatively rare event in this study. Furthermore, the presence of upstream exon skipping in these cases implies that splicing of the preceding intron has not yet been completed by the time the next intron is being spliced. If upstream splicing were complete, there would be no upstream donor splice site available to allow exon skipping to take place (Fig. 4), except in the setting of a recursive splicing mechanism.[24]

Splicing is known to occur co-transcriptionally and the choice of splice site depends not only on sequence but also on additional factors such as rate of transcription, RNA secondary

structure, chromatin conformation and the effects of splicing enhancers and silencers.[32] It may be that some of these factors are playing a role in driving the upstream exon skipping that predominates in this variant cohort. The timing of splicing events may also potentially be influenced somewhat by intron length. However, analysis of the intron-exon structure around these variants did not indicate any significant skewing of upstream versus downstream intron length.

Further work will be needed to better characterise the mechanistic and regulatory elements of the abnormal splicing seen in this study. Understanding the underlying mechanisms governing such splicing abnormalities is critical, not only to allow their better prediction but also to inform therapeutic approaches that aim to correct them. Splice-switching antisense oligonucleotide (ASO) therapies are increasingly being developed for clinical use and their design depends upon accurate targeting of disease-specific splice sites or splice-regulatory elements.[33,34] The sequence-specificity of this approach lends itself ideally to personalised medicine and indeed such a drug has recently been developed for an N-of-1 study in a single patient with a deep intronic variant affecting splicing.[35] In the appropriate disease settings, splice-affecting variants lying within deep intronic or exonic regions therefore represent particularly good candidates for the development of splice-switching ASO therapeutic approaches.

Conclusion

This variant cohort is among the largest and most diverse to have had experimentally determined RNA splicing effects analysed and published to date. While routine use of RNA analysis in genetic diagnostics requires further work to clarify the service implications, based on this study, we recommend that RNA-based splicing analysis be at least routinely

considered in genetic disease variant interpretation in order to improve diagnostic uplift. While bioinformatic splicing prediction tools, particularly SpliceAI, continue to improve in accuracy, there is still significant miscalling of predictions from all tools. Ideally, they should therefore not be relied upon in isolation in assessing a variant's effect on splicing and their predictions should not be a prerequisite line of evidence for classifying splice variants, should clear experimentally obtained RNA splicing evidence be available. Owing to the subtlety and complexity of RNA splicing, additional work will be required in order to determine how best to incorporate splicing predictions and experimental splicing analysis into variant classification guidelines.

In conclusion, this large study demonstrates the potential of blood RNA analysis in clarifying the effects of variants of unknown significance and the uplift of diagnostic rate.

ACKNOWLEDGEMENTS

This research was funded by NIHR and the NewLife Foundation. The Baralle lab. is supported by NIHR Research Professorship to DB RP-2016-07-011. The authors thank all patients and families taking part in this research, and acknowledge the NIHR Clinical Research Network (CRN) in recruiting patients and the Muskateers Memorandum, as well as support from the NIHR UK Rare Genetic Disease Consortium. We thank staff from regional genetics services who recruited patients: Birmingham Women's and Children's NHS Foundation Trust (Swati Naik, Nicola Ragge, Helen Cox, Jenny Morton, Mary O'Driscoll, Derek Lim, Deborah Osio, Camilla Huber, Julie Hewitt); St George's University Hospitals NHS Foundation Trust (Heidy Brandon, Meriel McEntagart, Sahar Mansour, Nayana Lahiri, Esther Dempsey, Merrie Manalo, Tessa Homfray, Anand Saggar); University Hospitals of Leicester

NHS Trust (Jin Li, Julian Barwell); Manchester University NHS Foundation Trust (Kate Chandler, Tracy Briggs, Sofia Douzgou), Leeds Teaching Hospital NHS Trust (Julian Adlard, Alison Kraus); Cambridge University Hospitals NHS Foundation Trust (Sarju Mehta); University Hospitals Bristol NHS Foundation Trust (Amy Watford, Alan Donaldson, Karen Low); Nottingham University Hospitals NHS Trust (Gabriela Jones, Abhijit Dixit, Elizabeth King, Nora Shannon); Great Ormond Street Hospital for Children NHS Foundation Trust (Marios Kaliakatsos); Guys and St Thomas' NHS Foundation Trust (Merrie Manalo); NHS Greater Glasgow and Clyde (Shelagh Joss); Sheffield Children's NHS Foundation Trust (Meena Balasubramanian, Diana Johnson); Royal Devon and Exeter NHS Foundation Trust (Sarah Everest); University Hospital Southampton NHS Foundation Trust (Claire Salter, Victoria Harrison, Gillian Wise, Audrey Torokwa, Victoria Sands, Esther Pyle, Tessy Thomas, Katherine Lachlan, Nicola Foulds, Diana Baralle, Andrew Lotery, Andrew Douglas, Simon Hammans, Emily Pond, Rachel Horton, Mira Kharbanda, David Hunt, Charlene Thomas, Lucy Side, Catherine Willis, Stephanie Greville-Heygate, Rebecca Mawby, Catherine Mercer, Karen Temple, Esther Kinning); University of Bergen, Norway (Ognjen Bojovic); L. Archer. 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.

REFERENCES

1. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46(D1):D1062-D1067. doi:10.1093/nar/gkx1153
2. Wright CF, FitzPatrick DR, Firth H V. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet*. 2018;19(5):253-268. doi:10.1038/nrg.2017.116

3. Clift K, Macklin S, Halverson C, McCormick JB, Abu Dabrh AM, Hines S. Patients' views on variants of uncertain significance across indications. *J Community Genet*. 2019. doi:10.1007/s12687-019-00434-7

4. López-Bigas N, Audit B, Ouzounis C, Parra G, Guigó R. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett*. 2005;579(9):1900-1903. doi:10.1016/j.febslet.2005.02.047

5. Richards S, Aziz N, Bale 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*. 2015;17(5):405-423. doi:10.1038/gim.2015.30

6. Jian X, Boerwinkle E, Liu X. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet Med*. 2014;16(7):497-503. doi:10.1038/gim.2013.176

7. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

8. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative Genomics Viewer. *Nat Biotechnol*. 2011;29(1):24-26. doi:10.1038/nbt.1754

9. Katz Y, Wang ET, Silterra J, et al. Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics*. 2015;31(14):2400-2402. doi:10.1093/bioinformatics/btv034

10. Schafer S, Miao K, Benson CC, Heinig M, Cook SA, Hubner N. Alternative splicing signatures in RNA-seq data: percent spliced in (PSI). *Curr Protoc Hum Genet*. 2015;87(October):11.16.1-11.16.14. doi:10.1002/0471142905.hg1116s87

11. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with

applications to RNA splicing signals. *J Comput Biol*. 2004;11(2-3):377-394. doi:10.1089/1066527041410418

12. Reese MG, Eechman FH, Kulp D, Haussler D. Improved splice site detection in Genie. *J Comput Biol*. 1997;4(3):311-323. doi:10.1089/cmb.1997.4.311
13. Shapiro MB, Senapathy P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res*. 1987;15(17):7155-7174. doi:10.1093/nar/15.17.7155
14. Houdayer C, Caux-Moncoutier V, Krieger S, 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*. 2012;33(8):1228-1238. doi:10.1002/humu.22101
15. Tang R, Prosser DO, Love DR. Evaluation of bioinformatic programmes for the analysis of variants within splice site consensus regions. *Adv Bioinformatics*. 2016;2016:5614058. doi:10.1155/2016/5614058
16. Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res*. 2009;37(9):e67-e67. doi:10.1093/nar/gkp215
17. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting splicing from primary sequence with deep learning. *Cell*. 2019;176(3):535-548. doi:10.1016/j.cell.2018.12.015
18. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77. doi:10.1186/1471-2105-12-77
19. R Core Team. R: A language and environment for statistical computing. 2018.

