**Short amplicon Reverse Transcription-Polymerase Chain Reaction (RT-PCR) detects aberrant splicing in genes with low expression in blood missed by Ribonucleic Acid sequencing (RNA-seq) analysis for clinical diagnosis**

Htoo A. Wai1, Matthew Constable1, Cosima Drewes1, Ian C. Davies1, Eliska Svobodova3, Esther Dempsey4, Anand Saggar4, Tessa Homfray4, Sahar Mansour4, Sofia Douzgou5,6, Kate Barr7, Stephanie Scott2, David Hunt1,2, Andrew G. L. Douglas1,2, Diana Baralle1,2

Corresponding author : Professor Diana Baralle, contact email : [d.baralle@soton.ac.uk](mailto:d.baralle@soton.ac.uk)

Funding information : National Institute for Health Research (NIHR) (RP-2016-07-011)

1. Human Development and Health, Faculty of Medicine, Southampton General Hospital,

University of Southampton, UK,

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

3. Department of Experimental Biology, Faculty of Science, Masaryk university, Brno 611 37, Czech Republic

4. Department of Clinical Genetics, St George's University of London, London, UK

5. Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University NHS Foundation Trust, Health Innovation Manchester, Manchester, UK

6. Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

7. Bristol Regional Clinical Genetics Service, St Michael's Hospital, Bristol, UK

HAW developed, designed,conducted the experiments, wrote and revised the manuscript . MC,CD,ICD and ES contributed to the experiments. ED,TH, AS,SM,SD,KB,SS and DH provided the patient samples. AGLD and DB supervised the study and edited, revised the manuscript.

**Abstract**

Use of blood RNA-seq as a splicing analysis tool for clinical interpretation of variants of uncertain significance (VUSs) found via whole genome and exome sequencing can be difficult for genes that have low expression in blood due to insufficient read count coverage aligned to specific genes of interest. Here, we present short amplicon RT-PCR for detection of genes with low blood expression. Short amplicon RT-PCR, is designed to span three exons where an exon harbouring a variant is flanked by one upstream and one downstream exon. We tested short amplicon RT-PCRs for genes which have median transcripts per million (TPM) values less than one according to the Genotype-Tissue Expression (GTEx) database. Median TPM values of genes analysed in this study are *SYN1*=0.8549, *COL1A1*=0.6275, *TCF4*=0.4009, *DSP=*0.2894, *TTN*=0.2851, *COL5A2*=0.1036, *TERT*=0.04452, *NTRK2*=0.0344, *ABCA4*=0.00744, *PRPH*= 0 and *WT1*=0. All these genes show insufficient exon-spanning read coverage in our RNA-seq data to allow splicing analysis. We successfully detected all genes tested except *PRPH* and *WT1*. Aberrant splicing was detected in *SYN1, TCF4, NTRK2, TTN* and *TERT* VUSs. Therefore, our results show short amplicon RT-PCR is a useful alternative for the analysis of splicing events in genes with low TPM in blood RNA for clinical diagnostics.

Key words: Aberrant splicing, blood RNA, RNA-seq, RT-PCR, VUS

**Introduction**

RNA splicing analysis is increasingly being used to aid rare disease diagnosis through detection of splicing variants (Murdock et al., 2021; Rowlands et al., 2020; H. A. Wai et al., 2020). This can be achieved either via targeted RT-PCR when testing the effects of known variants of uncertin significance (VUSs) or increasingly also via transcriptome-wide RNA sequencing (RNA-seq) (H. A. Wai et al., 2020). The limits of sensitivity of transcriptome-wide RNA-seq are yet to be fully defined or standardised but they depend upon tissue-specific gene expression values, levels of transcript degradation, library preparation methods, sequencing parameters and subsequent bioinformatic processing such as read filtering and alignment.

Transcripts per million (TPM) values can be calculated from RNA-seq data and can be used to provide an estimate of whether sufficient coverage is likely to be achieved for specific genes of interest (Wagner et al., 2012). Publicly available expression datasets such as the TPM values from the Genotype-Tissue Expression (GTEx) project are frequently referred to when trying to determine whether or not splicing can be assayed via whole blood RNA sampling (Lonsdale et al., 2013). However, significant experimental protocol differences between prior published data and ongoing current RNA analyses may lead to errors in predicting the detectability of abnormal splicing if such datasets are relied upon in isolation.

