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

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

[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

*To whom correspondence should be addressed.

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 other alternative models to detect the formation of aberrant mRNAs. Long-read sequencing is a suitable technique to identify and quantify mRNA isoforms. Available isoform clusterization and/or quantification tools are generally designed for the whole transcriptome analysis. Experiments focusing on a single locus analysis 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 and maturation in low-throughput splicing assays. This tool takes sequences aligned to a reference, defines consensus splice sites, and quantifies isoforms. Users can edit splice sites through dynamic and interactive graphics and tables as part of their manual curation. Known transcripts, or isoforms detected by other methods, can also be imported as references for comparison. Here, we explain VIsoQLR principles and features, and show its applicability in a case study example using Nanopore sequencing. VIsoQLR is available at https://github.com/TBLabFJD/VIsoQLR.

Keywords: isoforms, long-read sequencing, shiny app, alternative splicing, rare diseases, minigenes, graphical interface

INTRODUCTION

Spliceogenic DNA variants are an underestimated cause of genetic diseases (Lord & 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 to disease traits is a challenging task that requires *in vivo* or *in vitro* functional studies. Global transcriptomics approaches are available (Mehmood et

al., 2020); however, analysis is usually restricted to a single locus (or few loci) in Mendelian diseases. 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 & Monika, 2018) or minigenes-based exon trapping assays if the damaged primary tissue is not accessible (Cooper, 2005).

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. On the other hand, short-read sequencing of RT-PCR products 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).

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 Mandalorion (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 & Canzar, 2018). To the best of our knowledge, there is no interactive tool that allows close inspection and fine-tuning analysis of one gene at a time on data from low-throughput approaches.

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 also be downloaded from https://hub.docker.com/r/tblabfjd/visoqlr.

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 ~200bp intronic sequence on each side into the exon trapping expression pSPL3 vector. This construction was transfected in HEK-293T 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.

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 (Wu & Watanabe, 2005). GMAP parameters were: ‘-n1’ to avoid chimeric alignments, ‘--cross-species’ for a more sensitive search for canonical splicing, ‘--gff3-add-separators=0’ to prevent the “###” separator after each query sequence, and ‘-f 2’ to generate a GFF3 file containing exons coordinates. In the case study described here, VIsoQLR was applied using the automatic detection of splice sites with default parameters and filtering out truncated reads.

Data was also analyzed with StringTie2 (Kovaka et al., 2019) to detect splicing isoforms. In this analysis, we mapped reads using GMAP with the same parameters as the above, setting ‘-f samse’ to generate a SAM file, which was transformed into a BAM file, sorted, and indexed. 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. Here we suggest the GMAP aligner based on the performance of our own experience on Nanopore sequences. Still, other mappers, such as minimap2 (Li, 2018), can be used. Mapped data is uploaded to VIsoQLR. Using this input as GFF3 or BED6 format, VIsoQLR defines consensus exon coordinates (CECs) 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 3%); 2) a position window, in which the user can define a window, where other non-candidate CEC 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 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 discarded. Any change in exon coordinates makes isoforms redefined and quantified again on the fly.

User interface

The VIsoQLR control panel of the user interface is shown in Figure 2 (the complete VIsoQLR user interface is shown in Supplementary Figure 1). To start the analysis, the user must submit a GFF3 or BED6 file with aligned sequences at the top of the control panel (Figure 2A). Once reads are uploaded, the different sequences submitted are displayed in a drop-down menu in the analysis bounding section (Figure 2B). Sequences are analyzed one at a time. Also, the user can set the sequence boundaries from the same section to be explored. This is a useful option if the user wants to avoid working with 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 transcripts. VIsoQLR has a default configuration to detect splice sites that can be changed in the “Exon coordinates” menu (Figure 2C). 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 allows 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) (Figure 2D) and are displayed justified below 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 figure settings can be adapted to the user’s requirements using the “Display option” menu (Figure 2E). Lastly, all results, including figures and tables, can be downloaded by setting their prefix in the “Output prefix” menu (Figure 2F).

Figure 3A shows the collection of isoforms detected by VIsoQLR, including their exon configuration, coordinates, lengths and relative quantification. Here, isoforms detected by an external method using the same data uploaded by the user for reference or comparison purposes are shown. The exons of the “reference” isoforms are justified by those detected by VIsoQLR. Color coding is used 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 (Figure 3B). Their coordinates, as well as the window defining them, can be edited directly in these tables. Any edition made in these tables 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 (Figure 3C).