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

20. RStudio Team. RStudio: Integrated Development for R. 2015. http://www.rstudio.com/.
21. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2016.
22. Eilbeck K, Lewis SE, Mungall CJ, et al. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol*. 2005;6(5):R44. doi:10.1186/gb-2005-6-5-r44
23. Cassina M, Cerqua C, Rossi S, et al. A synonymous splicing mutation in the SF3B4 gene segregates in a family with highly variable Nager syndrome. *Eur J Hum Genet*. 2017;25(3):371-375. doi:10.1038/ejhg.2016.176
24. Sibley CR, Blazquez L, Ule J. Lessons from non-canonical splicing. *Nat Rev Genet*. 2016;17(7):407-421. doi:10.1038/nrg.2016.46
25. Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med*. 2017;9(386). doi:10.1126/scitranslmed.aal5209
26. Gonorazky HD, Naumenko S, Ramani AK, et al. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am J Hum Genet*. 2019;104(3):466-483. doi:10.1016/j.ajhg.2019.01.012
27. Liew CC, Ma J, Tang HC, Zheng R, Dempsey AA. The peripheral blood transcriptome dynamically reflects system wide biology: A potential diagnostic tool. *J Lab Clin Med*. 2006;147(3):126-132. doi:10.1016/j.lab.2005.10.005
28. GTEx Consortium. GTEx pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348(6235):648-660. doi:10.1126/science.1262110
29. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression

estimation with read mapping uncertainty. *Bioinformatics*. 2009;26(4):493-500. doi:10.1093/bioinformatics/btp692

30. Nguyen LS, Wilkinson M, Gecz J. Nonsense-mediated mRNA decay: inter-individual variability and human disease. *Neurosci Biobehav Rev*. 2014;46 Pt 2(1):175-186. doi:10.1016/j.neubiorev.2013.10.016

31. Miller JN, Pearce DA. Nonsense-mediated decay in genetic disease: friend or foe? *Mutat Res - Rev Mutat Res*. 2014;762(605):52-64. doi:10.1016/j.mrrev.2014.05.001

32. Douglas AGL, Wood MJA. RNA splicing: disease and therapy. *Brief Funct Genomics*. 2011;10(3):151-164. doi:10.1093/bfgp/elr020

33. Crooke ST, Witztum JL, Bennett CF, Baker BF. RNA-Targeted Therapeutics. *Cell Metab*. 2018;27(4):714-739. doi:10.1016/j.cmet.2018.03.004

34. Pitout I, Flynn LL, Wilton SD, Fletcher S. Antisense-mediated splice intervention to treat human disease: the odyssey continues [version 1; peer review: 3 approved]. *F1000Research*. 2019;8(F1000 Faculty Rev):710. doi:10.12688/f1000research.18466.1

35. Kim J, Hu C, El Achkar CM, et al. Patient-customized oligonucleotide therapy for a rare genetic disease. *N Engl J Med*. 2019;381(17):1644-1652. doi:10.1056/NEJMoa1813279

Figure 1. **Variant locations and effects on splicing.** **A**. Plot of the numbers of SNVs in this cohort (multi-nucleotide variants not included) present at each donor (D-3 to D+8) and acceptor (A-8 to A+3) splice region position, along with the numbers of these found to affect splicing. **B** and **C**. Position-weight matrices of nucleotide sequence across the splice donor (**B**) and acceptor (**C**) regions as determined for the specific exon-intron junctions analysed in

this study. In this representation, the donor splice site +1 position correlates to position 12 in **B**, while the acceptor splice site -1 position correlates to position 25 in **C**. **D**. Abnormal splicing effects plotted by SNV location. Sequence ontology defines the donor splice region as extending from the third last nucleotide of the exon (D-3) to the eighth nucleotide of the intron (D+8) and the acceptor splice region as extending form the eighth last nucleotide of the intron (A-8) to the third nucleotide of the exon (A+3).[22] **E**. Overall proportion of all variants affecting splicing in this cohort. **F**. Proportions of different abnormal splicing events identified in this cohort. SE, skipped exon; A5SS, alternative 5´ splice site; A3SS, alternative 3´ splice site; IR, intron retention.

Figure 2. Illustrative examples of variant splicing analysis. *DKC1* c.915+10G>A could not be identified by RT-PCR and Sanger sequencing but alternative donor splice site usage was identified by RNA-seq. *P3H1* (*LEPRE1*) c.1224-80G>A causes at least three abnormal splicing events using alternative splice donor and acceptor sites, as well as increasing levels of intron retention. *DCTN1* c.414+1G>A appears to alter a canonical splice donor site but exons 5-7, although annotated, are never expressed and are constitutively spliced out. *SF3B4* c.417C>T is a synonymous coding variant but causes formation of a 125 nt "exitron", an intronic region within an exon. Pt, patient; Ctrl, control; A5SS, alternative 5´ splice site; A3SS, alternative 3´splice site.

Figure 3. Bioinformatic tools for predicting abnormal splicing. Receiver operating characteristic (ROC) curves and area under the curve (AUC) comparing *in silico* methods for predicting splice disruption in overlapping set of experimentally validated variants scored by all measures (136 non-splice disrupting, 70 splice disrupting). HSF = human splicing finder,

MES = MaxEntScan (Alamut), NN = NNSplice (Alamut), SSF = SpliceSiteFinder (Alamut), Ala23 = number of Alamut tools exceeding specified thresholds.

Figure 4. A potential model of splicing disruption. Where an upstream splicing event is complete, a splice donor or acceptor site variant may lead to intron retention. Where a preceding splicing event remains incomplete, a splice donor variant may cause skipping of the upstream exon. Similarly, if a splice acceptor site variant causes an upstream splice donor site to remain unused then this may cause skipping of the exon downstream of the acceptor site variant. Exonic or intronic variants that create or strengthen cryptic splice sites can lead to use of alternative splice donor or acceptor sites.

Table 1. Performance assessment of *in silico* prediction tools on experimentally validated variants (n=257). Values have been calculated omitting the missing scores for each tool.

Scoring Metric	n missing	Sensitivity	Specificity	Accuracy	PPV	NPV
HSF (2%)	28	0.8941	0.3958	0.5808	0.4663	0.8636
SpliceAI (0.2)	11	0.8987	0.9162	0.9106	0.8353	0.9503
Alamut SSF (5%)	5	0.7317	0.9294	0.8651	0.8333	0.8778
Alamut MES (10%)	1	0.7381	0.9070	0.8516	0.7949	0.8764
Alamut NNSplice (5%)	11	0.6923	0.8631	0.8089	0.7013	0.8580
Alamut 2/3	14*	0.7237	0.9162	0.8560	0.7971	0.8793

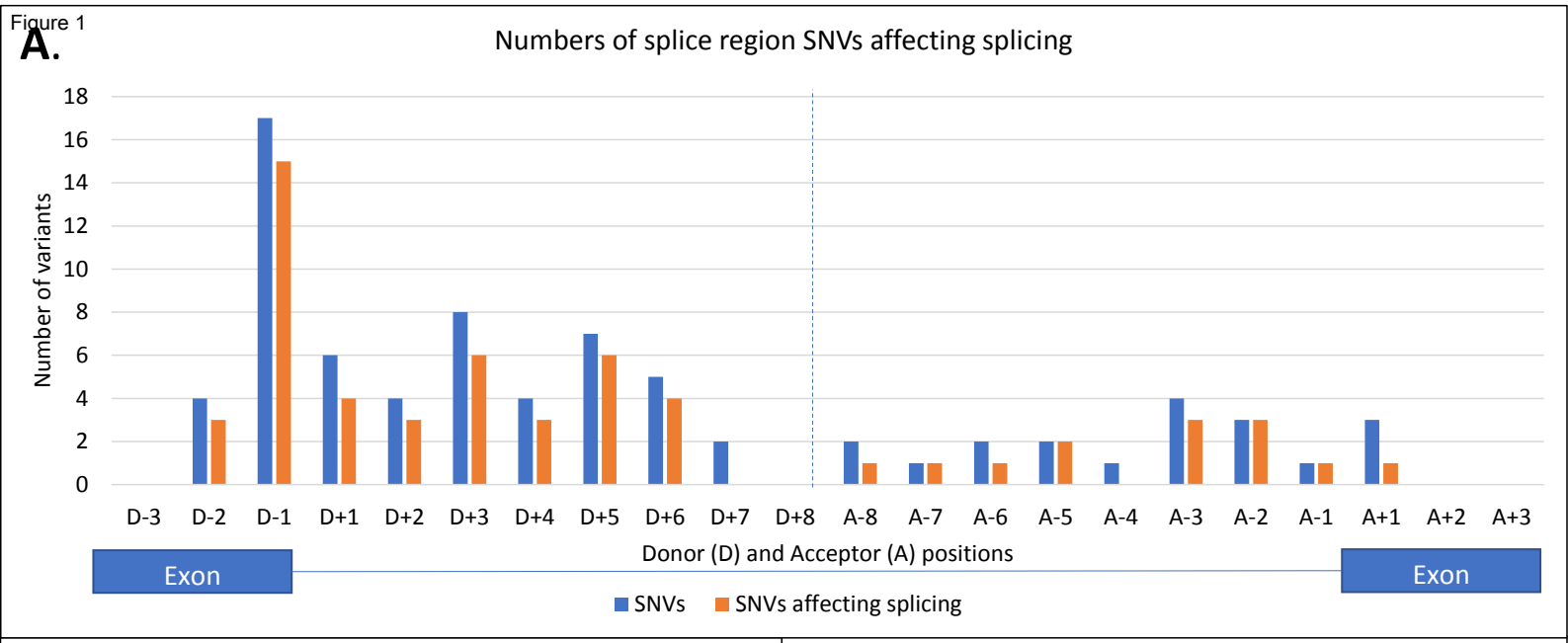
* 11 variants missing one score, three variants missing two scores

