Human Genetics (2023) 142:495–506
https://doi.org/10.1007/s00439-023-02539-z

ORIGINAL INVESTIGATION

VIsoQLR: an interactive tool for the detection, quantification and fine-tuning of isoforms in selected genes using long-read sequencing

Gonzalo Núñez-Moreno[1,2,3] · Alejandra Tamayo[1,3,4] · Carolina Ruiz-Sánchez[1] · Marta Cortón[1,3] · Pablo Mínguez[1,2,3]

Received: 2 December 2022 / Accepted: 23 February 2023 / Published online: 7 March 2023

Abstract

DNA variants altering the pre-mRNA splicing process represent an underestimated cause of human genetic diseases. Their association with disease traits should be confirmed using functional assays from patient cell lines or alternative models to detect aberrant mRNAs. Long-read sequencing is a suitable technique to identify and quantify mRNA isoforms. Available isoform detection and/or quantification tools are generally designed for the whole transcriptome analysis. However experiments focusing on genes of interest need more precise data fine-tuning and visualization tools.

Here we describe VIsoQLR, an interactive analyzer, viewer and editor for the semi-automated identification and quantification of known and novel isoforms using long-read sequencing data. VIsoQLR is tailored to thoroughly analyze mRNA expression in splicing assays of selected genes. Our tool takes sequences aligned to a reference, and for each gene, it defines consensus splice sites and quantifies isoforms. VIsoQLR introduces features to edit the splice sites through dynamic and interactive graphics and tables, allowing accurate manual curation. Known isoforms detected by other methods can also be imported as references for comparison. A benchmark against two other popular transcriptome-based tools shows VIsoQLR accurate performance on both detection and quantification of isoforms. Here, we present VIsoQLR principles and features and its applicability in a case study example using nanopore-based long-read sequencing. VIsoQLR is available at https://github.com/TBLabFJD/VIsoQLR.

Introduction

Spliceogenic DNA variants are an underestimated cause of genetic diseases (Lord and Baralle 2021). They disrupt canonical donor and acceptor splicing sites or introduce cryptic exonic or intronic sites leading to aberrant mRNA maturation processes. Detection of genomic variation modifying splicing patterns and their association with disease traits is a challenging task that requires in vivo or in vitro functional studies. Although global transcriptomics approaches and methods are available (Mehmood et al. 2020), monogenic diseases need to focus on a single locus (or few loci) to assess the relevance of potentially disease-causing variant effects. In these cases, some accurate and cost-effective procedures to detect altered expression in a targeted region are splicing RT-PCR assays applied to patient samples (Anna and Monika 2018) or, alternatively, minigenes-based exon trapping assays if the damaged primary tissue is not accessible (Cooper 2005). RNA sequencing approaches to identify splicing defects are now being applied as a complementary analysis for the genetic study

✉ Marta Cortón
mcorton@quironsalud.es

✉ Pablo Mínguez
pablo.minguez@quironsalud.es

Gonzalo Núñez-Moreno
gonzalo.nunezm@quironsalud.es

1 Department of Genetics and Genomics, Health Research Institute-Fundación Jiménez Díaz University Hospital, Universidad Autónoma de Madrid (IIS-FJD, UAM), Madrid, Spain

2 Bioinformatics Unit, Health Research Institute-Fundación Jiménez Díaz University Hospital, Universidad Autónoma de Madrid (IIS-FJD, UAM), Madrid, Spain

3 Center for Biomedical Network Research On Rare Diseases (CIBERER), Instituto de Salud Carlos III, Madrid, Spain

4 Department of Surgery, Medical and Social Sciences, Faculty of Medicine and Health Sciences, Science and Technology Campus, University of Alcalá, 28871 Alcalá de Henares, Spain

of Mendelian single-gene disorders, such as familial breast/ovarian cancer (Whiley et al. 2014; Fraile-Bethencourt et al. 2019), cystic fibrosis (Felício et al. 2016), Duchenne muscular dystrophy (Gonorazky et al. 2019; Okubo et al. 2022), neurofibromatosis type 1 (Evans et al. 2016; Koster et al. 2021), spinal muscular atrophy (Wadman et al. 2020) and Stargardt disease (Sangermano et al. 2018). Traditionally used analytical techniques have significant drawbacks in addressing the full spectrum of splicing events. First, Sanger sequencing or capillary electrophoresis assays are time-consuming techniques with laborious protocols that cannot assess the relative level of transcript isoforms. Another drawback is that it is necessary to infer the exon organization from the sizes obtained in the electropherogram peaks, since the sequence is not available. This makes it more difficult to report novel isoforms which are not predicted by splicing predictors. On the other hand, short-read sequencing of RT-PCR products can be used to discover splicing sites, but cannot determine the exon organization of the full-length isoforms. The recent advent of long-read sequencing (LRS) appears as an alternative to characterize and quantify the isoform spectrum in splicing assays (Amarasinghe et al. 2020; Helman et al. 2021; Dai et al. 2022; Jurkute et al. 2022). This latter approach has the potential to cover entire transcripts in a single read, allowing the study of splice sites and complete exon organization of all sequenced transcripts.

Several bioinformatics tools are available for the quantification and analysis of transcript isoforms. Most of them define isoforms as clusters of reads. In addition, they may require: (1) a reference sequence(s), such as StringTie2 (Kovaka et al. 2019); (2) the reference and an annotation file with exon coordinates, as is the case of Mandalorian (Byrne et al. 2017), TALON (Wyman et al. 2019), FLAIR (Tang et al. 2020) and LIQA (Hu et al. 2021); (3) a reference and short-read sequencing data to infer splice sites, such as FLAIR and IDP (Fu et al. 2018); or 4) none of the above, such as Oxford Nanopore algorithm (https://github.com/epi2me-labs/wf-transcriptomes). Pacific Biosciences (PacBio) sequencing data have its own collection of available tools, including PacBio IsoSeq3 (Gonzalez-Garay 2016), IsoCon (Sahlin et al. 2018), SQANTI (Tardaguila et al. 2018), TAPIS (Abdel-Ghany et al. 2016), and SpliceHunter (Kuang and Canzar 2018). To the best of our knowledge, there is no interactive tool that allows a close inspection and fine-tuning analysis of one gene at a time on data from long-reads RNA sequencing. This feature is needed to be applied in the study of variants affecting splicing sites causing monogenic diseases.

Herein, we present VIsoQLR, an interactive analyzer, viewer and editor for identifying and quantifying isoforms obtained from LRS data without prior knowledge of splice sites. VIsoQLR is designed to characterize aberrant mRNAs detected by functional assays targeting a single *locus* linked to specific phenotypes.

Materials and methods

Software implementation and availability

VIsoQLR is implemented in R (R Core Team 2020) using the Shiny package (Chang et al. 2021) to build the interactive local web-based application. Figures displayed in the app are rendered using the plotly package (Sievert 2020). VIsoQLR code, installation, and user manual are available at https://github.com/TBLabFJD/VIsoQLR. A docker image can be downloaded from https://hub.docker.com/r/tblabfjd/visoqlr.

SIRV isoform detection and quantification

LRS data were downloaded from a public data set provided by Pacific Biosciences at https://github.com/PacificBiosciences/DevNet/wiki/Sequel-II-System-Data-Release:-Universal-Human-Reference-(UHR)-Iso-Seq. This consists of RNASeq data using the Iso-Seq™ method of the Universal Human Reference RNA (Agilent) plus SIRV Isoform Mix E0 (Lexogen). SIRV (Spike-In RNA Variants) Isoform Mix E0 contains synthetic transcripts that mimic the expression of 69 transcripts derived from seven human model genes (Paul et al. 2016). All transcripts are present in equimolar concentrations. Full-length reads in the provided BAM file were transformed to FASTQ and subsequently mapped using GMAP aligner (Wu and Watanabe 2005). GMAP parameters were: '-n1' to avoid chimeric alignments, '–cross-species' for a more sensitive search for canonical splicing, and '-f samse' to generate a SAM file, which was transformed into a BAM file, sorted, and indexed. We ran VIsoQLR with default parameters and with no further edition of the detected consensus exon coordinates. Other two methods were selected to assess VIsoQLR performance, StringTie2, having the same requirements of our tool, and FLAIR, as representative of algorithms that need previous splice sites definition. StringTie2 (Kovaka et al. 2019) was run using long reads mode (using '-L' parameter) and setting '-f 0' to output all isoforms. FLAIR (Tang et al. 2020) was run using default parameters providing the GTF with exon coordinates of all transcripts. Benchmark was also performed using minimap2 (Li 2018) as mapper with the options "-ax splice:hq -uf –secondary = no".