Conventional RT-PCR is known to be useful for the analysis of splicing as an alternative cost-effective technique to RNA-seq (Macken et al., 2021; H. A. Wai et al., 2020). Here, we sought to examine whether genes with low TPM values on RNA-seq can still be assayed for potentially abnormal splicing through the use of RT-PCR. We confirmed that exonic and junction-spanning read coverages are poor in genes with low TPM values. We then used RT-PCR to test VUSs for their effect on splicing in low TPM genes which cannot be readily analysed using RNA-seq, designing short-amplicon PCR assays to see if they could outperform blood RNA transcriptome analysis for such genes and produce clinically actionable results.

**Materials and Methods**

**Patient recruitment procedure and RNA extraction**

A total of 13 samples with low TPM genes from patients with VUSs were identified through the Splicing and Disease research study at the University of Southampton, UK, ethics approved by the Health Research Authority (IRAS Project ID 49685, REC 11/SC/0269) and by the University of Southampton (ERGO ID23056). Informed consent for splicing studies was provided for all patients from whom samples were obtained. Patient blood samples were collected using PAXgene Blood RNA tubes and intracellular RNA extraction from whole blood was performed with the PAXgene Blood RNA Kit (PreAnalytix, Switzerland). All the patient RNA samples quality and concentration were checked with bioanalyser before performing RT-PCR and RNA-seq.

**RNA sequencing and data analysis**

Whole transcriptome sequencing with ribosomal RNA depletion and stranded library preparation was carried out via Novogene (Hong Kong) using NovoSeq 600 PE150. A minimum coverage depth of 70M 150 base pair paired-end reads were produced for each sample. Distribution of sequencing quality, error rate distribution, A/T/G/C distribution, composition of raw data and data quality summary were mentioned in the supplementary data (Supp. Data D1). Data analysis was performed on the IRIDIS 4 high performance computer cluster, University of Southampton. FASTQ reads were aligned to the reference human genome version 38 (GRCh38) using the STAR alignment tool (2.5.2b version) (Dobin et al., 2013) with read level filters : reads corresponding to a mapping quality of 255 and maximum mismatch of 6. These filtering parameters were used to match those employed by GTEx. Filtered aligned BAM files were visualised and sashimi plots were generated using IGV viewer (Ttir et al., 2013).

**RT-PCR and PCR**

All the experiments were conducted in the laboratory of Human Development and Health, Southampton General Hospital, University of Southampton, UK, SO16 6YD. cDNA was synthesised using the High-Capacity cDNA Reverse Transcription Kit (ThermoFisher Scientific, USA) using random hexamers. Primer pairs (Supp Data D2) were designed using online Primer3web tool (primer3.ut.ee), selecting binding sites in exons adjacent to variants and standard desalted synthetic single stranded DNA oligo primers were ordered from Integrated DNA Technologies (IDT, USA). PCR experiments were performed using the GoTaq G2 Polymerase PCR system (Promega, USA) and GoTaq Hot Start Polymerase (Promega, USA) according to the manufacturer's protocol. PCR cycles were set up at 95 degree Celsius for 5 minutes followed by 35 and 40 repetitive cycles of 95 degree Celsius for 30 seconds, 60 degree Celsius or the annealing temperature of individual primer pairs for 30 seconds and 72 degree Celsius for 1 minute per kilo base pairs before ending the cycle with 72 degree Celsius for 10 minutes and cooling down at 4 degree Celsius. RT-PCR products were purified by GeneJET PCR Purification Kit (ThermoFisher Scientific, USA) and bidirectional Sanger sequencing was carried out by SourceBioscience (Nottingham, UK). Where indicated, amplicons for further analysis were gel-purified by GeneJET Gel Extraction Kit (ThermoFisher Scientific, USA) and sequenced at SourceBioscience (Nottingham, UK) for Sanger sequencing. All the PCR experiments were repeated at least twice for reproducibility and all expected positive amplicons were confirmed by Sanger sequencing.