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 Figure 4A. We sequenced the amplified cDNA on a MinION flow cell and mapped the generated reads with GMAP. The aligned reads were uploaded into VIsoQLR, and isoforms were calculated using default parameters and keeping only full-length transcripts. To have an external reference of the methodology to detect exons and characterize isoforms, we applied StringTie2 (Kovaka et al., 2019) to the same data (see Material and Methods). In Figure 4C, 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. 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 Figure 4B.

VIsoQLR detects 40 isoforms (Supplementary File 1), nine having an abundance above 1% of reads. In the case of StringTie2, it detects 58 isoforms (Supplementary File 1), 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 776bp. StringTie2 also reports this as the most abundant isoform (60.5%), but different start and end positions make the transcript larger (816bp). 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 772bp (Figure 4B).

The second most abundant isoform (17.4%) with 575bp 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 transcript also corresponds to the second most abundant isoform in StringTie2 (15.0 %) with a size of 615bp, and to the second highest emission peak at 569bp 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 42bp (Figure 4A) is included (SD6-EXON5-EXON5a-EXON6-EXON7-SA2), making the isoform 818bp long in VIsoQLR and 858bp in StringTie2. This isoform also corresponds to the electropherogram's third highest peak of 816bp.

Both methods also called some minor isoforms below 5%. Iso6 (3.6% and 617bp) and Iso7 (2.5% and 892bp) detected by VIsoQLR were also called by StringTie2, as STRG.1.4 (3.3% and 657bp) and STRG.1.5 (3.3% and 932bp), and correspond to in the electrophoretic peaks of 612bp and 892bp, respectively. However, four other minor isoforms were called by VIsoQLR but not by StringTie2. Iso4 (5.0% and 645bp) and Iso9 (1.3% and 560bp) appear in the electropherogram at

642bp and 552bp peaks, respectively. But, Iso5 and Iso8 were detected neither by StringTie2 nor CE. Finally, three isoforms (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 by involving canonical splicing sites but also by exonic and intronic non-canonical sites in increasing cases (Riolo et al., 2021). Therefore, they can be missed during conventional DNA screening or misinterpreted 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; Hu et al., 2021; Kovaka et al., 2019; Tang et al., 2020; Wyman et al., 2019). 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 performing low-throughput experiments. The study of Mendelian diseases is a clear example that requires this kind of facility. In their diagnosis, the seek for pathogenic spliceogenic variants usually is restricted to genes with a known association with the disease. In many of them, 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). Functional characterization is crucial for correctly interpreting potential spliceogenic variants, especially in clinically actionable genes.

In addition to describing the VIsoQLR method and its features, we present a case study using LRS to illustrate its application. LRS is the state-of-the-art technology to study splicing (Amarasinghe et al., 2020). It allows the analysis of full transcripts, reaching sequences of about 10 Kb. It can report the whole repertoire of isoforms and provide their sequences. 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.

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 multi-exon *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 (Supplementary Figure 2) with an abundance of 4% and 1.7%, respectively, but hardly distinguishable from the noise in the electropherogram (Figure4).

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 (Kovaka et al., 2019). 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 seems to be an artifact as it is not

present in BAM files used by StringTie2 nor in the GFF3 file used in VIsoQLR (Supplementary Figure 3 and Supplementary File 2). This can also be why StringTie2 reports systematically longer isoforms than electropherogram and VIsoQLR. 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 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 in bioinformatics to analyze their data in a user-friendly program.

Author contributions

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

Data availability

VisoQLR code is available at https://github.com/TBLabFJD/VIsoQLR, docker image is available at https://hub.docker.com/r/tblabfjd/visoqlr. Case study data is available at https://github.com/TBLabFJD/VIsoQLR/example

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]; 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].

Conflict of Interest: The authors declare no conflicts of interest.