Isoforms detected by the three algorithms were compared to the transcript coordinates provided by Lexogen. Transcripts intersecting more than 99% of the bases reciprocally were considered true positives. The SIRV502 transcript was removed from the analysis following the recommendations of the batch amendment (https://www.lexog

en.com/wp-content/uploads/2021/06/025UI079V0110_SIRV-Set-1_Amendment_Batch-No.-216652830_2021-06-08.pdf). The abundance of all isoforms was calculated relative to each SIRV and compared between the three methods and the gold standard using cosine similarity.

Scripts used for data preprocessing, isoform calling and isoform comparison of VIsoQLR, FLAIR and StringTie2 are available at https://github.com/TBLabFJD/VIsoQLR_benchmark.

PAX6 and *TP53* analysis from complete RNASeq data

Human reference unspliced sequences from all protein-coding genes present in the Matched Annotation from NCBI and EMVL-EBI (MANE) resource, were retrieved from BioMart (Ensembl Genes 108, Human genes (GRCh38.p13)). The LRS data used in the previous section were mapped using GMAP with the same parameters as above to this reference resulting in 6,226,430 mapped reads. *PAX6* (ENSG00000007372) and *TP53* (ENSG00000141510) were analyzed using VIsoQLR setting the "read threshold" for the automatic detection of exon coordinates at 3%. In the case of *PAX6*, exon coordinates at the beginning of the first exon were merged. Same was done for the stop of the last exon.

Minigene splicing assay and RT-PCR

We cloned the region from exon 5 to 7 of the gene *PAX6* (RefSeq transcript NM_000280.4), including ~ 200 bp intronic sequence on each side into the exon trapping expression pSPL3 vector. This construction was transfected in HEK-293 T cells. Total RNA was isolated and retrotranscribed to cDNA using random hexamers, as previously described (Tarilonte et al. 2022). The obtained cDNA was further amplified using a primer pair that hybridized with the exons SD6 and SA2 from the expression vector.

Case study long-read sequencing and analysis

The amplified *PAX6* cDNA was sequenced on a MinION Mk1B device (Oxford Nanopore Technologies, ONT, UK) using a SpotOn Flow Cell (R9.4.1). Library preparation was carried out using the SQK-LSK109 sequencing kit (ONT) following the recommended protocol "Native barcoding amplicons" (version NBA_9093_v109_revF_12Nov2019). The library was sequenced until 5000 reads were obtained. Base-calling was performed using Guppy v5.0.16.

Reads were mapped using the GMAP aligner with the same parameters as above. In the case study described here, VIsoQLR was applied to set the "read threshold" for the automatic detection of exon coordinates at 3% after an initial exploration with default parameters without editing any of the detected coordinates, and selecting the "Select only complete PCR sequences" option. Data were also analyzed with StringTie2 to detect splicing isoforms. StringTie2 was run using long reads mode (using '-L' parameter). The detected and quantified isoforms were retrieved in a GTF file and visualized using VIsoQLR.

Semi-quantitative capillary electrophoresis

PCR was performed as described above, but the reverse primer targeting SA2 was HEX-labelled. Fluorescent amplicons were run together with ROX1000 size standard (Asuragen, USA) under denaturing conditions in an ABI3130xl Genetic Analyzer (Thermo Fisher Scientific, USA). The results were analyzed with GeneMapper software (Thermo Fisher Scientific, USA).

Results

Software scope and description

VIsoQLR has been developed to provide users with a graphical and interactive tool for isoform identification and quantification using LRS data generated by Nanopore or PacBio technologies. It gives a single-locus at a time analysis as the data exploration focuses on the graphical display of splice site distribution across the gene. VIsoQLR automatically detects splice site coordinates that can be fine-tuned according to the user's expertise.

Figure 1 shows the workflow for isoform analysis using VIsoQLR. First, raw reads need to be aligned to a reference sequence. VIsoQLR has built-in options for mapping reads using GMAP (Wu and Watanabe 2005) or minimap2 (Li 2018) aligners. Next, mapped reads are uploaded, and consensus exon coordinates (CECs) are defined based on the frequency of the reads' exon coordinates (start and end positions). Start and end positions are treated independently. Some facilities for selecting final CECs are provided: (1) a frequency filter, in which start and end positions above a configurable frequency are selected as candidate CECs (by default 2%); (2) a position window, in which the user can define a window, where other non-candidate CEC (below range in the frequency filter) are merged (by default 5 nucleotides (nt) on both sides of each candidate CEC); and 3) a filter in which candidate CECs are merged into the most frequent one if they are closer than a given distance (by default 3 bases). In addition, VIsoQLR allows the user to change any parameter that defines CECs automatically and to add, delete and edit them manually. Known splice sites can also be uploaded in a file. Thus, once the consensus start and end positions are selected, the exons are defined accordingly, and reads

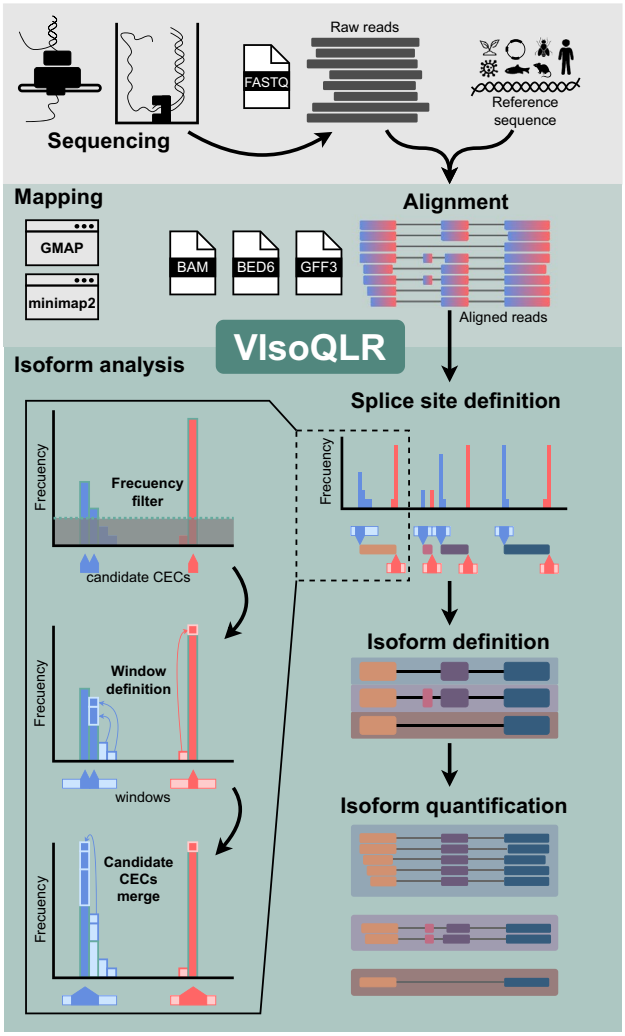


Fig. 1 Workflow for isoform detection and quantification using VIsoQLR. Sequenced long-reads are mapped to generate a BAM, GFF3 or BED6 file containing the coordinates of all transcripts and exons. The frequency of each exon's start and end coordinates are calculated with all reads. The selection of the consensus exon coordinates (CEC) includes the application of several optional features: (1) frequency threshold, which selects the most frequent ones; (2) window definition, where a window is defined around each candidate CECs so that close non-candidate CES are assigned to the nearest candidate CECs; (3) candidate CECs merge, where very close candidate CECs are merged into the most frequent one. Once the final CECs are defined, they are assigned to all read coordinates to define consensus exons and isoforms. Transcripts with all exons fully delimited by CECs are grouped into isoforms for quantification

with the same exons are considered the same isoform. The final isoform collection is defined, presented and quantified. Reads that do not fit into the delimited exons are not considered in this isoform collection, although they could be recovered if other VIsoQLR configuration is applied. Any change in exon coordinates makes isoforms redefined and quantified again on the fly.

User interface

VIsoQLR user interface has two working tabs, the mapping tab, where users can obtain mapped sequences from raw sequencing files using two aligners (GAMP or minimap2), and the isoforms analysis tab, which performs the VIsoQLR tasks providing its main features. In addition, external mapping can be uploaded as BAM, BED6, or GFF3 format.