**Results**

**RNA-seq sensitivity is reduced for low TPM genes**

TPM is a normalised RNA abundance measurement of high-throughput RNA sequencing data. The values are calculated using the number of reads aligned to a gene of interest and they come from their respective genes or transcripts for every one million RNA molecules in the RNA-seq sample (Wanger *et al*.,2013). It is a useful indicator of gene expression level and for predicting the detectability of splicing events using reads spanning exon-exon junctions. Exon-spanning reads are crucial for splicing analysis using RNA-seq and these are abundant in genes where the median TPM value is more than one (Figure 1). Aberrant splicing events can also be readily visualised in genes where the TPM value is one or more in RNA-seq of blood RNA. However, exonic and exon-spanning reads of genes with median TPM below 1 show lower or no read coverage (Figure 1). Therefore, aberrant splicing events can be missed due to insufficient read coverage where TPM values are lower than 1 and cannot be easily analysed using the RNA sequencing.

**Short amplicon RT-PCR is more sensitive than standard RT-PCR in detecting genes with low TPM in blood.**

The primers we designed for conventional RT-PCR for the detection of aberrant splicing are located so as to span at least seven exons, where the exon harbouring the variant is at the centre of the amplicon. In this way, aberrant splicing events such as one or two exon skipping affecting upstream or downstream exons with respect to the variant of interest, can be detected by gel electrophoresis. This design of PCR for aberrant splicing detection is sensitive for genes which have high TPM values in blood (Figure 2A). However, it fails to produce any amplicons for genes which have lower TPM values (less than one) (Figure 2B-G). On the other hand, a short amplicon RT-PCR design, in which primers span only three exons, is sensitive enough to amplify the targeted cDNA (Figure 2B-G) when performed at 35 cycles using GoTaq G2 DNA polymerase. As only three exons are spanned in short amplicon PCR, only single-exon-skipping aberrant splicing events can be detected and two-exon-skipping events may be missed.

Both short and long amplicons were further analysed using different PCR conditions such as amplification cycles and polymerases. Two different amplification cycles, 35 and 40 cyles and two different Taq DNA polymerase, GoTaq G2 and GoTaq Hot Start polymerases were tested for the three genes with low TPM values. Only short amplicons were detected in all different PCR conditions whereas non-specific amplicons or no amplicons were detected in long amplicon PCR reactions, confirming that short amplicon PCR has higher sensitivity than long one (Supp Figure S1).

**Short amplicon RT-PCR detects normal and aberrant transcripts of genes with low TPM values**

Four different categories of low TPM genes were analysed using short amplicon RT-PCRs. These categories were based on the GTEx database median TPM values: 1 to 0.1 (*COL1A1, COL5A2*, *DSP, SYN1*, *TCF4*, *TTN*), 0.1 to 0.01 (*NTRK2*, *TERT)*, 0.01 to 0.001 (*ABCA4*), 0 (*PRPH, WT1*). Amplicons were successfully detected in genes with median TPM ranges from 1 to 0.001 using short amplicon RT-PCR (Figure 3). Aberrant splicing events were detected in *SYN1* c.838-2 A>G (median TMP=0.8549) (Figure 3A), *TCF4* c.550-3C>G(median TPM=0.4009), *NTRK2* c.287+3G>C (median TPM=0.03444) (Figure 3D), *TERT* c.3295+5G>T (median TPM=0.04452), *TERT* c.5137+3A>G (median TPM=0.04452) (Figure 3E), *TTN* c.49346-1G>A (median TPM=0.2851) and *TTN* c.63793G>A (median TPM=0.2851). In comparison, sashimi plots of RNA-seq read coverage data across splice junction for low TPM gene regions show poor or no exonic or exon-spanning reads covering these areas. The results of all VUSs tested are summarised in table 1. Amino acid changes caused by aberrant splicing are also shown in table 1.

**Discussion**

RNA-seq is being integrated into clinical diagnostic services as a tool for the identification of pathogenic sequence variants not identified by standard exome or genome data filtering (Douglas & Baralle, 2021). However, using RNA-seq to detect aberrant splicing patterns in genes not well expressed in the tissue source can be difficult due to low read coverage of exons and splice junctions. In our RNA-seq analysis, blood RNA was used to generate 70 million, 150bp paired-end reads per sample. However, 70 million read coverage is not enough to generate exonic and junction spanning reads for low TPM genes. A possible solution to overcome low read counts is to increase the total read counts generated by RNA-seq. However, the cost of sequencing increases prohibitively for large-scale testing in a healthcare system. In addition, the generation of more read counts per sample may still not guarantee successful splicing analysis in low TPM genes, due to individual differences in gene expression (Lonsdale et al., 2013). A solution would be to source the relevant cells or tissues that have high gene expression levels of the genes of interest. However, alternative tissue sources are often not feasible or pleasant for the patient, highlighting the importance of a reliable source of RNA for splicing analysis.