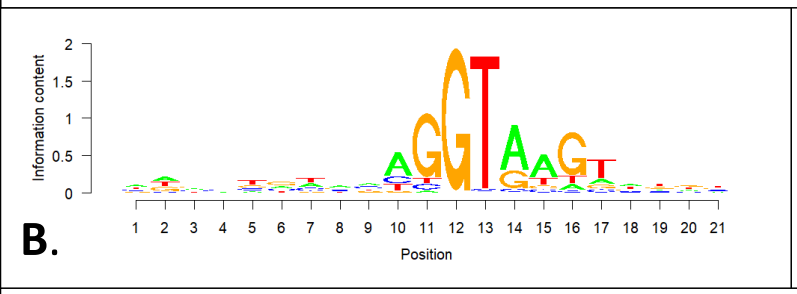
17th December 2019

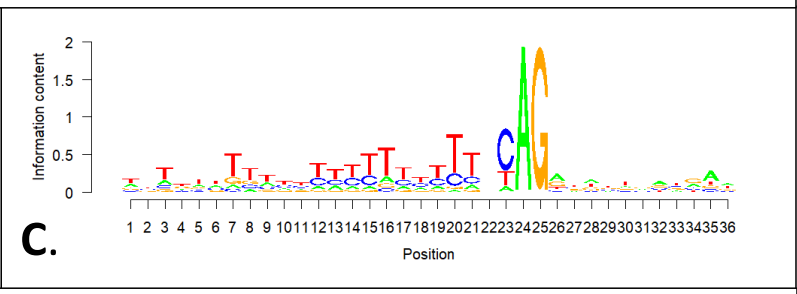
The authors declare no conflict of interest.

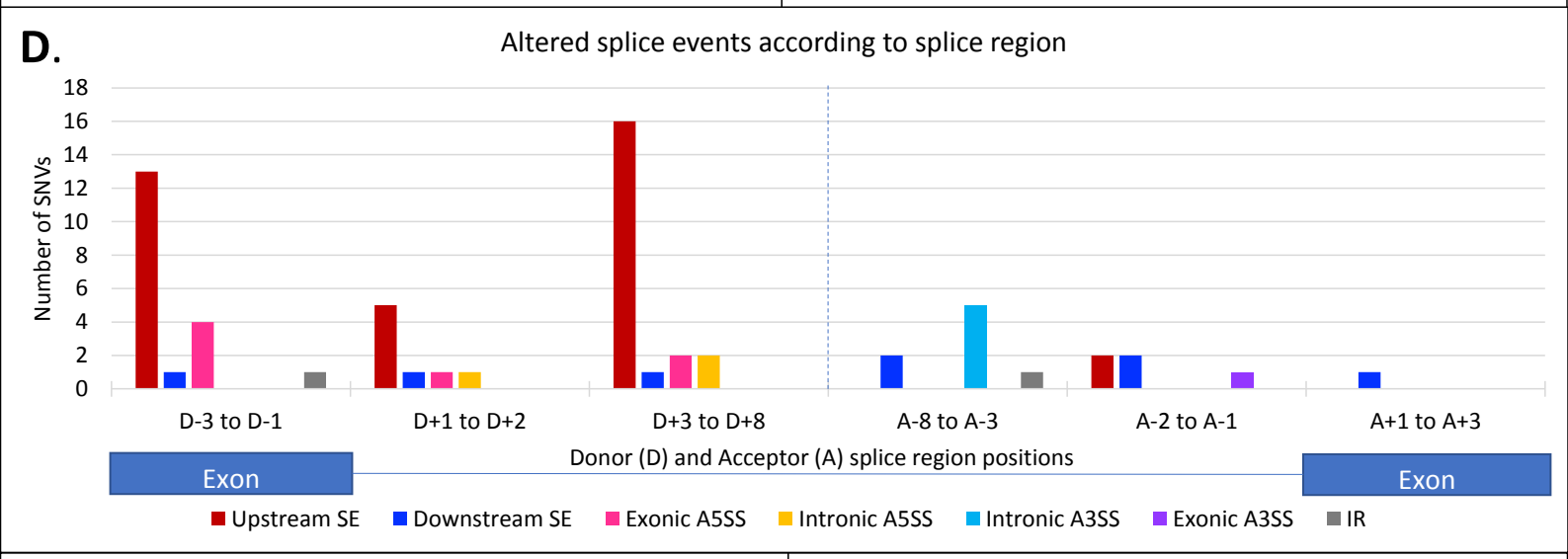
Prof Diana Baralle
For and on behalf of the authors

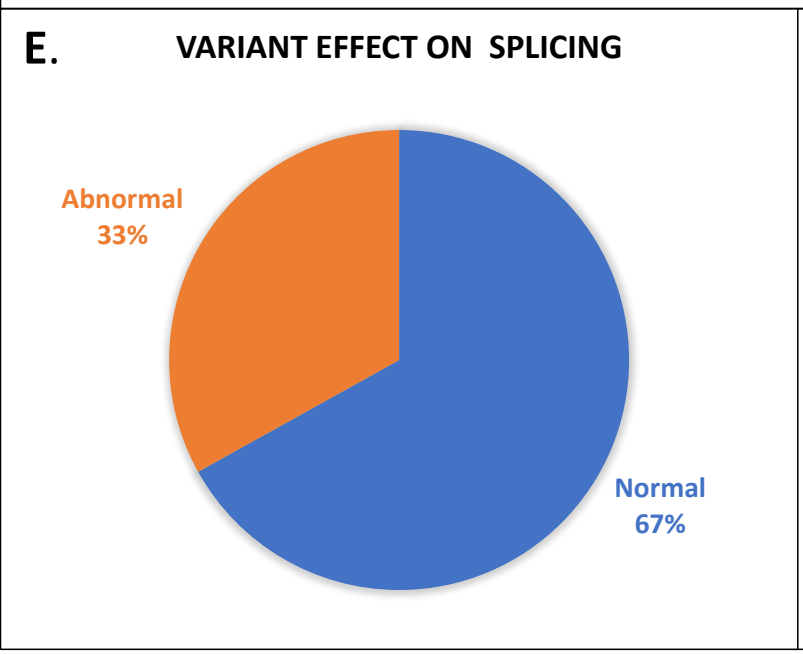
Figure 1


A.
Numbers of splice region SNVs affecting splicing
Number of variants
18
16
14
12
10
8
6
4
2
0
D-3 D-2 D-1 D+1 D+2 D+3 D+4 D+5 D+6 D+7 D+8 A-8 A-7 A-6 A-5 A-4 A-3 A-2 A-1 A+1 A+2 A+3
Donor (D) and Acceptor (A) positions
Exon
Exon
SNVs
SNVs affecting splicing


Information content
2
1.5
1
0.5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Position
B.


Information content
2
1.5
1
0.5
0
Position
C.


D.
Altered splice events according to splice region
Number of SNVs
18
16
14
12
10
8
6
4
2
0
D-3 to D-1
D+1 to D+2
D+3 to D+8
A-8 to A-3
A-2 to A-1
A+1 to A+3
Donor (D) and Acceptor (A) splice region positions
Exon
Exon
Upstream SE
Downstream SE
Exonic A5SS
Intronic A5SS
Intronic A3SS
Exonic A3SS
IR