The control panel of the isoform analysis tab is shown in Fig. 2 (the complete VIsoQLR user interface of the isoform analysis tab is shown in Figure S1). To start the analysis, the user submits the aligned sequences (GFF3, BED6, or BAM formats are allowed) at the top of the control panel (Fig. 2a). Next, the different sequences submitted are displayed in a drop-down menu in the analysis bounding section (Fig. 2b). Sequences are analyzed one at a time on user request. VIsoQLR also allows to restrict analysis to specified regions, for instance to exclude exons coming from a vector or to focus on the splicing events of specific exons. Lastly, in this section, the user has the option to use only full-length PCR sequences. In the "Exon coordinates" menu (Fig. 2c), user can fine-tune the detection of splice sites. Here, options include different approaches to select CECs, such as setting up a minimum percentage of reads supporting the start and end exon coordinates, defining the size of the windows, or merging close CECs. In this control panel section, the user can upload previously defined exon coordinates ("Custom exon coordinates" menu) that replace or are merged with the automatically detected CECs performed by VIsoQLR.

VIsoQLR also support plotting known or previously analyzed transcripts for visual comparison with the isoforms detected in the experiment under analysis. They can be submitted in GTF, or GFF3 format ("Load transcripts from file" menu) (Fig. 2d) and are displayed justified above isoforms with the same exon color codes to facilitate a direct comparison. Along with the transcript ID and size, it is possible to specify additional information about the transcript to be displayed. This information is stored in the last column of these two file formats. Visualizing a reference set of isoforms can help compare new results with previous data using VIsoQLR, those obtained with other software, or known transcripts.

The visual representation of isoforms can be adapted to the user's requirements using the "Display option" menu (Fig. 2e). Lastly, all results, including figures and tables, can be downloaded by setting their prefix in the "Output prefix" menu (Fig. 2f).

Figure 3a shows the collection of isoforms detected by VIsoQLR, including their exon configuration, coordinates, lengths and relative quantification as provided by the software. Here, isoforms detected by an external method uploaded by the user for reference or comparison purposes are shown. The exons of the uploaded isoforms

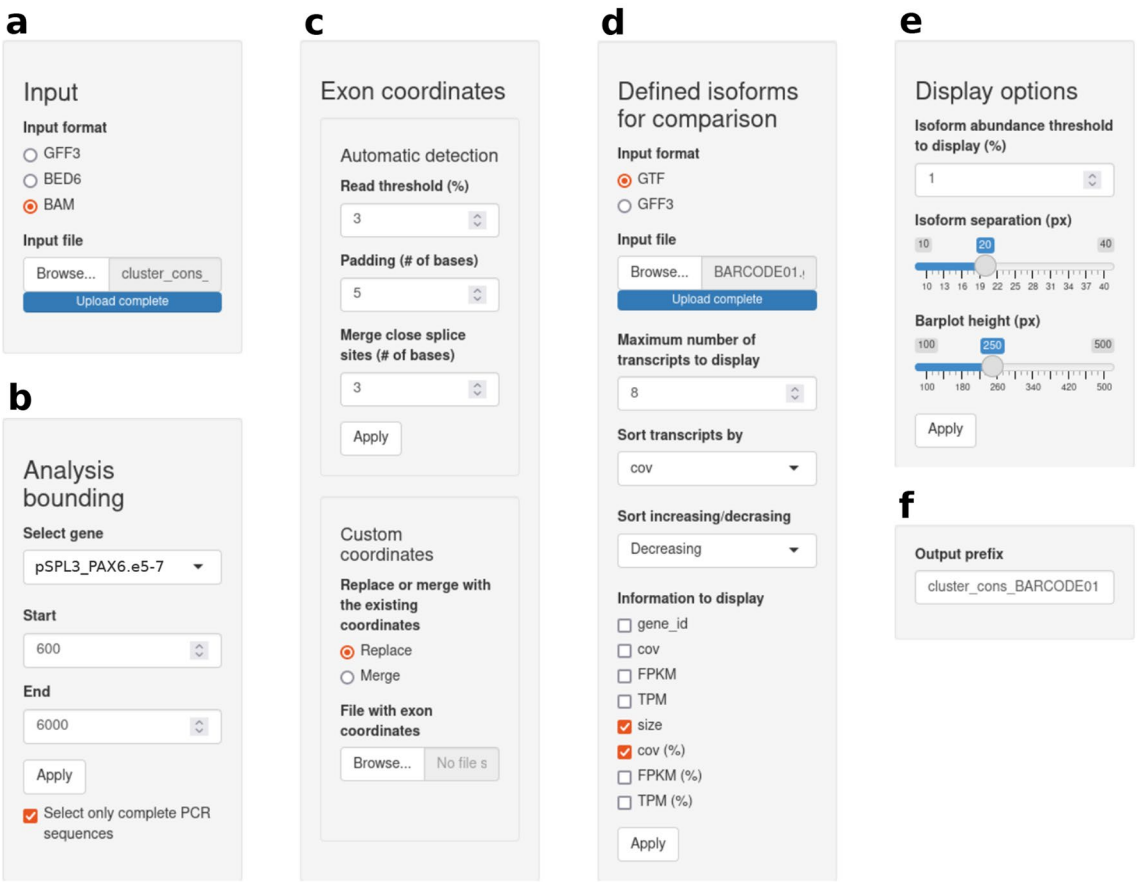


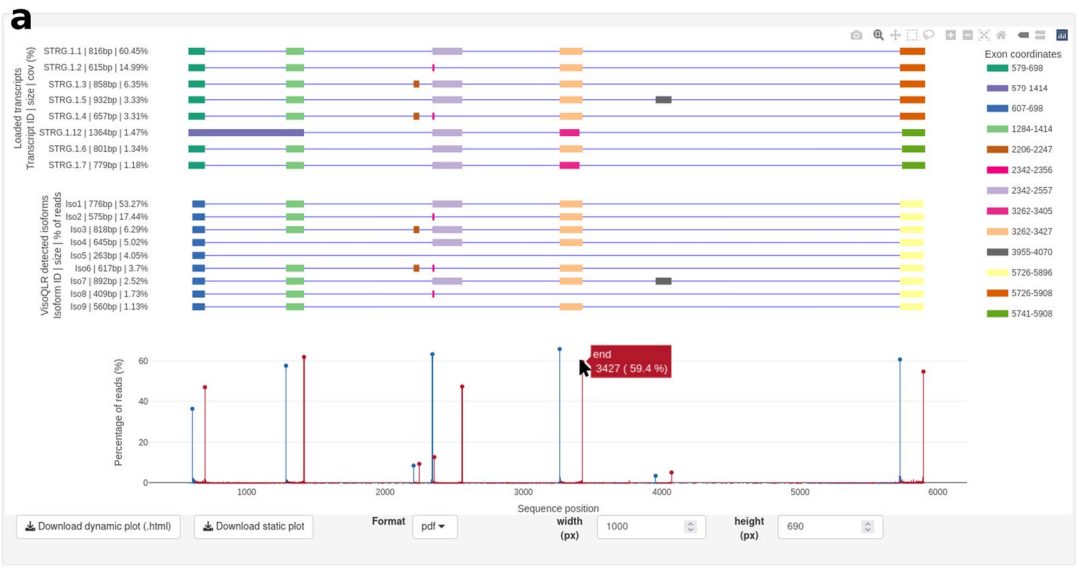
Fig. 2 Control panel of the user interface of VIsoQLR. For visualization purposes, the control panel shown in the application as a single column is here split into six subpanels. **a** Input subpanel where users can select the input file format and upload the aligned sequences. **b** Analysis bounding subpanel allows the analysis of the gene and sequence area. It also contains an option to analyze only the full-length PCR sequences. **c** Exon coordinates subpanel, with the option to automatically detect consensus exon coordinate (CEC). **d** External isoforms subpanel, where users can upload known or previously defined transcripts as a reference to curate the isoforms detected by VIsoQLR. **e** Display options subpanel. Here the user can filter the isoforms to be displayed based on their abundance and fine-tune the graphics. **f** Download prefix subpanel is used to indicate the prefix of all downloadable tables and figures