Conventional RT-PCR-based splicing analysis has been traditionally used for the assessment of VUSs and their effect on splicing (H. Wai et al., 2019; H. A. Wai et al., 2020). However, even this technique can fail in genes where there is low TPM. In contrast, we found that our redefined short amplicon RT-PCR, spanning only three exons, is sensitive enough to detect low TPM genes, providing crucial clinically useful information. However, one drawback is that unlike long amplicon PCR, short amplicon RT-PCR may miss multi exon skipping events.

Although aberrant splicing is associated with disease phenotype, the phenotypic severity may depend on the type of aberrant splicing. For example, in-frame aberrant splicing may have milder phenotypic effects than a frameshift or termination codon induced by aberrant splicing, depending on the relevant gene's pathogenetic mechanisms (Lord & Baralle, 2021). However, further studies are needed on a gene-by-gene basis in order to determine what usage level of alternative splicing is pathogenically significant.

RNA-seq read coverage for low TPM genes becomes sparse (TPM value 1 to 0.01) and completely disappears for genes in which TPM values are lower than 0.01. We found that short amplicon RT-PCR can detect genes beyond this threshold with TPM ranging from 1 to 0.001. Minigene assays are an alternative method for analysing variants not detected by either RNA-seq or RT-PCR but these are time-consuming to set up and are not easily translated into a high-throughput diagnostic service (Pagani & Baralle, 2009).In contrast, this short RT-PCR method can be applied in a more high-throughput manner for aberrant splicing detection in the majority of genes of interest.

In conclusion, our findings hightlight the limits of diagnostic RNA-seq for low TPM genes. Redefined short amplicon RT-PCR is a method that is cheaper, simpler and quicker. Our findings show that short RT-PCR amplicons compensate for the shortfalls of RNA sequencing in assessing splicing in low TPM genes, down to 0.001TPM but not where TPM is zero, as a useful adjunct to large-scale transcriptomics for genomic diagnostic services.

**Data Availability Statement**

VUS information were submitted to ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) with accession numbers from SCV002106369 to SCV002106379.

**Figure 1. Sashimi plots showing junction-spanning read comparison between genes with median TPM value more than one and genes with median TPM value less than one in data from three randomly selected blood RNA-seq samples.** Genes with median TPM value more than one have more exon spanning reads whereas genes with median TPM value less than one have less or no exon spanning reads.

**Figure 2. Short amplicon RT-PCR has higher sensitivity than long amplicon RT-PCR.** (A) Long amplicon primers are designed to span at least 7 exons whereas short amplion primers are designed to span only three exons in RT-PCR to detect low TPM genes. Red exons represent VUS harbouring exons and the green are the primer-targeting exons (B-G)Short amplicon and long amplicon PCRs are compared using three controls and one no template control (NTC) in *SYN1, COL1A1, COL5A2, TERT, NTRK2* and *ABCA4.* Expected short and and long amplicon sizes are indicated with green and blue arrows respectively. Amplified genomic DNA fragments are indicated with red arrows. All PCR reactions were performed using GoTaq G2 DNA Polymerase (Promega) at 35 amplification cycles. .

**Figure 3. Gel electrophoresis and Sanger sequencing results of short amplicon RT-PCR for low TPM genes.** (A to C) Examples of three genes with TPM ranging from 1 to 0.1, *SYN1* (TPM=0.8459), *COL1A1* (TPM=0.6275) and *COL5A2* (TPM=0.1036). (D and E) Examples of two genes with TPM ranging from 0.1 to 0.01, *TERT* (TPM=0.04452) and *NTRK2* (TPM= 0.0334) (F) Example of a gene with TPM ranging from 0.01 to 0.001, *ABCA4* (TPM=0.00744). Green arrows and arches indicate the normal splicing amplicons and events whereas the red arrow and arches hightlights the aberrant splicing amplicons and events. Gels were spliced to form the images shown in Figure 3 and the editors have reviewed the respective raw data.