References

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A., & Reddy, A. S. N. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms11706

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 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. *Journal of Applied Genetics*, *59*(3), 253–268. https://doi.org/10.1007/s13353-018-0444-7

Blanco-Kelly, F., Tarilonte, M., Villamar, M., Damián, A., Tamayo, A., Moreno-Pelayo, M. A., Ayuso, C., & Cortón, M. (2021). Genetics and epidemiology of aniridia: Updated guidelines for genetic study. *Archivos de La Sociedad Espanola de Oftalmologia*, *96 Suppl 1*, 4–14. https://doi.org/10.1016/J.OFTALE.2021.02.002

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., DuBois, R. M., Forsberg, E. C., Akeson, M., & Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms16027

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *shiny: Web Application Framework for R*. https://cran.r-project.org/package=shiny

Cooper, T. A. (2005). Use of minigene systems to dissect alternative splicing elements. *Methods*, *37*(4), 331–340. https://doi.org/10.1016/j.ymeth.2005.07.015

Cremers, F. P. M., Lee, W., Collin, R. W. J., & Allikmets, R. (2020). Clinical spectrum, genetic complexity and therapeutic approaches for retinal disease caused by ABCA4 mutations. *Progress in Retinal and Eye Research*, *79*. https://doi.org/10.1016/J.PRETEYERES.2020.100861

Fu, S., Ma, Y., Yao, H., Xu, Z., Chen, S., Song, J., & Au, K. F. (2018). IDP-denovo: De novo transcriptome assembly and isoform annotation by hybrid sequencing. *Bioinformatics*, *34*(13), 2168–2176. https://doi.org/10.1093/bioinformatics/bty098

Gonzalez-Garay, M. L. (2016). *Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq)* (pp. 141–160). https://doi.org/10.1007/978-94-017-7450-5_6

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. *European Journal of Human Genetics: EJHG*, *7*(3), 274–286. https://doi.org/10.1038/SJ.EJHG.5200308

Hu, Y., Fang, L., Chen, X., Zhong, J. F., Li, M., & Wang, K. (2021). LIQA: long-read isoform quantification and analysis. *Genome Biology*, *22*(1). https://doi.org/10.1186/s13059-021-02399-8

Koster, R., Brandão, R. D., Tserpelis, D., van Roozendaal, C. E. P., van Oosterhoud, C. N., Claes, K. B. M., Paulussen, A. D. C., Sinnema, M., Vreeburg, M., van der Schoot, V., Stumpel, C. T. R. M., Broen, M. P. G., Spruijt, L., Jongmans, M. C. J., Lesnik Oberstein, S. A. J., Plomp, A. S., Misra-Isrie, M., Duijkers, F. A., Louwers, M. J., … Blok, M. J. (2021). Pathogenic

neurofibromatosis type 1 (NF1) RNA splicing resolved by targeted RNAseq. *NPJ Genomic Medicine*, *6*(1). https://doi.org/10.1038/S41525-021-00258-W

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*(1). https://doi.org/10.1186/S13059-019-1910-1

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

Landsend, E. C. S., Lagali, N., & Utheim, T. P. (2021). Congenital aniridia – A comprehensive review of clinical features and therapeutic approaches. *Survey of Ophthalmology*, *66*(6), 1031–1050. https://doi.org/10.1016/J.SURVOPHTHAL.2021.02.011

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/BIOINFORMATICS/BTY191

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

Mehmood, A., Laiho, A., Venälaïnen, M. S., McGlinchey, A. J., Wang, N., & Elo, L. L. (2020). Systematic evaluation of differential splicing tools for RNA-seq studies. *Briefings in Bioinformatics*, *21*(6), 2052–2065. https://doi.org/10.1093/bib/bbz126

R Core Team. (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. https://www.r-project.org/

Riolo, G., Cantara, S., & Ricci, C. (2021). What's Wrong in a Jump? Prediction and Validation of Splice Site Variants. *Methods and Protocols*, *4*(3). https://doi.org/10.3390/MPS4030062

Sahlin, K., Tomaszkiewicz, M., Makova, K. D., & Medvedev, P. (2018). Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. *Nature Communications*, *9*(1), 4601. https://doi.org/10.1038/s41467-018-06910-x

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC. https://plotly-r.com

Tang, A. D., Soulette, C. M., van Baren, M. J., Hart, K., Hrabeta-Robinson, E., Wu, C. J., & Brooks, A. N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications*, *11*(1), 1438. https://doi.org/10.1038/s41467-020-15171-6

Tardaguila, M., De La Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., Edelmann, M., Ezkurdia, I., Vazquez, J., Tress, M., Mortazavi, A., Martens, L., Rodriguez-Navarro, S., Moreno-Manzano, V., & Conesa, A. (2018). SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Research*, *28*(3), 396–411. https://doi.org/10.1101/gr.222976.117