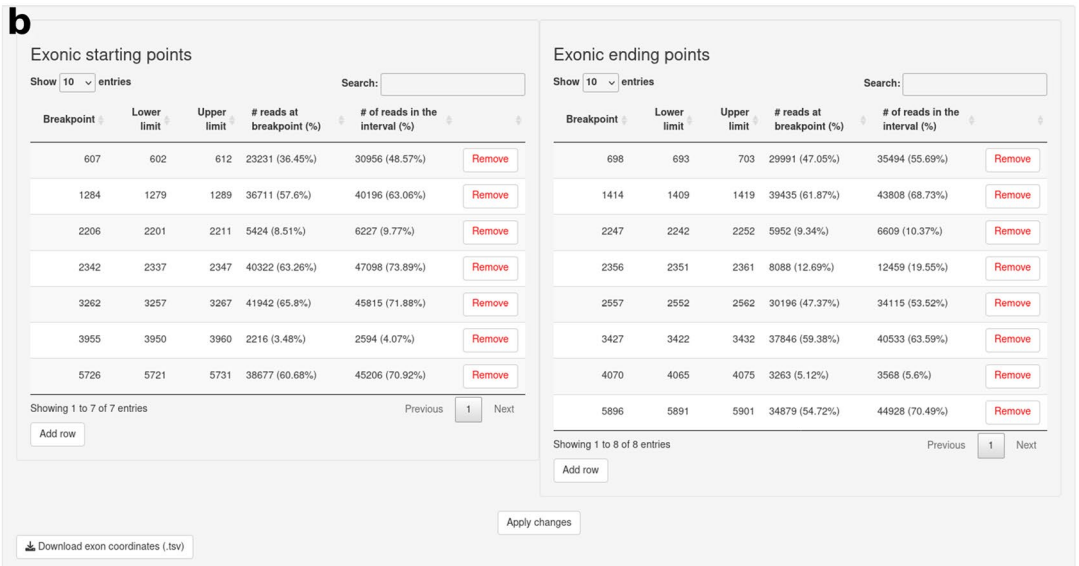
are justified by those detected by VIsoQLR. Color coding is applied to represent identical exons. Below, the start (in blue) and end (in red) exon sites are shown as peaks. The height of the peaks represents the percentage of reads mapped with the same position. A zoom and graphical selection of regions is also provided for close inspection of details. Selected CECs are marked with a dot, and when hovered cursor over, these dots display a tag with the exact coordinates and the percentage of reads supporting it. The figure can be downloaded as an HTML file maintaining all the dynamic properties and a static figure in many formats in the desired size.

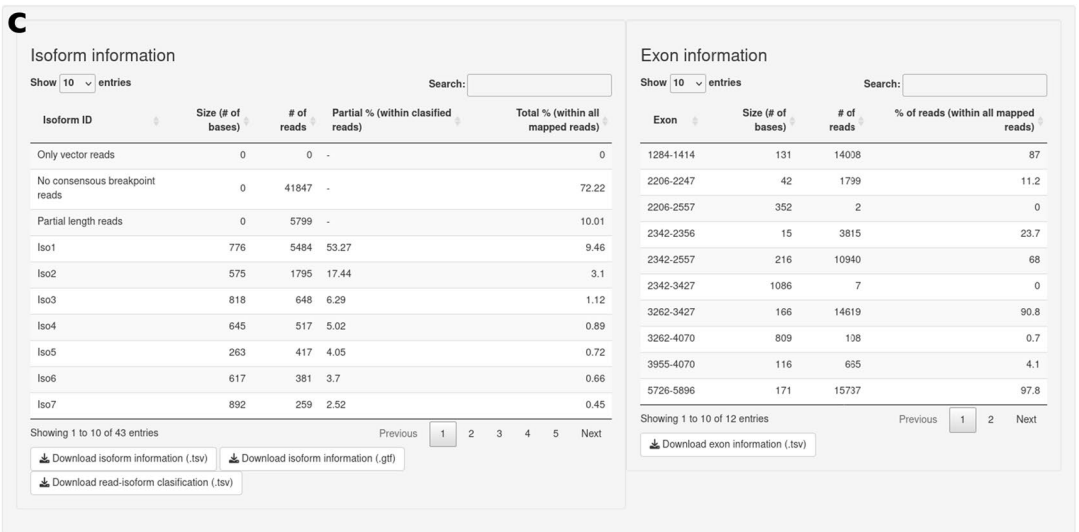
The consensus positions of the start and end exon are shown in two separate tables (Fig. 3b). Their coordinates, as well as the window defining them, can be edited directly in these tables. The manual edition of start and end positions automatically redefines the exons and isoforms and recalculates their relative abundance. These tables provide additional information on the size and abundance of isoforms and exons (Fig. 3c).

A benchmark of isoform detection and quantification using SIRVs

Spike-In RNA Variants (SIRVs) are a collection of synthetic transcripts with known concentrations used for quality control in RNA sequencing. To assess the detection and quantification of isoforms of our software, we analyzed public PacBio long-read sequencing data and tested the performance of three software (VIsoQLR, FLAIR and StringTie2) in detecting isoforms from the SIRV Isoform Mix E0 (Lexogen), containing 69 transcripts from seven genes (see Table S1 for mapped read sequencing metrics). FLAIR (Tang et al. 2020) and StringTie2 (Kovaka et al. 2019) are the most popular tools in whole transcriptome analysis. FLAIR requires exon coordinates to be provided as input.


a
Loaded transcripts
Transcript ID | size | cov (%)
STRG.1.1 | 816bp | 60.45%
STRG.1.2 | 615bp | 14.99%
STRG.1.3 | 858bp | 6.35%
STRG.1.5 | 932bp | 3.33%
STRG.1.4 | 657bp | 3.31%
STRG.1.12 | 1364bp | 1.47%
STRG.1.6 | 801bp | 1.34%
STRG.1.7 | 779bp | 1.18%
VisoQLR detected isoforms
Isoform ID | size | % of reads
Iso1 | 776bp | 53.27%
Iso2 | 575bp | 17.44%
Iso3 | 818bp | 6.29%
Iso4 | 645bp | 5.02%
Iso5 | 263bp | 4.05%
Iso6 | 617bp | 3.7%
Iso7 | 892bp | 2.52%
Iso8 | 409bp | 1.73%
Iso9 | 560bp | 1.13%
Exon coordinates
579-698
570 1414
607-698
1284-1414
2206-2247
2342-2356
2342-2557
3262-3405
3262-3427
3955-4070
5726-5896
5726-5908
5741-5908
Percentage of reads (%)
end
3427 (59.4 %)
Sequence position
Download dynamic plot (.html)
Download static plot
Format
pdf
width (px)
1000
height (px)
690


b
Exonic starting points
Show 10 entries
Search:
Breakpoint | Lower limit | Upper limit | # reads at breakpoint (%) | # of reads in the interval (%)
607 602 612 23231 (36.45%) 30956 (48.57%) Remove
1284 1279 1289 36711 (57.6%) 40196 (63.06%) Remove
2206 2201 2211 5424 (8.51%) 6227 (9.77%) Remove
2342 2337 2347 40322 (63.26%) 47098 (73.89%) Remove
3262 3257 3267 41942 (65.8%) 45815 (71.88%) Remove
3955 3950 3960 2216 (3.48%) 2594 (4.07%) Remove
5726 5721 5731 38677 (60.68%) 45206 (70.92%) Remove
Showing 1 to 7 of 7 entries
Previous 1 Next
Add row
Exonic ending points
Show 10 entries
Search:
Breakpoint | Lower limit | Upper limit | # reads at breakpoint (%) | # of reads in the interval (%)
698 693 703 29991 (47.05%) 35494 (55.69%) Remove
1414 1409 1419 39435 (61.87%) 43808 (68.73%) Remove
2247 2242 2252 5952 (9.34%) 6609 (10.37%) Remove
2356 2351 2361 8088 (12.69%) 12459 (19.55%) Remove
2557 2552 2562 30196 (47.37%) 34115 (53.52%) Remove
3427 3422 3432 37846 (59.38%) 40533 (63.59%) Remove
4070 4065 4075 3263 (5.12%) 3568 (5.6%) Remove
5896 5891 5901 34879 (54.72%) 44928 (70.49%) Remove
Showing 1 to 8 of 8 entries
Previous 1 Next
Add row
Apply changes
Download exon coordinates (.tsv)


c
Isoform information
Show 10 entries
Search:
Isoform ID | Size (# of bases) | # of reads | Partial % (within clasified reads) | Total % (within all mapped reads)
Only vector reads 0 0 - 0
No consensous breakpoint reads 0 41847 - 72.22
Partial length reads 0 5799 - 10.01
Iso1 776 5484 53.27 9.46
Iso2 575 1795 17.44 3.1
Iso3 818 648 6.29 1.12
Iso4 645 517 5.02 0.89
Iso5 263 417 4.05 0.72
Iso6 617 381 3.7 0.66
Iso7 892 259 2.52 0.45
Showing 1 to 10 of 43 entries
Previous 1 2 3 4 5 Next
Download isoform information (.tsv)
Download isoform information (.gtf)
Download read-isoform clasification (.tsv)
Exon information
Show 10 entries
Search:
Exon | Size (# of bases) | # of reads | % of reads (within all mapped reads)
1284-1414 131 14038 87
2206-2247 42 1799 11.2
2206-2557 352 2 0
2342-2356 15 3815 23.7
2342-2557 216 10940 68
2342-3427 1086 7 0
3262-3427 166 14619 90.8
3262-4070 809 108 0.7
3955-4070 116 655 4.1
5726-5896 171 15737 97.8
Showing 1 to 10 of 12 entries
Previous 1 2 Next
Download exon information (.tsv)