E.
VARIANT EFFECT ON SPLICING
Abnormal
33%
Normal
67%

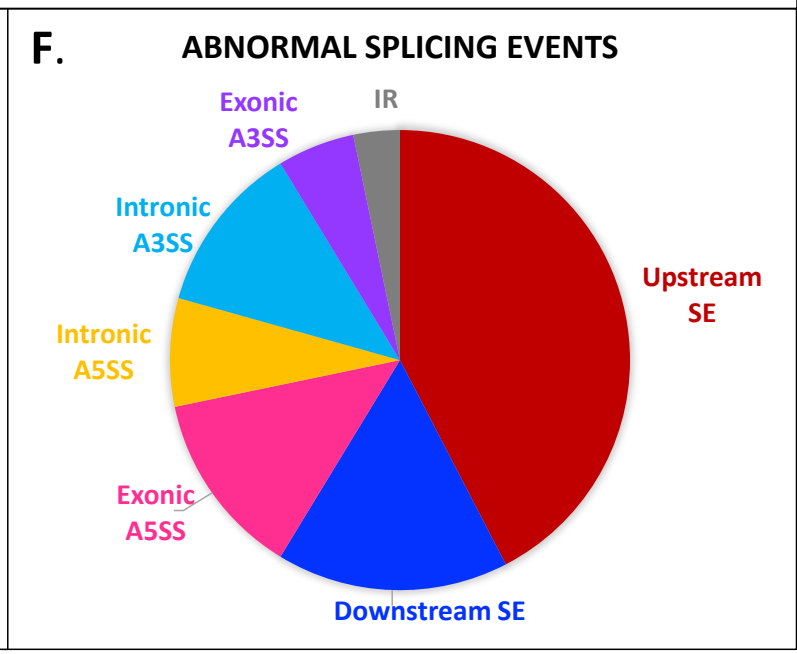
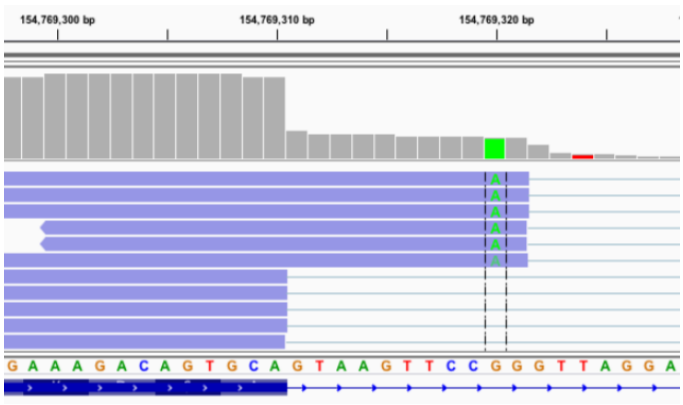
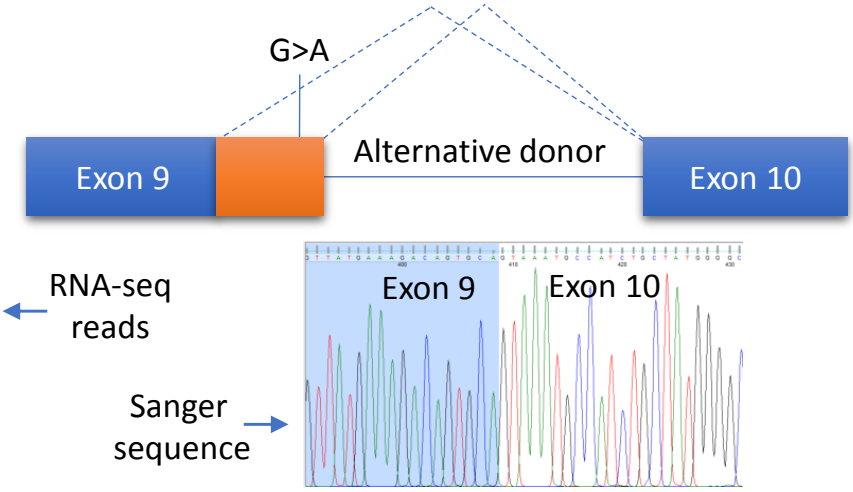