Tarilonte, M., Ramos, P., Moya, J., Fernandez-Sanz, G., Blanco-Kelly, F., Swafiri, S. T., Villaverde, C., Romero, R., Tamayo, A., Gener, B., Calvas, P., Ayuso, C., & Corton, M. (2022). Activation of cryptic donor splice sites by non-coding and coding PAX6 variants contributes to congenital aniridia. *Journal of Medical Genetics*, *59*(5), 428–437. https://doi.org/10.1136/jmedgenet-2020-106932

Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*(9), 1859–1875. https://doi.org/10.1093/bioinformatics/bti310

Wyman, D., Balderrama-Gutierrez, G., Reese, F., Jiang, S., Rahmanian, S., Zeng, W., Williams, B., Trout, D., England, W., Chu, S., Spitale, R. C., Tenner, A., Wold, B., & Mortazavi, A. (2019). A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *BioRxiv*. https://doi.org/10.1101/672931

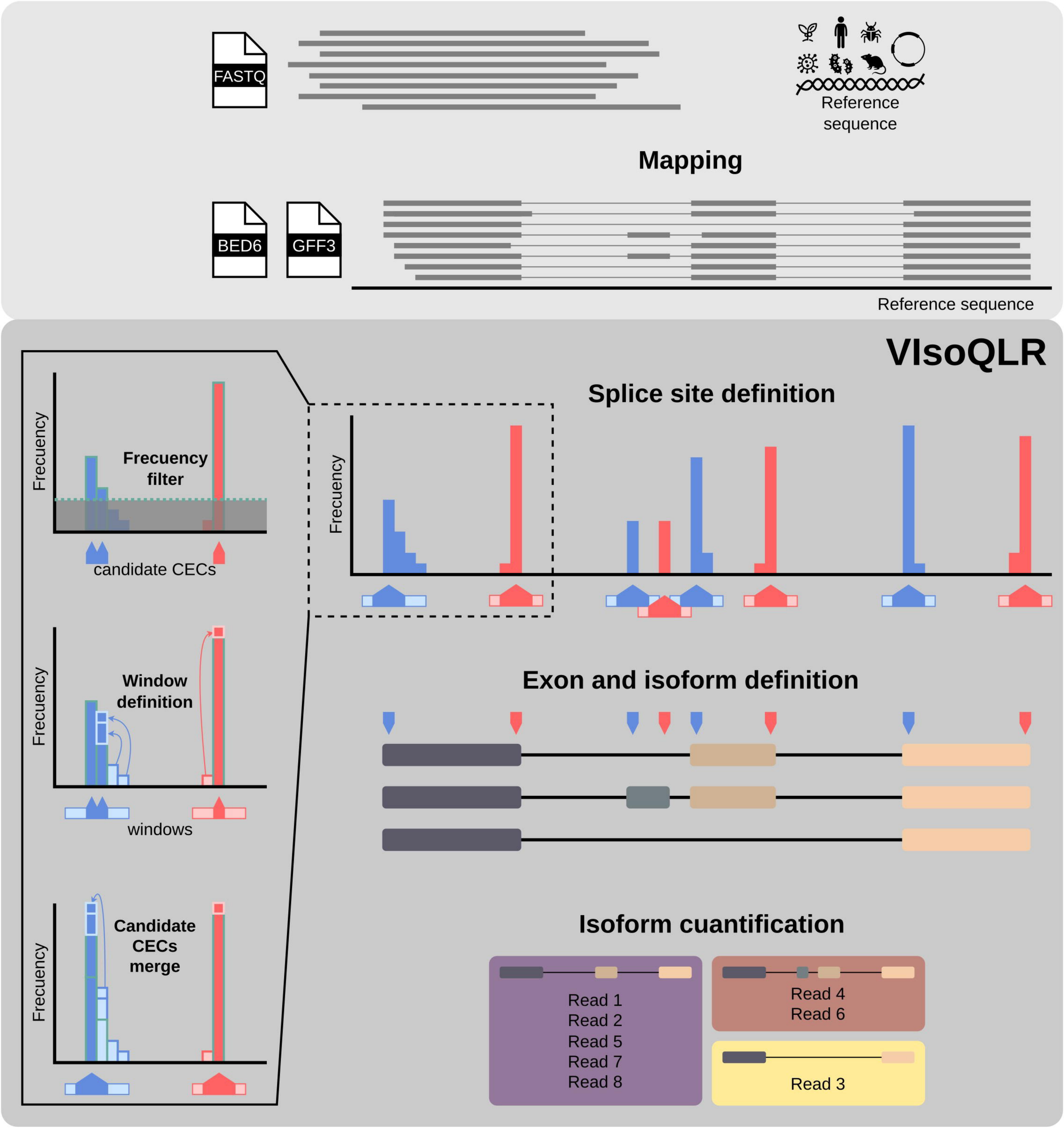
Figure 1. Workflow for isoform detection and quantification using VIsoQLR. Sequenced long-reads are mapped to generate a GFF3 or BED6 file containing the coordinates of all transcripts and exons. The frequency of each exon's start and end coordinates are calculated with all reads. The selection of the 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.