◂**Fig. 3** VIsoQLR results panel. **a** Display subpanel containing the isoforms detected by VIsoQLR, including their exon configuration, coordinates, lengths and relative quantification. If uploaded by the user, externally defined isoforms are displayed. The color code is used to identify identical exons. Below isoforms, the frequency of start (blue) and end (red) coordinates are shown. The consensus exon coordinates (CECs) are marked with a dot on each bar, and the exact coordinate and frequency are displayed with the cursor over. All the plots are aligned on the x-axis. This plot can be downloaded as a dynamic figure in HTML or as a static figure in multiple formats in a configurable size. **b** CECs are displayed in two tables (for start and end coordinates) with "Breakpoint", "Lower limit", and "Upper Limit" information that can be edited. The number of reads at the exact CECs and corresponding intervals are displayed. These coordinates can be downloaded as a single table. **c** Extra isoform and exon information regarding their lengths and abundances is displayed and can be downloaded in multiple formats

VIsoQLR and StringTie2 infer them from the sequence distribution. File S1, File S2 and File S3 contain the isoforms detected by VIsoQLR, FLAIR, and StringTie2, respectively.

Figure 4 shows the abundance of each isoform (relative to the gene) of the detected transcripts, which intersect more than 99% of the base positions reciprocally with the gold standard (theoretical coordinates) (Table S2 contains the absolute values). Isoforms SIRV701 and SIRV705, and SIRV604 and SIRV612 were merged into two unique isoforms as they intersect in 99.92% and 99.22%, respectively, and our evaluation criteria cannot distinguish transcripts matching either of these isoforms. VIsoQLR detected 49 (72%) out of the 68 isoforms from the seven SIRVs, FLAIR detected 37 (54%), and StringTie2 12 (18%) (see Table S3 containing the total number of detected isoforms and for each SIRV). The cosine similarity of the abundances between the gold standard and VIsoQLR, FLAIR, and StringTie2 was 0.78, 0.68, and 0.42, respectively, with the partial highest similarity being 0.97 achieved with VIsoQLR in SIRV7 (Table S4). Using minimap2 as sequence mapper for all three software, the results remain similar: VIsoQLR detected 48 out of 68 isoforms (71%), FLAIR 35 (51%), and StringTie2 15 (22%), see Figure S2.

The relative isoform abundance for different values of the "read threshold" in VIsoQLR is represented in Figure S3. The number of isoforms detected out of the 68 SIRV transcripts was 65 (96%), 63 (93%), 59 (87%), 49 (72%), and 41 (60%), with cosines similarities being 0.81, 0.81, 0.80, 0.78, 0.74 for "read threshold" values of 0.25%, 0.5%, 1%, 2%, and 3% respectively (Table S5, Table S6 and Table S7).

With these results on hand, we now illustrate the usage of VIsoQLR in the same PacBio RNASeq experiment analyzing isoforms of two genes, *PAX6* as an example of a low expressed gene, and *TP53*, having an average expression. For both genes we show results with a "read threshold" of 3%. In the case of *PAX6* we manually merged the exon coordinates at the beginning, and at the end of the first and last exon respectively. *PAX6* has a total of 45 mapped reads and the most abundant detected isoform correspond to the canonical transcript (9 reads, 45%). The rest of the isoforms are supported by 2 (10%) or 1 (5%) reads (Figure S4a). *TP53* has 450 reads mapped. The most abundant detected isoform has 166 reads (88.8%) and again corresponds to the canonical transcript. The rest of the isoforms are supported with less than 4 reads (Figure S4b).

Case study

To show the applicability of VIsoQLR, we performed a multi-exonic minigene assay for exons 5 to 7 of the *PAX6* gene using DNA from a healthy individual. The design of our cloned sequence in the exon trapping expression pSPL3 vector is depicted in Fig. 5a. We sequenced the amplified cDNA on a MinION flow cell and mapped the generated reads with GMAP (see Table S1 for mapped read sequencing metrics). The aligned reads were uploaded into VIsoQLR, and isoforms were calculated using default parameters and keeping only full-length PCR transcripts. To have an external reference of the methodology to detect exons and characterize isoforms, we applied StringTie2 to the same data (see Materials and methods). In Fig. 5b, we show isoforms detected by VIsoQLR at the top, justified with the results obtained by StringTie2 that were uploaded using the "Load transcripts from file" menu. Both tracks show isoforms with an abundance above 1%, and they are sorted decreasingly by this field (see Table S8 and Table S9 with the absolute abundance values for all detected isoforms). The splicing isoforms from this sample were additionally analyzed by semi-quantitative capillary electrophoresis (CE) of fluorescent amplicons to estimate the proportion of each isoform and compare it with the results obtained by VIsoQLR. The fluorescent emission peaks of each isoform are represented in the electropherogram in Fig. 5c.

VIsoQLR detects 40 isoforms (File S4), nine having an abundance above 1% of reads. In the case of StringTie2, it detects 58 isoforms (File S4), eight of them composed of more than 1% of the reads. The most abundant isoform detected by VIsoQLR (53.1%) corresponds to the canonical transcript. This is composed of five exons (SD6-EXON5-EXON6-EXON7-SA2), including the two specific exons of the vector and three exons of *PAX6*, with a size of 776 bp. StringTie2 also reports this as the most abundant isoform (60.5%), but different start and end positions make the transcript larger (816 bp). In fact, all StrinTie2 isoforms have extra bases at the begging and end compared to VIsoQLR isoforms. This canonical isoform corresponds to CE's highest peak at 772 bp (Fig. 5b).

The second most abundant isoform (17.4%) with 575 bp detected by VIsoQLR is an alternative transcript in which the EXON6 was partially skipped due to the use of an alternative exonic donor site (Grønskov et al. 1999). This