F.
ABNORMAL SPLICING EVENTS
Exonic
A3SS
IR
Intronic
A3SS
Upstream
SE
Intronic
A5SS
Exonic
A5SS
Downstream SE

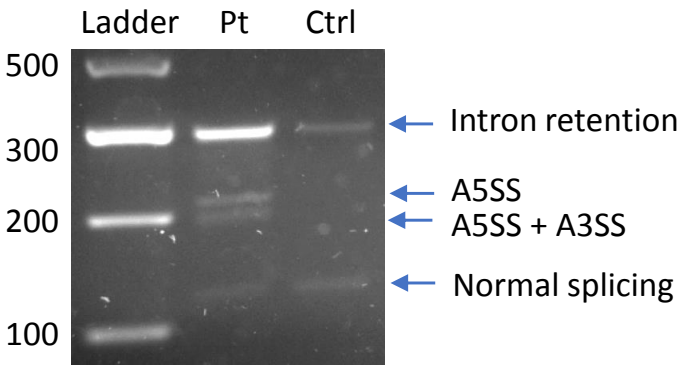
Figure 2

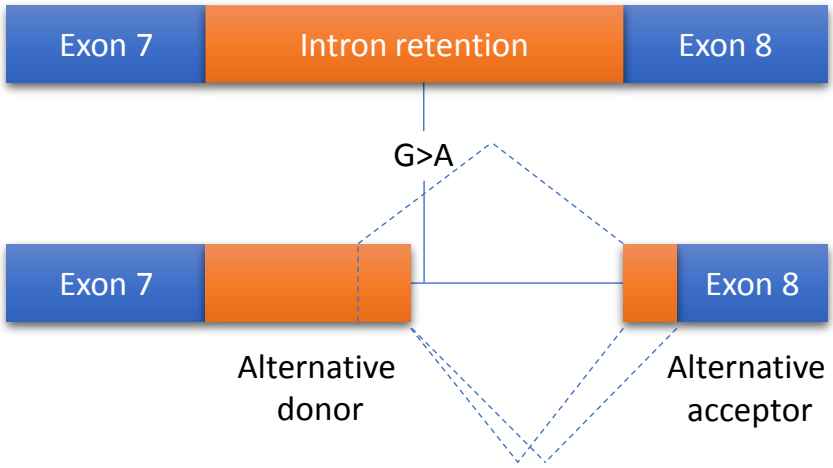
DKC1 c.915+10G>A



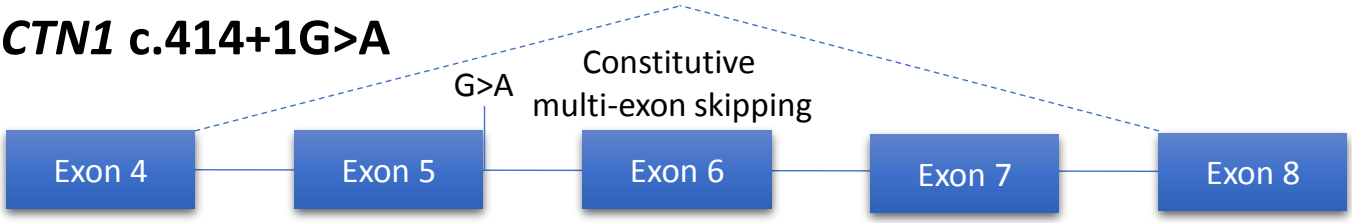


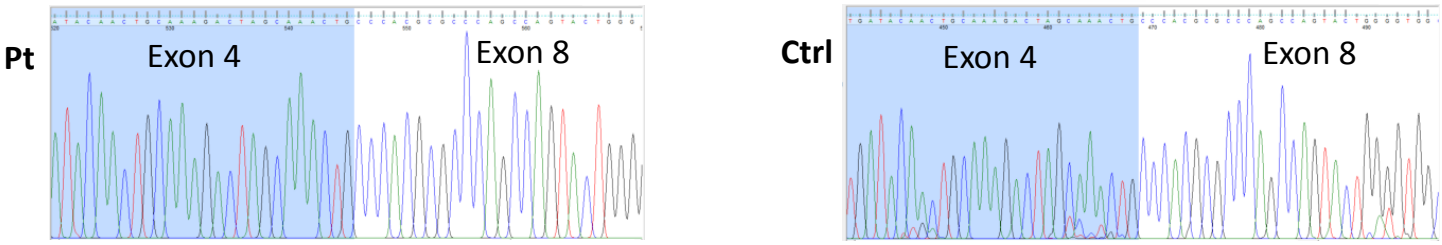
P3H1 (LEPRE1) c.1224-80G>A



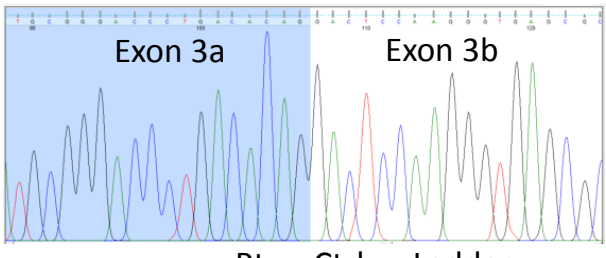


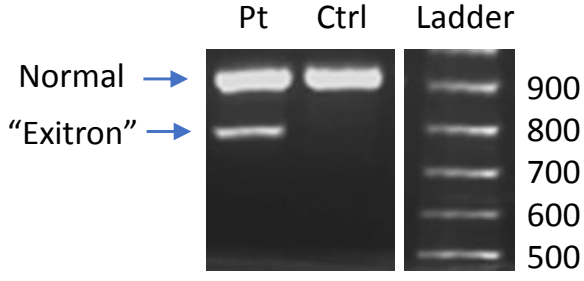
DCTN1 c.414+1G>A





SF3B4 c.417C>T





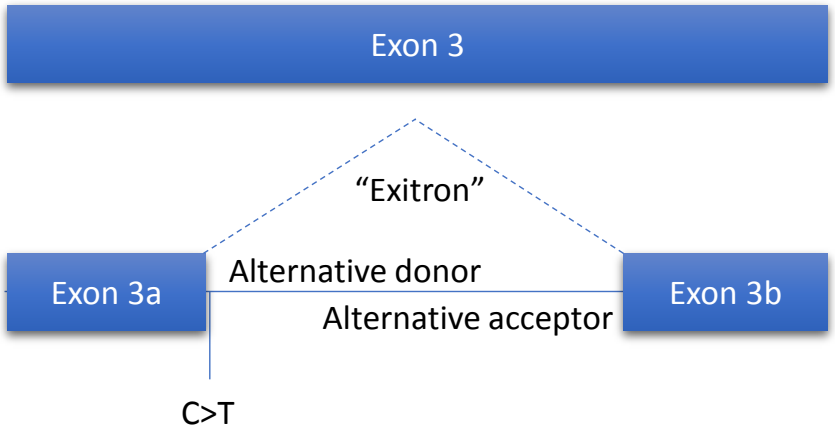
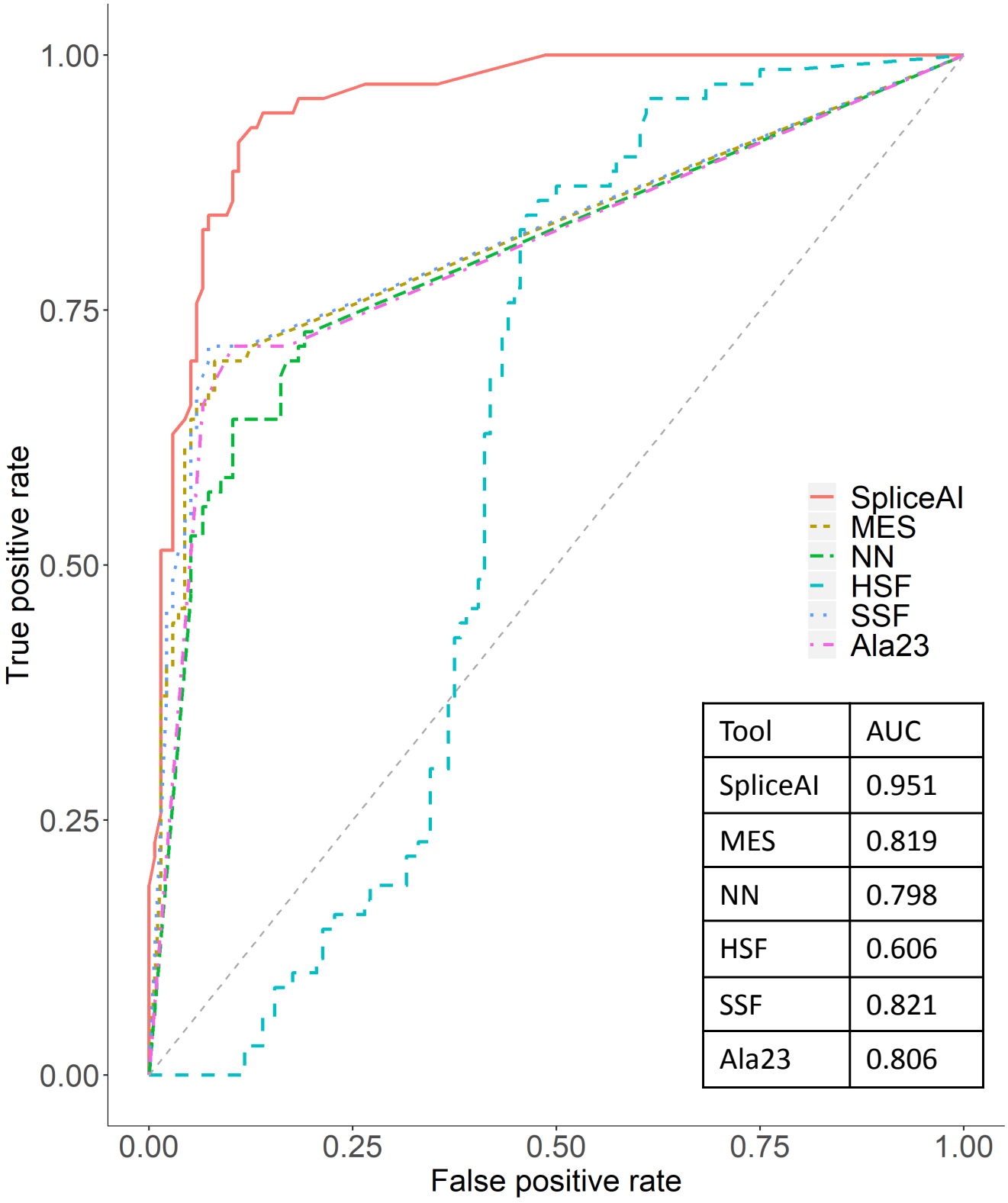
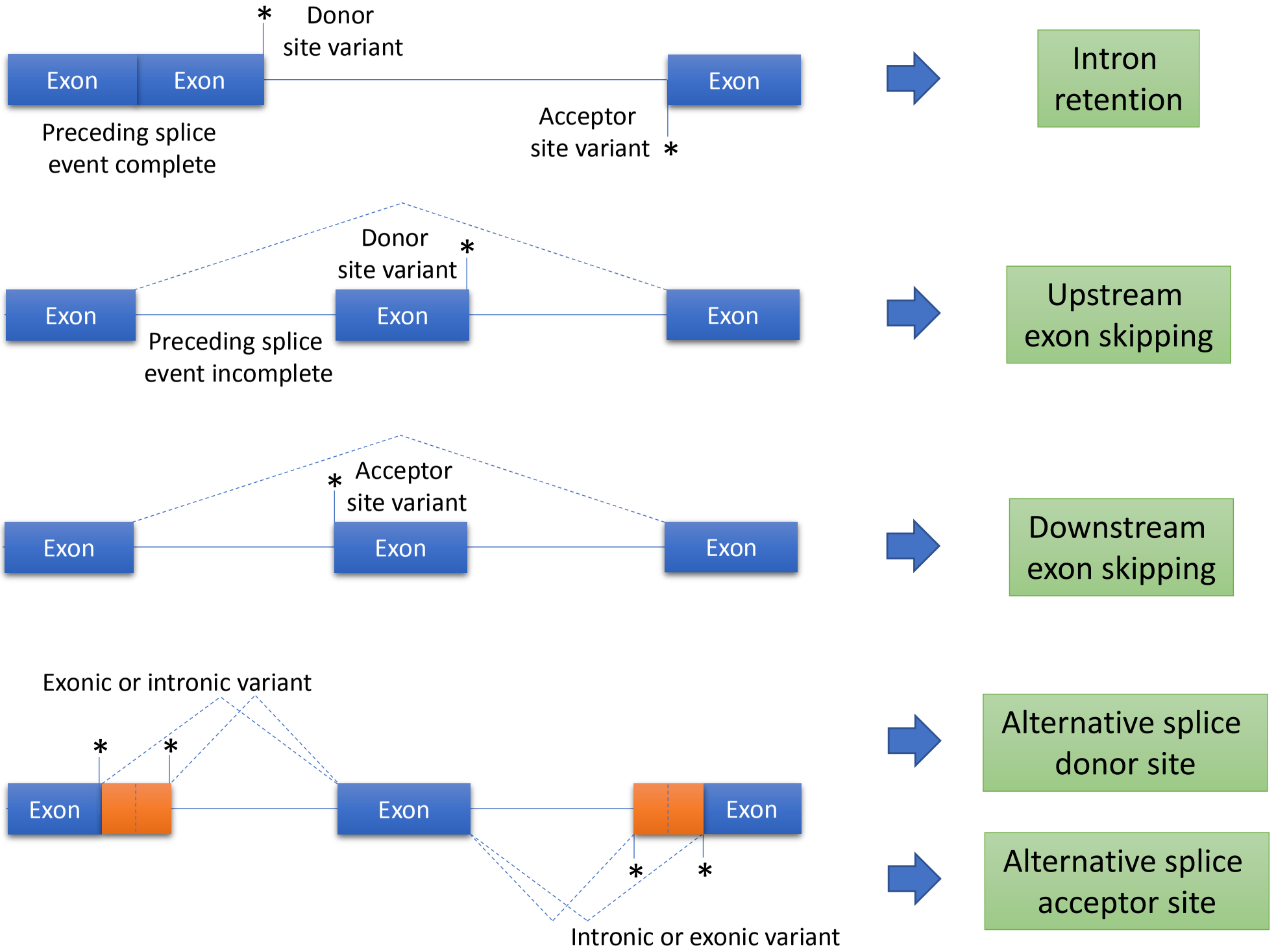