Figure 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 transcripts. **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.

Figure 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.

Figure 4. Minigene design and isoforms detected from the splicing assay. A. The exon trapping vector pSPL3 contains exons 5, alternative 5 (EX5a), 6, and 7 of *PAX6* (NM_000280.4). This vector contains a SV40 promoter, SD6 (splice donor 6) and SA2 (splice acceptor 2), and

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. **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 with an abundance above 1% are represented 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. * Coordinates without well-defined peaks.


FASTQ
Reference sequence
Mapping
BED6
GFF3
Reference sequence
VIsoQLR
Frecuency
Frecuency filter
candidate CECs
Splice site definition
Frecuency
Frecuency
Window definition
windows
Exon and isoform definition
Frecuency
Candidate CECs merge
Isoform cuantification
Read 1
Read 2
Read 5
Read 7
Read 8
Read 4
Read 6
Read 3

A

Input

Input format

- (•) GFF3
- ( ) BED6

Input file

Browse... | cluster_cons_I

Upload complete

B

Analysis bounding

Select gene

Secuencia

Start

0

End

8106

Apply

[x] Select only full length transcripts

C

Exon coordinates

Automatic detection

Read threshold (%)

3

Padding (# of bases)

5

Merge close splice sites (# of bases)

3

Apply

Custom coordinates

Replace or merge with the existing coordinates

- (•) Replace
- ( ) Merge

File with exon coordinates

Browse... | No file s

D

Defined isoforms for comparison

Input format

- (•) GTF
- ( ) GFF3

Input file

Browse... | BARCODE01.

Upload complete

Maximum number of transcripts to display

8

Sort transcripts by

cov

Sort increasing/decrasing

Decreasing

Information to display

- [ ] gene_id
- [ ] cov
- [ ] FPKM
- [ ] TPM
- [x] size
- [x] cov (%)
- [ ] FPKM (%)
- [ ] TPM (%)

Apply

E

Display options

Isoform abundance threshold to display (%)

1

Isoform separation (px)

10 20 40

10 13 16 19 22 25 28 31 34 37 40

Barplot height (px)