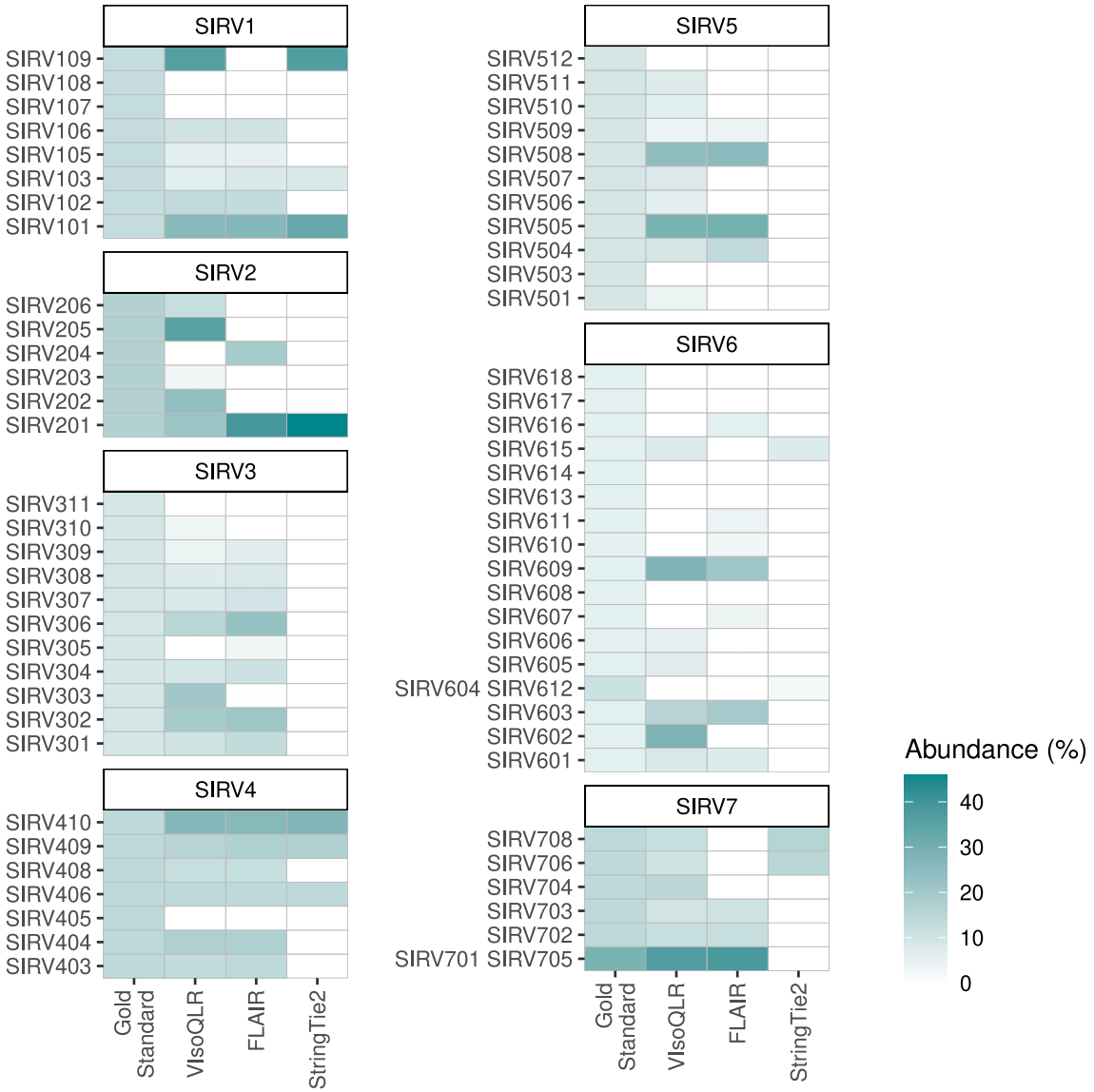


Fig. 4 Isoform abundance in the gold standard, VIsoQLR, FLAIR, and StringTie2. The relative abundance of 68 isoforms in seven Spike-In RNA Variants (SIRVs) sequenced in a RNASeq experiment using PacBio is shown as detected by the three programs together with their theoretical concentration. All transcripts have equimolar concentrations. Transcripts were considered identical if they intersected 99%. SIRV701 and SIRV705, and SIRV604 and SIRV612 were merged as the comparison methodology used does not differentiate transcripts matching either of these isoforms, as they intersect over 99% of their bases

transcript also corresponds to the second most abundant isoform in StringTie2 (15.0%), with a size of 615 bp and the second highest emission peak at 569 bp in CE. The third most abundant isoform called by both methods also coincides (6.3% and 6.4% in VIsoQLR and StringTie2, respectively). This is a second canonical transcript for *PAX6* in which an alternative exon 5a of 42 bp (Fig. 5a) is included (SD6-EXON5-EXON5a-EXON6-EXON7-SA2), making the isoform 818 bp long in VIsoQLR and 858 bp in StringTie2. This isoform also corresponds to the electropherogram's third highest peak of 816 bp.

Both methods also called some minor isoforms below 5%. Iso6 (3.6% and 617 bp) and Iso7 (2.5% and 892 bp) detected by VIsoQLR were also called by StringTie2, as STRG.1.4 (3.3% and 657 bp) and STRG.1.5 (3.3% and 932 bp), and correspond to in the electrophoretic peaks of 612 bp and 892 bp, respectively. However, four other minor isoforms were called by VIsoQLR but not by StringTie2. Iso4 (5.0% and 645 bp) and Iso9 (1.3% and 560 bp) appear in the electropherogram at 642 bp and 552 bp peaks, respectively. But, Iso5 and Iso8 were detected neither by StringTie2 nor CE. Finally, three isoforms

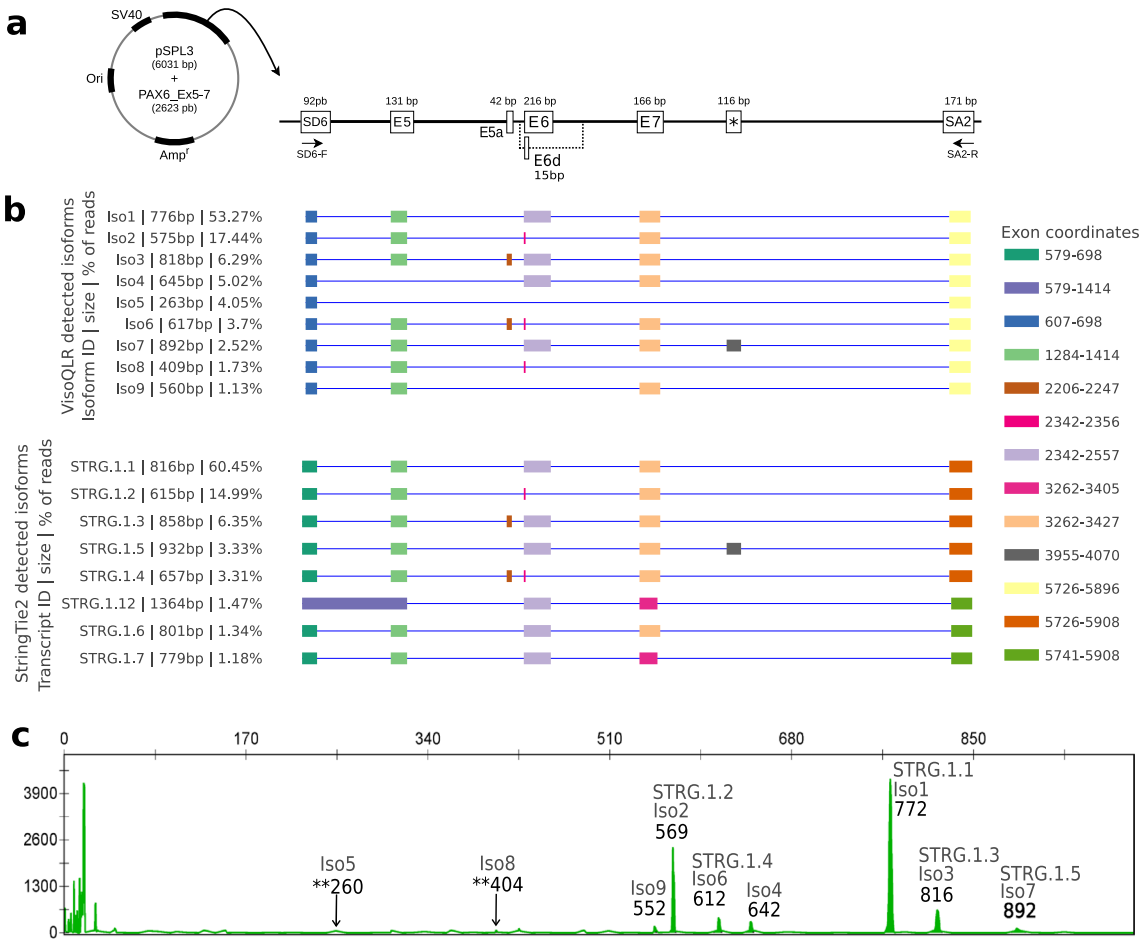


Fig. 5 Minigene design and isoforms detected from the splicing assay. **a** The exon trapping vector pSPL3 contains exons 5 to 7 of *PAX6* (NM_000280.4). This vector contains a SV40 promoter, SD6 (splice donor 6) and SA2 (splice acceptor 2), and the Ampicillin resistance gene (Amp^r). The size in base pairs (bp) of the *PAX6* insert and the pSPL3 vector are shown. The forward (SD6-F) and reverse (SA2-R) primers are indicated. The amplified transcripts of the minigene are composed of SD6, E5 (exon5), E5a (alternative exon5), E6 (exon6), E6d (partial deleted exon 6), E7 (exon7), * (artifact exon), and SA2. **b** The isoforms detected by VIsoQLR are shown on the top track, including their exon configuration, coordinates, lengths and relative quantification. On the bottom track, isoforms detected by StringTie2 are shown. Only isoforms above 1% are represented and sorted by abundance. The color code is used to identify identical exons. **c** Semi-quantitative PCR electropherogram. The x-axis represents the migration time, which correlates with the size of the molecules. The y-axis represents the absorbance intensity in Relative Fluorescence Units (RFU). The size, in base pairs (bp), is shown at the top of each peak, along with the corresponding isoforms detected by VIsoQLR and StringTie2. ** Coordinates without well-defined peaks

(STRG.1.12, STRG.1.6, and STRG.1.7) were called only by StringTie2.