**Table 1. VUS information of low TPM genes and their aberrant splicing results**

**References**

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Douglas, A. G. L., & Baralle, D. (2021). Translating RNA splicing analysis into diagnosis and therapy. In *OBM Genetics* (Vol. 5, Issue 1). LIDSEN Publishing Inc. https://doi.org/10.21926/obm.genet.2101125

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. In *Nature Genetics* (Vol. 45, Issue 6, pp. 580–585). Nature Publishing Group. https://doi.org/10.1038/ng.2653

Lord, J., & Baralle, D. (2021). Splicing in the Diagnosis of Rare Disease: Advances and Challenges. *Frontiers in Genetics*, *12*, 1146. https://doi.org/10.3389/FGENE.2021.689892/BIBTEX

Macken, W. L., Godwin, A., Wheway, G., Stals, K., Nazlamova, L., Ellard, S., Alfares, A., Aloraini, T., AlSubaie, L., Alfadhel, M., Alajaji, S., Wai, H. A., Self, J., Douglas, A. G. L., Kao, A. P., Guille, M., & Baralle, D. (2021). Biallelic variants in COPB1 cause a novel, severe intellectual disability syndrome with cataracts and variable microcephaly. *Genome Medicine*, *13*(1), 1–19. https://doi.org/10.1186/s13073-021-00850-w

Murdock, D. R., Dai, H., Burrage, L. C., Rosenfeld, J. A., Ketkar, S., Müller, M. F., Yépez, V. A., Gagneur, J., Liu, P., Chen, S., Jain, M., Zapata, G., Bacino, C. A., Chao, H. T., Moretti, P., Craigen, W. J., Hanchard, N. A., & Lee, B. (2021). Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. *Journal of Clinical Investigation*, *131*(1). https://doi.org/10.1172/JCI141500

Pagani, F., & Baralle, F. E. (2009). Analysis of Human Splicing Defects Using Hybrid Minigenes. In *Molecular Diagnostics: Second Edition* (pp. 155–169). Elsevier Ltd. https://doi.org/10.1016/B978-0-12-374537-8.00011-0

Rowlands, C., Thomas, H., Lord, J., Wai, H., Arno, G., Beaman, G., Sergouniotis, P., Gomes-Silva, B., Campbell, C., Gossan, N., Hardcastle, C., Webb, K., O’Callaghan, C., Hirst, R., Ramsden, S., Jones, E., Clayton-Smith, J., Webster, A., Douglas, A., … Ellingford, J. (2020). Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. *Authorea Preprints*. https://doi.org/10.22541/AU.160157595.59675486

Ttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

Wagner, G. P., Kin, K., & Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, *131*(4), 281–285. https://doi.org/10.1007/s12064-012-0162-3

Wai, H. A., Lord, J., Lyon, M., Gunning, A., Kelly, H., Cibin, P., Seaby, E. G., Spiers-Fitzgerald, K., Lye, J., Ellard, S., Thomas, N. S., Bunyan, D. J., Douglas, A. G. L., Baralle, D., Naik, S., Ragge, N., Cox, H., Morton, J., O’Driscoll, M., … Archer, L. (2020). Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genetics in Medicine*, *22*(6), 1005–1014. https://doi.org/10.1038/s41436-020-0766-9

Wai, H., Douglas, A. G. L., & Baralle, D. (2019). RNA splicing analysis in genomic medicine. *International Journal of Biochemistry and Cell Biology*, *108*, 61–71. https://doi.org/10.1016/j.biocel.2018.12.009

Acknowledgements

This study was funded by the National Institute for Health Research (NIHR) (RP-2016-07-011 research professorship awarded to DB). The authors thank all the patients recruited to this research and the CRN Musketeers’ Memorandum. The authors acknowledge the IRIDIS4 High Capacity Performance Computer and the supporting team at the University of Southampton. The authors also thank the technical teams of Duthie and IDS buildings, University of Southampton, for their support in day-to-day lab work.

Disclosure

The authors declare no conflict of interest.