100 250 500

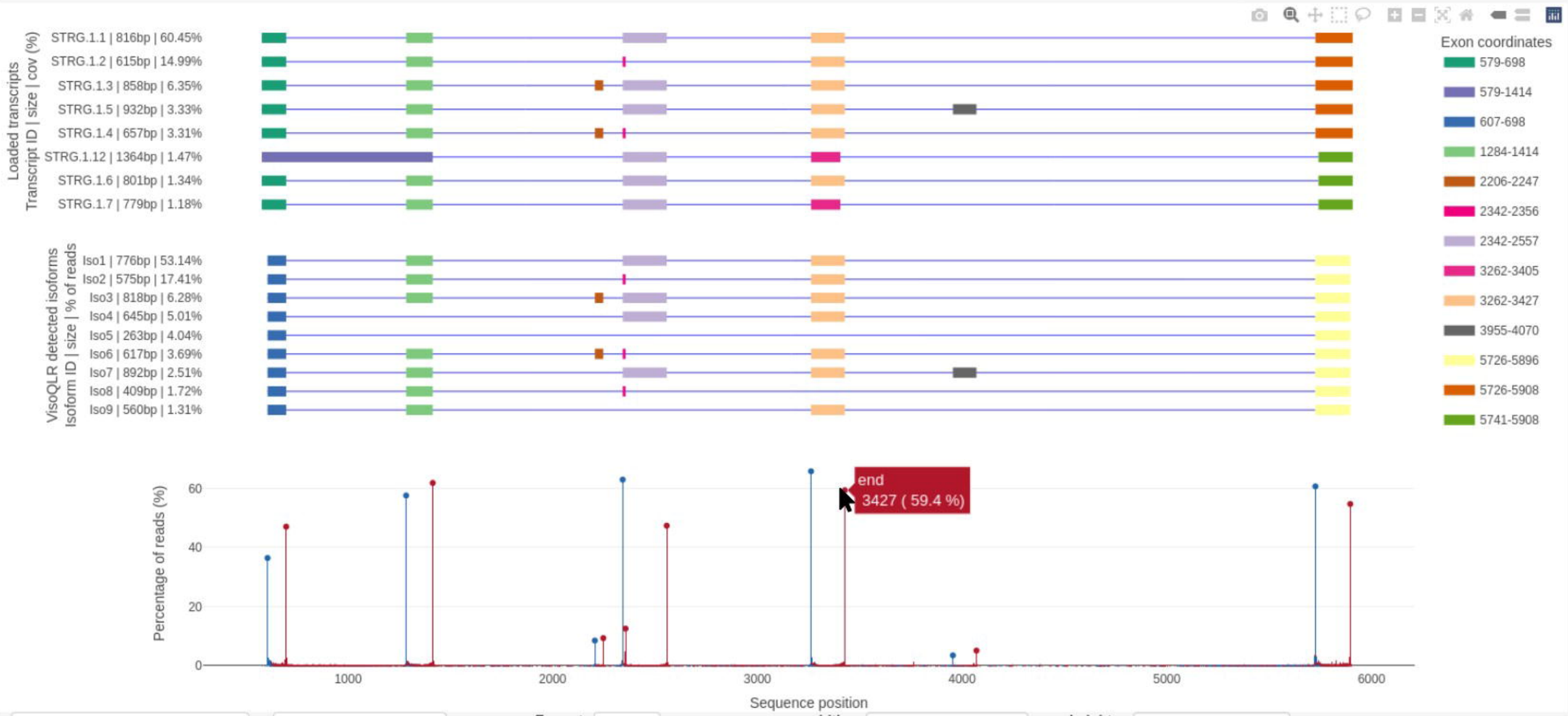
100 180 260 340 420 500

Apply

F

Output prefix

cluster_cons_BARCODE01



Download dynamic plot (.html) | Download static plot | Format: png | width (px): 1000 | height (px): 690

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	36708 (57.59%)	40193 (63.06%)	Remove
2206	2201	2211	5409 (8.49%)	6209 (9.74%)	Remove
2342	2337	2347	40127 (62.96%)	46881 (73.55%)	Remove
3262	3257	3267	41932 (65.79%)	45793 (71.84%)	Remove
3955	3950	3960	2216 (3.48%)	2594 (4.07%)	Remove
5726	5721	5731	38673 (60.67%)	45202 (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	29980 (47.04%)	35482 (55.67%)	Remove
1414	1409	1419	39420 (61.85%)	43766 (68.66%)	Remove
2247	2242	2252	5939 (9.32%)	6589 (10.34%)	Remove
2356	2351	2361	7991 (12.54%)	12300 (19.3%)	Remove
2557	2552	2562	30195 (47.37%)	34113 (53.52%)	Remove
3427	3422	3432	37840 (59.37%)	40525 (63.58%)	Remove
4070	4065	4075	3247 (5.09%)	3550 (5.57%)	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)

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	41784	-	72.13
Partial length reads	0	5814	-	10.04
Iso1	776	5488	53.14	9.47
Iso2	575	1798	17.41	3.1
Iso3	818	649	6.28	1.12
Iso4	645	517	5.01	0.89
Iso5	263	417	4.04	0.72
Iso6	617	381	3.69	0.66
Iso7	892	259	2.51	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)