Discussion

Up to 15% of the pathogenic variants affect RNA splicing not only in canonical splicing sites but also in exonic and intronic non-canonical sites (Riolo et al. 2021), which can hinder their detection during conventional DNA screening or misinterpret their clinical significance by in silico splicing predictions. Many algorithms have been developed to analyze and quantify the splicing pattern of genes using RNA sequencing (Byrne et al. 2017; Fu et al. 2018; Wyman et al. 2019; Kovaka et al. 2019; Tang et al. 2020; Hu et al. 2021). Although most can be used to analyze a single locus, they lack visualization and editing options that allow close exploration of regions of interest. VIsoQLR has been developed to fill this gap that is specifically required when studying a reduced set of genes of interest. The study of Mendelian diseases is a clear example that requires this kind of feature. In their diagnosis, the seek for pathogenic spliceogenic variants usually is restricted to genes with a known association with the phenotype. In many of them, mutations in a single gene can explain most cases, e.g., *PAX6* and aniridia (Landsend et al. 2021), *ABCA4* and Stargardt disease (Cremers et al. 2020), or *NF1* and Neurofibromatosis type 1 (Koster et al. 2021). In addition, to obtain a conclusive diagnosis,

functional characterization is needed for a correct interpretation of their effect in the splicing event. Here, LRS has become the state-of-the-art technology to study splicing (Amarasinghe et al. 2020) allowing the analysis of full transcripts due to the length of the sequences obtained, about 10 Kb. In contrast, classical approaches, such as capillary electrophoresis, only detect the most abundant isoforms and need a manual inspection of lengths to associate absorption peaks with known or in silico-predicted isoforms.

Although the VIsoQLR main and significant feature is the possibility to curate isoforms interactively, we wanted to assess its initial automatic analysis compared with other tools. Thus, we have performed a benchmark to determine the quality of the isoforms detected by VIsoQLR, StringTie2 and FLAIR software in a typical LRS experiment. Out of the three tools, VIsoQLR detects with 99% accuracy of exon coordinates 72% of the known isoforms using SIRVs, outperforming the other tools. Remarkably, VIsoQLR performs better than FLAIR which needs the exon coordinates to provide quantification. In the case of StringTie2, the missed isoforms have mainly two observed behaviors: it extends the boundaries of some isoforms, and it does not call isoforms without certain exons. An example of the isoforms extension is seen in SIRV3 where the longest isoforms (SIRV301, SIRV302, SIRV303, SIRV304, SIRV306, SIRV307) have an extension between 43 and 62 bp (see Figure S5a). This extension makes these isoforms more different than the 1% allowed in the comparison. SIRV5 represents an example of how StringTie2 calls isoforms with extra exons (see Figure S5b). In this case the final exons of SIRV509 and SIRV510 are added to other isoforms (SIRV501, SIRV508, SIRV505). Moreover, there are two not detected isoforms (SIRV505 and SIRV5010) that have a central exon missing that StringTie2 seems to force to report. Finally, although the last exon of SIRV509 has been added to other isoforms, this is not present in StringTie2 results as detected transcripts with this exon start 8 kb before the theoretical SIRV509 does.

Focusing now on VIsoQLR results, as a proof of concept, we performed SIRV isoform detection using different values of the threshold to define CECs (read threshold) and observed that most missing isoforms were below 2% (default value). At the same time, lowering this threshold has the cost of detecting more isoforms: 452 isoforms (55 with an abundance above 1%) detected with the read threshold at 2% and 833 (68 with an abundance above 1%) at 0.25% (see Table S6). In contrast, in other cases such as the analysis of *PAX6* and *TP53* in a PacBio dataset, and the *PAX6* minigene, best performances are obtained with values above 2%. Several factors may affect the results obtained for particular genes, including the gene sequencing depth, complexity of the splicing isoforms, and the quality of the data. Thus, the description of the isoforms collection of individual genes seems to require a close inspection and sometimes a manual configuration of certain parameters that would allow the user to focus the analysis on main or rare isoforms. An example of this is shown with *PAX6* and *TP53* from PacBio dataset, where the canonical transcripts are detected, but the rest of the transcripts have a very low number of reads supporting them to make quantification conclusions.

In addition, we present a case study using LRS to illustrate how the interactive graphical interface and the editing features allow a thorough analysis of the set of isoforms produced by gene using a minigene experiment. Our case study aims to provide a comprehensive report of the *PAX6* isoforms in a healthy individual. Mutations in *PAX6* are responsible for nearly 100% of the cases of congenital aniridia, a rare developmental disease characterized by abnormalities in the iris and fovea (Blanco-Kelly et al. 2021). Two hotspot exons, EX5 and EX6, are prone to suffer naturally alternative splicing, resulting in a mixture of different splicing isoforms (Tarilonte et al. 2022). We compared results for a multiexon *PAX6* minigene splicing assay provided by VIsoQLR using its automatic isoform detection with those detected by semi-quantitative CE. This comparison includes features not only specific to VIsoQLR but also those of LRS. VIsoQLR detected all isoforms present in the electropherogram, and there is a correlation in the abundance of the ranked isoforms provided by both methods. As expected, VIsoQLR provides a more extensive set of isoforms detected. However, some caveats in the abundance of small isoforms should be considered, as LRS tends to overrepresent them (Amarasinghe et al. 2020). An example of this might be the isoforms Iso5 and Iso8 detected by VIsoQLR (Figure S6) with an abundance of 4% and 1.7%, respectively, but hardly distinguishable from the noise in the electropherogram (Fig. 5).

In addition, we also provide an example of how VIsoQLR can be used to visualize and compare results coming from other isoform detection algorithms. In this example, we chose StringTie2, a popular transcriptome analysis for short and long reads. The first evident difference is that StringTie2 reports a few extra bases for all isoforms at the outer boundary of the first and last exons compared to VIsoQLR. This extension (also seen in the SIRVs analysis) seems to be an artifact as it is not present in BAM files used by StringTie2 and VIsoQLR (Figure S7 and File S5). This could also explain that StringTie2 reports systematically longer isoforms than electropherogram and VIsoQLR. Its missed isoforms (Iso4 and Iso9), appearing in the electropherogram and detected by VIsoQLR, have missed exons 5 and 6, respectively. This pattern is also observed in the SIRVs analysis where it did not call isoforms without certain exons. Moreover this could also explain why it did not detect Iso5 and Iso8, where the first one missed exons 5 to 7, and the second one, the exon 7. In any case, the relative proportion of the most abundant isoforms detected by StingTie2

correlates with the isoforms reported with VIsoQLR and CE.

VIsoQLR demonstrates an accurate isoform automatic detection using LRS data. On top of this, it provides a flexible, interactive and editable visualization framework for the manual inspection of potential gene splice sites. This feature allows a fast custom analysis of single-locus that is not available in any other tool. VIsoQLR can also be used to manually curate results from other isoform detection algorithms by adding its complete visualization and editing features. The docker containerization plus user interface allows users without deep knowledge of bioinformatics to analyze their data in a user-friendly program.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00439-023-02539-z.

Author contributions GNM has conceived, designed and coded the tool and analyzed the data. MC, CRS and AT designed and performed experiments on the case study. GNM, PM and MC wrote the paper. PM and MC conceived the work and obtained the funding. All authors reviewed and approved the manuscript.

Funding This work was supported by the Instituto de Salud Carlos III (ISCIII) of the Spanish Ministry of Health [PI17/00164, PI18/00579, PI20/00851, IMP/00019], co-funded by European Regional Development Fund (FEDER funds) "A way to make Europe"; Centro de Investigación Biomédica en Red en Enfermedades Raras (CIBERER) [06/07/0036]; Comunidad de Madrid (CAM) [RAREGenomics Project, B2017/BMD-3721]; and Organización Nacional de Ciegos Españoles (ONCE). GNM is supported by a contract of the Comunidad de Madrid [PEJ-2020-AI/BMD-18610]. PM and MC are supported by a Miguel Servet program contract from ISCIII [CP16/00116, CPII21/00015 and CPII17/00006, respectively].