Figure 3



Tool	AUC
SpliceAI	0.951
MES	0.819
NN	0.798
HSF	0.606
SSF	0.821
Ala23	0.806

figure 4


Donor
site variant
Exon
Exon
Exon
Preceding splice
event complete
Acceptor
site variant
Intron
retention
Donor
site variant
Exon
Exon
Exon
Preceding splice
event incomplete
Upstream
exon skipping
Acceptor
site variant
Exon
Exon
Exon
Downstream
exon skipping
Exonic or intronic variant
Exon
Exon
Exon
Intronic or exonic variant
Alternative splice
donor site
Alternative splice
acceptor site

SUPPLEMENTARY INFORMATION

SUPPLEMENTARY METHODS:

Salisbury RNA extraction and RT-PCR protocol

Blood samples were collected into PAXgene RNA collection tubes (Qiagen, UK) and RNA was extracted using the QIAcube Connect extraction machine (Qiagen, UK) and the standard protocol of the RNeasy kit (Qiagen, UK). cDNA preparation was then carried out using 1.5µl of 10mM dNTP mix (Promega, UK), 2µl of 0.1M Dithiothreitol (Invitrogen, UK), 1µl of 40 U/µl Moloney Murine Leukemia Virus Reverse Transcriptase (Invitrogen, UK), 1µl of 40 U/µl RNAaseOut ribonuclease inhibitor (Invitrogen, UK), 4µl of Reverse Transcriptase buffer (Invitrogen, UK), 1µl of 10ng/µl random hexamer primers (Thermo Fisher, UK) and 10µl of RNA. Samples were incubated for one hour at 37^{o}C followed by 10mins at 65^{o}C. For each variant, where possible, RNA analysis was carried out by using a forward PCR primer situated at least two exons upstream from the exon (or flanking intronic sequence) containing the variant and a reverse primer at least two exons downstream from the exon of interest. This was subject to the resultant PCR fragment being of a reasonable size for Sanger sequencing (ideally below about 600 base pairs), and for those genes with relatively small exons the primers were situated further away where possible (all primer sequences available upon request). PCR reactions were carried out in a 20µl volume containing 1.5µl of cDNA, 10nM of each primer (Promega, UK), 2µl of 10x Platinum Taq buffer (Invitrogen, UK), 0.2mM of each dNTP, 1.5mM $MgCl_2$ and 0.5 units of Platinum Taq polymerase (Invitrogen, UK). Cycling parameters were 94^{o}C for 12 minutes followed by 35 cycles of 94^{o}C for 30 seconds, 60^{o}C for 30 seconds and 72^{o}C for 30 seconds. PCR products were checked by gel electrophoresis and then bi-directionally sequenced using the standard protocol of the

Big-Dye® Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems, USA) and separated on an ABI 3130*xl* Genetic Analyzer (Applied Biosystems, USA). Subsequent data were analysed using the Mutation Surveyor (version 3.1) software (SoftGenetics, USA).

Exeter RNA extraction and RT-PCR protocol

RNA was extracted from whole blood (PAXgene Blood RNA tube; Qiagen 762165) on the QIAcube automated nucleotide extraction robot using the PAXgene blood RNA kit (Qiagen 762174) following the manufacturer's protocol. Reverse transcription PCR was done using a random hexamer primer mix and the VILO SuperScript III RT-PCR system (Life Technologies; 11754250) following the manufacturer's protocol. Primers were designed manually using the Primer3Plus software (National Human Genome Research Institute, USA); where possible PCR primers were designed to span exon-exon boundaries. PCR amplification was performed using the Megamix Royal PCR master mix (MicroZone; 2MMR-10). PCR products were visualised by gel electrophoresis (3% agarose) before bi-directional Sanger sequencing. Sanger sequencing was performed using BigDye terminator v3.1 (Applied Biosystems; 4337456) and the Agencourt automated clean-up system (AMPure (Beckman Coulter; A63881), CleanSEQ (Beckman Coulter; A29154)), following the manufacturer's protocol, and sequenced using the ABI 3730 DNA analyser. Sanger sequencing products were visualised using Mutation Surveyor v5.1.2 (SoftGenetics).

Southampton RNA extraction and RT-PCR protocol

Blood was collected in PAXgene Blood RNA tubes (PreAnalytiX, Switzerland). RNA was then extracted from blood samples using PAXgene Blood RNA Kit (PreAnalytiX, Switzerland) and quality control was performed using a 2100 Bioanalyzer instrument (Agilent, UK). RNA extracted from blood samples was converted to cDNA using the High-Capacity cDNA Reverse

Transcription Kit (ThermoFisher Scientific, UK) using random hexamers. Primer pairs were designed manually depending on the genomic locations of variants and ordered from Integrated DNA Technologies (IDT, UK). PCR experiments were performed using GoTaq G2 Polymerase PCR system (Promega, UK) according to the manufacturer's protocol. RT-PCR products were purified by GeneJET PCR Purification Kit (ThermoFisher Scientific, UK) and bidirectional Sanger sequencing was carried out by SourceBioscience (Nottingham, UK). Amplicons were also analysed by agarose gel electrophoresis and imaged by Chemidoc XRS+ (Bio-Rad, USA). Where indicated, amplicons for further analysis were gel-purified by GeneJET Gel Extraction Kit (ThermoFisher Scientific, UK) and cloned into plasmids using a TA cloning kit, with the pCR 2.1 vector (ThermoFisher Scientific, UK). Plasmids carrying inserts were sent to SourceBioscience (Nottingham, UK) for Sanger sequencing.

RNA-seq analysis

QC of read data:

QC was performed on sequencing reads received from Novogene with FastQC (v0.11.3) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and compiled and visualised with MultiQC (v1.5) (https://multiqc.info/).

Alignment with STAR (v2.6.1c)[1]:

The STAR index was created with STAR's genomeGenerate using GRCh38.primary_assembly.genome.fa and gencode.v30.annotation.gtf, both downloaded from GENCODE[2] (https://www.gencodegenes.org/human/) with --sjdbOverhang 149 and all other settings as default. Samples were individually aligned in twopass Basic mode with the following parameters specified, and everything else as default: --outSAMmapqUnique 60, --outFilterType BySJout, --outReadsUnmapped Fastx, --outSAMtype BAM Unsorted.

Samtools[3] (v1.3.2) was used to sort, index and extract regions (using Samtools view) corresponding to the gene harbouring the VOUS.

QC of aligned data:

QC was performed on aligned data using the following components of RSeQC[4] (v2.6.4) (http://rseqc.sourceforge.net/): bam_stat.py, infer_experiment.py, geneBody_coverage.py, junction_annotation.py and junction_saturation.py. Results were compiled and visualised with MultiQC (v1.5) (https://multiqc.info/).

***In silico* splicing predictions**

Equations for sensitivity, specificity, accuracy, positive and negative predictive values:

Sensitivity = true positives / (true positives + false negatives)

Specificity = true negatives / (true negatives + false positives)

Accuracy = (true positives + true negatives) / (all variants)

Positive predictive value = true positives / (true positives + false positives)

Negative predictive value = true negatives / (true negatives + false negatives)

SUPPLEMENTARY FIGURE AND TABLE LEGENDS:

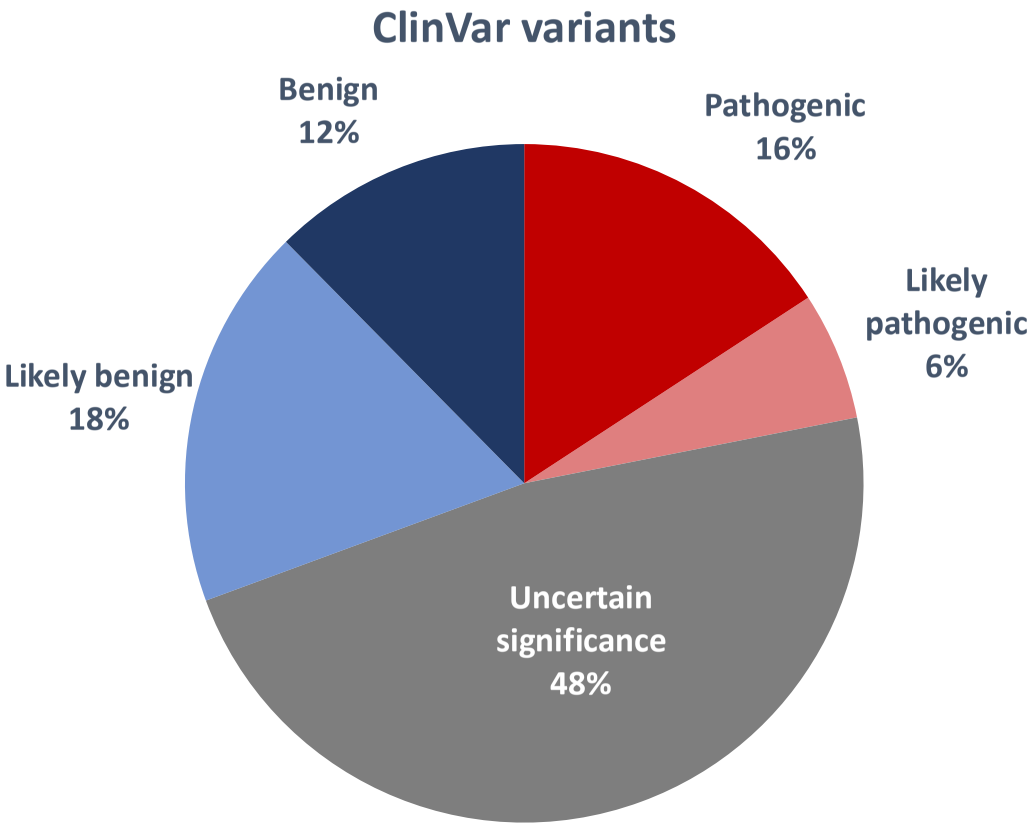
Figure S1. **Variant classifications in ClinVar (as of 13 November 2019).** Proportions and numbers of variants listed in ClinVar under each of the five ACMG classifications.

Table S1. **Tabulated list of variants, splicing effects and bioinformatic predictions.** Variants highlighted in red were found to affect splicing, while blue-highlighted variants were not. SNV position is in reference to the nearest annotated donor (D) or acceptor (A) splice site in the listed transcript, where D-1 is the final nucleotide of an exon and A+1 is the first nucleotide of an exon. Genomic coordinates are listed based on results obtained from the Ensembl Variant Effect Predictor (VEP) downloaded in VCF format.[5] Bioinformatic predictions taken to indicate a splice-altering effect (within applied thresholds) are highlighted in red. For splicing result: IR, intron retention; SE, skipped exon; A5SS, alternative 5´ splice site; A3SS, alternative 3´splice site. For HSF predicted effects: DSB, donor site broken; NDS, new donor site; NAS, new acceptor site; ASB, acceptor site broken; "no result" refers to where HSF made erroneous calls of input sequence elements, leading to inappropriate predictions. For SSF, MaxEntScan, NNSPLICE: NSS, native splice site. SpliceAI values were obtained from a pre-computed score file (v1.3). *These two *MKKS* variants were present *in trans* in a single patient. **This sample contained two separate but closely linked monoallelic *NF1* variants, which have been considered together as a single variant for the purposes of this study.

REFERENCES

1. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635
2. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*. 2019;47(D1):D766-D773. doi:10.1093/nar/gky955
3. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352
4. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*. 2012;28(16):2184-2185. doi:10.1093/bioinformatics/bts356
5. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122. doi:10.1186/s13059-016-0974-4

Supp Figure



Classification	ClinVar variants
Pathogenic	80506
Likely pathogenic	31132
Uncertain significance	242247
Likely benign	92949
Benign	63225

LICENCE TO PUBLISH – OPEN ACCESS

SPRINGER NATURE

Manuscript Number: GIM-D-19-01085

Journal Name: Select a Journal GENETICS IN MEDICINE (the "Journal")

Proposed Title of the Article: Blood RNA analysis can uplift clinical diagnostic rate and resolves variants of (the "Article")

Author(s) [Please list all named authors, continuing on a separate sheet if necessary]: Htoo A. Wai PhD1, Jenny Lord PhD1, Matthew Lyon MSc2, Adam Gunning (the "Author(s)")

Miscellaneous [for office use only]:

Licence options applicable to the Article:

☑ **CC BY**

This licence allows readers to copy, distribute and transmit the Article as long as it is attributed back to the author. Readers are permitted to alter, transform or build upon the Article, and to use the Article for commercial purposes. Please read the full licence for further details at - http://creativecommons.org/licenses/by/4.0/

☐ **CC BY-NC-SA**

This licence allows readers to copy, distribute and transmit the Article as long as it is attributed back to the author. Readers are permitted to alter, transform or build upon the Article as long as the resulting work is then distributed under this or a similar licence. Readers are not permitted to use the Article for commercial purposes. Please read the full licence for further details at - http://creativecommons.org/licenses/by-nc-sa/4.0/

☐ **CC BY-NC-ND**

This licence allows readers to copy, distribute and transmit the Article as long as it is attributed back to the author. Readers may not alter, transform or build upon the Article, or use the article for commercial purposes. Please read the full licence for further details at - http://creativecommons.org/licenses/by-nc-nd/4.0/

will consider publishing this article, including any supplementary information and graphic elements therein (e.g. illustrations, charts, moving images) (the 'Article'), including granting readers rights to use the Article on an open access basis under the terms of the selected Creative Commons licence. Headings are for convenience only.

2 Grant of Rights
Subject to editorial acceptance of the Article, it will be published under the Creative Commons licence shown above. In consideration of the Licensee evaluating the Article for publication, the Author(s) grant the Licensee a non-exclusive, irrevocable and sub-licensable right, unlimited in time and territory, to copy-edit, reproduce, publish, distribute, transmit, make available and store the Article, including abstracts thereof, in all forms of media of expression now known or developed in the future, including pre- and reprints, translations, photographic reproductions and extensions. Furthermore, to enable additional publishing services, such as promotion of the Article, the Author(s) grant the Licensee the right to use the Article (including any graphic elements on a stand-alone basis) in whole or in part in electronic form, such as for display in databases or data networks (e.g. the Internet), or for print or download to stationary or portable devices. This includes interactive and multimedia use as well as posting the Article in full or in part on social media, and the right to alter the Article to the extent necessary for such use. Author(s) grant to Licensee the right to re-license Article metadata without restriction, including but not limited to author name, title, abstract, citation, references, keywords and any additional information as determined by Licensee.

3 Copyright
Ownership of copyright in the Article shall vest in the Author(s). When reproducing the Article or extracts from it, the Author(s) acknowledge and reference first publication in the Journal.

4 Self-Archiving
The rights and licensing terms applicable to the version of the Article as published by the Licensee are set out in sections 2 and 3 above. The following applies to versions of the Article preceding publication by the Licensee and/or copyediting and typesetting by the Licensee. Author(s) are permitted to self-archive a pre-print and an Author's accepted manuscript version of their Article.
a) A pre-print is the Author's version of the Article before peer-review has taken place ("Pre-Print"). Prior to acceptance for publication, Author(s) retain the right to make a Pre-Print of their Article available on any of the following: their own personal, self-maintained website; a legally compliant pre-print server such as but not limited to arXiv and bioRxiv. Once the Article has been published, the Author(s) should update the acknowledgement and provide a link to the definitive version on the publisher's website: "This is a pre-print of an article published in [insert journal title]. The final authenticated version is available online at: https://doi.org/[insert DOI]".
b) An Author's Accepted Manuscript (AAM) is the version accepted for publication in a journal following peer review but prior to copyediting and typesetting. Author(s) retain the right to make an AAM of their Article available on any of the following, provided that they are not made publicly available until after first publication: their own personal, self-maintained website; their employer's internal website; their institutional and/or "This is a post-peer-review, pre-copyedit version of an article published in [insert journal title]. The final authenticated version is available online at: http://dx.doi.org/[insert DOI]".

5 Warranties
The Author(s) warrant and represent that:
a) (i) the Author(s) are the sole copyright owners or have been authorised by any additional copyright owner to grant the rights defined in clause 2, (ii) the Article does not infringe any intellectual property rights (including without limitation copyright, database rights or trade mark rights) or other third party rights and no licence from or payments to a third party are required to publish the Article, (iii) the Article has not been previously published or licensed, (iv) if the Article contains materials from other sources (e.g. illustrations, tables, text quotations), Author(s) have obtained written permissions to the extent necessary from the copyright holder(s), to license to the Licensee the same rights as set out in clause 2 and have cited any such materials correctly;
b) all of the facts contained in the Article are according to the current body of science true and accurate;
c) nothing in the Article is obscene, defamatory, violates any right of privacy or publicity, infringes any other human, personal or other rights of any person or entity or is otherwise unlawful and that informed consent to publish has been obtained for all research participants;
d) nothing in the Article infringes any duty of confidentiality which any of the Author(s) might owe to anyone else or violates any contract, express or implied, of any of the Author(s). All of the institutions in which work recorded in the Article was created or carried out have authorised and approved such research and publication; and
e) the signatory (the Author or the employer) who has signed this agreement has full right, power and authority to enter into this Agreement on behalf of all of the Author(s).