Data availability VisoQLR code is available at https://github.com/TBLabFJD/VIsoQLR. The docker image is available at https://hub.docker.com/r/tblabfjd/visoqlr. Public RNAseq PacBio long-read sequencing data is available at https://github.com/PacificBiosciences/DevNet/wiki/Sequel-II-System-Data-Release:-Universal-Human-Reference-(UHR)-Iso-Seq. SIRV reference sequences and transcript annotation (gold standard) are available at https://www.lexogen.com/sirvs/download/. Case study data is available at https://github.com/TBLabFJD/VIsoQLR/tree/main/example.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Abdel-Ghany SE, Hamilton M, Jacobi JL et al (2016) A survey of the sorghum transcriptome using single-molecule long reads. Nat Commun. https://doi.org/10.1038/ncomms11706

Amarasinghe SL, Su S, Dong X et al (2020) Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21:1–16. https://doi.org/10.1186/s13059-020-1935-5

Anna A, Monika G (2018) Splicing mutations in human genetic disorders: examples, detection, and confirmation. J Appl Genet 59:253–268. https://doi.org/10.1007/s13353-018-0444-7

Blanco-Kelly F, Tarilonte M, Villamar M et al (2021) Genetics and epidemiology of aniridia: updated guidelines for genetic study. Arch Soc Esp Oftalmol 96(Suppl 1):4–14. https://doi.org/10.1016/J.OFTALE.2021.02.002

Byrne A, Beaudin AE, Olsen HE et al (2017) Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat Commun. https://doi.org/10.1038/ncomms16027

Chang W, Cheng J, Allaire J, et al (2021) shiny: Web Application Framework for R

Cooper TA (2005) Use of minigene systems to dissect alternative splicing elements. Methods 37:331–340. https://doi.org/10.1016/j.ymeth.2005.07.015

Cremers FPM, Lee W, Collin RWJ, Allikmets R (2020) Clinical spectrum, genetic complexity and therapeutic approaches for retinal disease caused by ABCA4 mutations. Prog Retin Eye Res. https://doi.org/10.1016/J.PRETEYERES.2020.100861

Dai M, Xu Y, Sun Y et al (2022) Revealing diverse alternative splicing variants of the highly homologous SMN1 and SMN2 genes by targeted long-read sequencing. Mol Genet Genomics 297:1039–1048. https://doi.org/10.1007/S00438-022-01874-6

Evans DGR, Bowers N, Burkitt-Wright E et al (2016) Comprehensive RNA Analysis of the NF1 gene in classically affected NF1 affected individuals meeting NIH criteria has high sensitivity and mutation negative testing is reassuring in isolated cases with pigmentary features only. EBioMedicine 7:212–220. https://doi.org/10.1016/J.EBIOM.2016.04.005

Felício V, Ramalho AS, Igreja S, Amaral MD (2016) mRNA-based detection of rare CFTR mutations improves genetic diagnosis of cystic fibrosis in populations with high genetic heterogeneity. Clin Genet 91:476–481. https://doi.org/10.1111/cge.12802

Fraile-Bethencourt E, Valenzuela-Palomo A, Díez-Gómez B et al (2019) Minigene splicing assays identify 12 Spliceogenic Variants of BRCA2 Exons 14 and 15. Front Genet. https://doi.org/10.3389/FGENE.2019.00503

Fu S, Ma Y, Yao H et al (2018) IDP-de novo: de novo transcriptome assembly and isoform annotation by hybrid sequencing. Bioinformatics 34:2168–2176. https://doi.org/10.1093/bioinformatics/bty098

Gonorazky HD, Naumenko S, Ramani AK et al (2019) Expanding the Boundaries of RNA sequencing as a diagnostic tool for rare mendelian disease. Am J Hum Genet 104:466–483. https://doi.org/10.1016/J.AJHG.2019.01.012

Gonzalez-Garay ML (2016) Introduction to Isoform Sequencing Using Pacific Biosciences Technology. Transcriptomics and Gene Regulation. Springer, Netherlands, Dordrecht, pp 141–160

Grønskov K, Rosenberg T, Sand A, Brøndum-Nielsen K (1999) Mutational analysis of PAX6: 16 novel mutations including 5 missense mutations with a mild aniridia phenotype. Eur J Hum Genet 7:274–286. https://doi.org/10.1038/SJ.EJHG.5200308

Helman G, Compton AG, Hock DH et al (2021) Multiomic analysis elucidates Complex I deficiency caused by a deep intronic variant in NDUFB10. Hum Mutat 42:19–24. https://doi.org/10.1002/HUMU.24135

Hu Y, Fang L, Chen X et al (2021) LIQA: long-read isoform quantification and analysis. Genome Biol. https://doi.org/10.1186/s13059-021-02399-8

Jurkute N, Cancellieri F, Pohl L et al (2022) Biallelic variants in coenzyme Q10 biosynthesis pathway genes cause a retinitis pigmentosa phenotype. NPJ Genomic Med. https://doi.org/10.1038/S41525-022-00330-Z

Koster R, Brandão RD, Tserpelis D et al (2021) Pathogenic neurofibromatosis type 1 (NF1) RNA splicing resolved by targeted RNAseq. NPJ Genomic Med. https://doi.org/10.1038/S41525-021-00258-W

Kovaka S, Zimin AV, Pertea GM et al (2019) Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. https://doi.org/10.1186/S13059-019-1910-1

Kuang Z, Canzar S (2018) Tracking alternatively spliced isoforms from long reads by SpliceHunter. Methods Mol Biol 1751:73–88. https://doi.org/10.1007/978-1-4939-7710-9_5

Landsend ECS, Lagali N, Utheim TP (2021) Congenital aniridia – a comprehensive review of clinical features and therapeutic approaches. Surv Ophthalmol 66:1031–1050. https://doi.org/10.1016/J.SURVOPHTHAL.2021.02.011

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/BIOINFORMATICS/BTY191

Lord J, Baralle D (2021) Splicing in the diagnosis of rare disease: advances and challenges. Front Genet 12:1146. https://doi.org/10.3389/fgene.2021.689892

Mehmood A, Laiho A, Venäläinen MS et al (2020) Systematic evaluation of differential splicing tools for RNA-seq studies. Brief Bioinform 21:2052–2065. https://doi.org/10.1093/bib/bbz126

Okubo M, Noguchi S, Awaya T et al (2022) RNA-seq analysis, targeted long-read sequencing and in silico prediction to unravel pathogenic intronic events and complicated splicing abnormalities in dystrophinopathy. Hum Genet. https://doi.org/10.1007/S00439-022-02485-2

Paul L, Kubala P, Horner G et al (2016) SIRVs: Spike-In RNA Variants as External Isoform Controls in RNA-Sequencing. Biorxiv. https://doi.org/10.1101/080747

R Core Team (2020) R: A language and environment for statistical computing. R Found Stat Comput Vienna, Austria

Riolo G, Cantara S, Ricci C (2021) What's wrong in a jump? Prediction and validation of splice site variants, Methods Protoc, p 4

Sahlin K, Tomaszkiewicz M, Makova KD, Medvedev P (2018) Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. Nat Commun 9:4601. https://doi.org/10.1038/s41467-018-06910-x

Sangermano R, Khan M, Cornelis SS et al (2018) ABCA4 midigenes reveal the full splice spectrum of all reported noncanonical splice site variants in Stargardt disease. Genome Res 28:100–110. https://doi.org/10.1101/GR.226621.117/-/DC1

Sievert C (2020) Interactive Web-Based Data Visualization with R plotly, and shiny. Chapman and Hall/CRC, Boca Raton

Tang AD, Soulette CM, van Baren MJ et al (2020) Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun 11:1438. https://doi.org/10.1038/s41467-020-15171-6

Tardaguila M, De La Fuente L, Marti C et al (2018) SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res 28:396–411. https://doi.org/10.1101/gr.222976.117

Tarilonte M, Ramos P, Moya J et al (2022) Activation of cryptic donor splice sites by non-coding and coding PAX6 variants contributes to congenital aniridia. J Med Genet 59:428–437. https://doi.org/10.1136/jmedgenet-2020-106932

Wadman RI, Jansen MD, Stam M et al (2020) Intragenic and structural variation in the SMN locus and clinical variability in spinal muscular atrophy. Brain Commun. https://doi.org/10.1093/braincomms/fcaa075

Whiley PJ, De La Hoya M, Thomassen M et al (2014) Comparison of mRNA splicing assay protocols across multiple laboratories: recommendations for best practice in standardized clinical testing. Clin Chem 60:341–352. https://doi.org/10.1373/CLINCHEM.2013.210658

Wu TD, Watanabe CK (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859–1875. https://doi.org/10.1093/bioinformatics/bti310

Wyman D, Balderrama-Gutierrez G, Reese F et al (2019) A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. Biorxiv. https://doi.org/10.1101/672931