6 Cooperation
a) The Author(s) shall cooperate fully with the Licensee in relation to any legal action that might arise from the publication of the Article, and the Author(s) shall give the Licensee access at reasonable times to any relevant accounts, documents and records within the power or control of the Author(s). The Author(s) agree that the distributing entity is intended to have the benefit of and shall have the right to enforce the terms of this agreement.
b) The Author(s) authorise the Licensee to take such steps as it considers necessary at its own expense in the Author(s)' name and on their behalf if the Licensee believes that a third party is infringing or is likely to infringe copyright in the Article including but not limited to initiating legal proceedings.

7 Author List
After signing, changes of authorship or the order of the authors listed will not be accepted unless formally approved in writing by the Licensee.

Signed for and on behalf of the Author(s): P.P. C.A. DOUGLAS

Print Name: Diana Baralle

Date: 17.01.20

(PLEASE NOTE, ONLY HANDWRITTEN SIGNATURES ARE ACCEPTED)

Address: Faculty of Medicine, University of Southampton, Tremona Road, Southampton, UK.

Corresponding Author Name: Diana Baralle
Manuscript Number: GIM-D-19-01084

Genetics in Medicine

Reporting Checklist* (Please see page 3 for instructions on uploading this file)

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. **Please respond completely to all questions relevant to your manuscript. For sections that are not applicable please fill in NA.** For more information, please read the journal's Guide to Authors.

■ Check here to confirm that the following information is available in the Material & Methods section:

- the **exact sample size (*n*)** for each experimental group/condition, given as a number, not a range;
- a **description of the sample collection** allowing the reader to understand whether the samples represent **technical or biological replicates** (including how many animals, litters, culture, etc.);
- a **statement of how many times the experiment shown was replicated in the laboratory**;
- **definitions of statistical methods and measures**: (For small sample sizes (n<5) descriptive statistics are not appropriate, instead plot individual data points)
    - very common tests, such as *t*-test, simple χ^2 tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;
    - are tests one-sided or two-sided?
    - are there adjustments for multiple comparisons?
    - **statistical test results**, e.g., ***P* values**;
    - definition of **'center values'** as **median or mean**;
    - definition of **error bars as s.d. or s.e.m. or c.i.**

Please ensure that the answers to the following questions are reported both **in the manuscript itself and in the space below.** We encourage you to include a specific subsection in the methods section each for statistics, reagents and animal models. Below, provide the text as it appears in the manuscript as well as the page number.

Statistics and general methods	**Text AND page number from manuscript**
1. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size? (Give text and page #)	NA
For animal studies, include a statement about sample size estimate even if no statistical methods were used.	NA
2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established? (Give text and page #)	Any sample sent from a clinician who thought phenotype fitted.
3. If a method of randomization was used to determine how samples/animals were allocated to experimental groups and processed, describe it. (Give text and page #)	NA
For animal studies, include a statement about randomization even if no randomization was used.	NA

4. If the investigator was blinded to the group allocation during the experiment and/or when assessing the outcome, state the extent of blinding. (Give text and page #)	NA
For animal studies, include a statement about blinding even if no blinding was done.	NA
5. For every figure, are statistical tests justified as appropriate?	NA
Do the data meet the assumptions of the tests (e.g., normal distribution)?	NA
Is there an estimate of variation within each group of data?	NA
Is the variance similar between the groups that are being statistically compared? (Give text and page #)	NA

Reagents	**Text AND page number from manuscript**
6. Report the source of antibodies (vendor and catalog number)	NA
7. Identify the source of cell lines and report if they were recently authenticated (e.g., by STR profiling) and tested for mycoplasma contamination	NA

Animal Models	**Text AND page number from manuscript**
8. Report species, strain, sex and age of animals	NA
9. For experiments involving live vertebrates, include a statement of compliance with ethical regulations and identify the committee(s) approving the experiments.	NA

10. We recommend consulting the ARRIVE guidelines (*PLoS Biol.* **8**(6), e1000412,2010) to ensure that other relevant aspects of animal studies are adequately reported.

Human subjects	**Text AND page number from manuscript**
11. Identify the committee(s) approving the study protocol.	Page 5 - Health Research Authority (IRAS Project ID 49685, REC 11/SC/0269) and by the University of Southampton (ERGO ID 23056)
12. Include a statement confirming that informed consent was obtained from all subjects.	Page 5 - Informed consent for splicing studies was provided for all patients from whom samples were obtained.
13. For publication of patient photos, include a statement confirming that consent to publish was obtained. For more information, please see http://www.icmje.org/recommendations/browse/roles-and-responsibilities/protection-of-research-participants.html.	NA
14. Report the clinical trial registration number (at ClinicalTrials.gov or equivalent).	NA

15. For phase II and III randomized controlled trials, please refer to the CONSORT statement and submit the CONSORT checklist with your submission.

16. For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines.

Data deposition	**Text AND page number from manuscript**
17. Provide accession codes for deposited data. Data deposition in a public repository is recommended for: a. Protein, DNA and RNA sequences b. Macromolecular structures c. Crystallographic data for small molecules d. Microarray data	NA

Deposition is strongly recommended for many other datasets for which structured public repositories exist; more details on our data policy are available in the Guide to Authors. We encourage the provision of other source data in supplementary information or in unstructured repositories such as Figshare and Dryad. We encourage publication of Data Descriptors (see Scientific Data) to maximize data reuse.

18. If computer code was used to generate results that are central to the paper's conclusions, include a statement in the Methods section under "**Code availability"** to indicate whether and how the code can be accessed. Include version information as necessary and any restrictions on availability.	No code was used that was central to the paper's conclusions.

* If an error is shown on the PDF built in Editorial Manager after including this Reporting Checklist take the following steps:
1. Open the original filled GIM-reporting-checklist.pdf with Mac Preview App
2. Goto Tray -> Export as PDF
3. Save as: at your desk (local computer) with a different file name
4. Close the app
5. In Editorial Manager: Edit the Submission which displayed Error or Incomplete Status and remove the file, then upload the new one, and again select to build the merged PDF

Supp table

Click here to access/download

Large Excel File

Table_S1_final.xlsx

COLOR ARTWORK FORM

This form must be completed for all papers. We will be unable to process your paper through to production until we receive instructions concerning color files. Please upload this form with your revised files.

JOURNAL: ***Genetics in Medicine***

ARTICLE TITLE BLOOD RNA ANALYSIS CAN UPLIFT CLINICAL DIAGNOSTIC RATE AND RESOLVE VARIANTS OF UNCERTAIN SIGNIFICANCE

MANUSCRIPT NUMBER GIM-D-19-01085

CORRESPONDING AUTHOR NAME DIANA BARALLE

Genetics in Medicine has charges for figures printed in color: $500 for the first color figure and $250 for each subsequent figure. Current ACMG members (excluding Student Members) who are first or senior/corresponding authors are exempt,* as are authors who have opted for Open Access.*

Please note that figures can appear online in color in the HTML version of your manuscript, and in black and white in the PDF/print version of the manuscript. Color figures will also be at the discretion of the editorial office.

Please check:

☐ Yes, my manuscript contains material that must be printed in color. I agree to pay the color charges in full and hereby authorize Nature Publishing Group to invoice me for the cost of reproducing color artwork in print.

☑ Yes, my manuscript contains material that should be in color in the online HTML version, but in black and white in the PDF/print version of the manuscript. No charges incurred.

☐ Yes, my manuscript contains material that must be printed in color, but I have chosen Open Access for my article or am an ACMG member, so am exempt.

☐ No, my manuscript does not contain material that must be printed in color.

Which figures should be printed in color? (e.g. Figures 1a, 2, 3b)

_____

College Member's Name: _____

Membership will be verified by the ACMG, and false claims will be liable for the full color charge rate. Signed completion of this form constitutes a full and total acceptance of the terms listed. College membership has no bearing on the review process at *Genetics in Medicine*, and papers from members and nonmembers alike will be given equal consideration.

Color figures will be set close to the citation and in the best possible position.

20% VAT will be added to the total charge amount upon invoicing. This applies to all EU authors who do not provide a valid VAT number upon returning this form. Customers outside of the EU will not be charged VAT but local taxes will be added where applicable.

Signature: pp AM (A. DOUGLAS) VAT no. (if applicable): _____

Print Name: DIANA BARALLE Date: 17/01/2020

nature publishing group npg

Updated 8/16/16