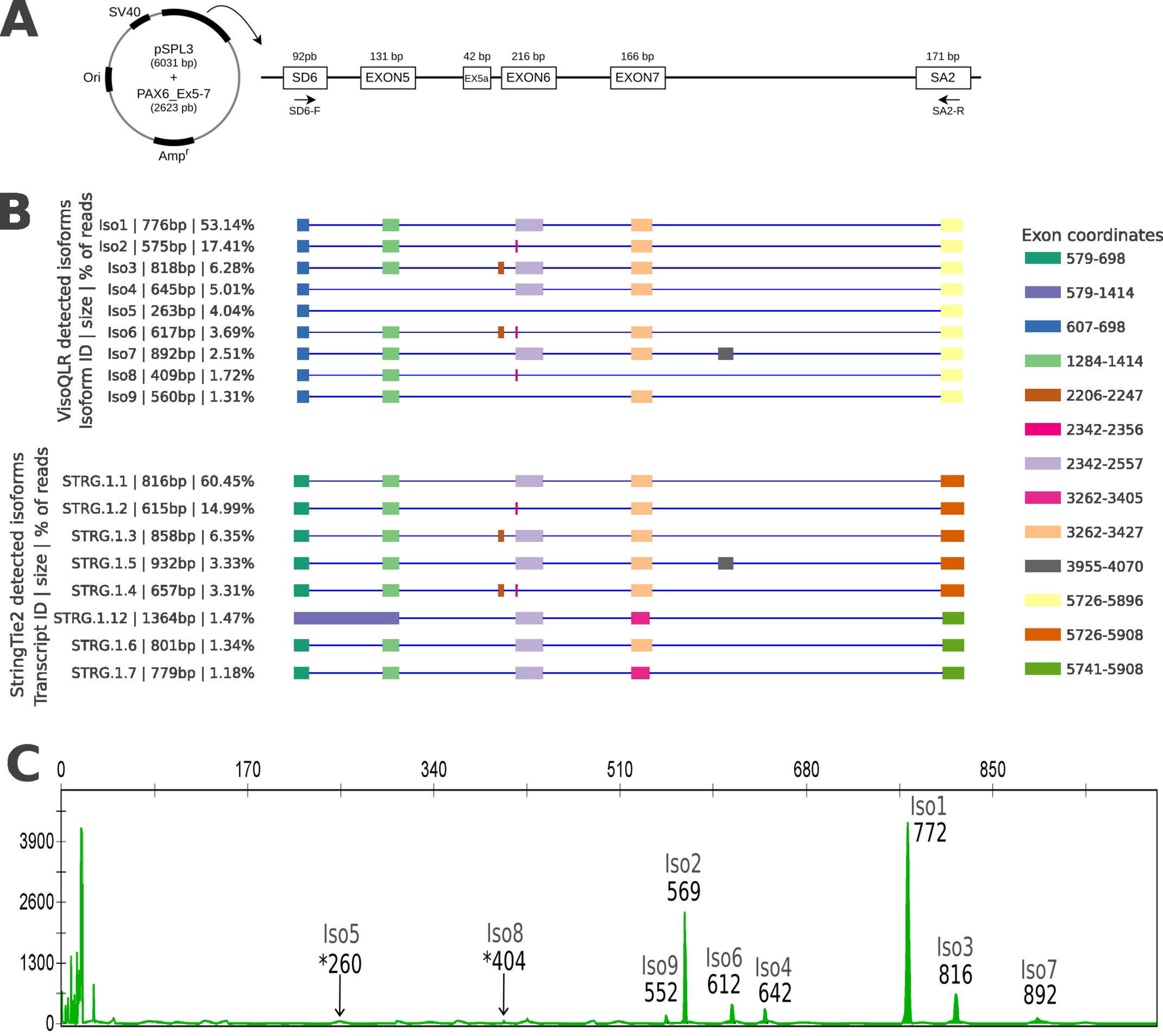
Exon information

Show 10 entries | Search:

Exon	Size (# of bases)	# of reads	% of reads (within all mapped reads)
1284-1414	131	14056	87.1
2206-2247	42	1801	11.2
2206-2557	352	2	0
2342-2356	15	3820	23.7
2342-2557	216	10946	67.8
2342-3427	1086	7	0
3262-3427	166	14659	90.8
3262-4070	809	109	0.7
3955-4070	116	665	4.1
5726-5896	171	15783	97.8

Showing 1 to 10 of 12 entries | Previous 1 2 Next

Download exon information (.tsv)


A
SV40
pSPL3
(6031 bp)
+
PAX6_Ex5-7
(2623 pb)
Ori
Amp^r
92pb
SD6
SD6-F
131 bp
EXON5
42 bp
EX5a
216 bp
EXON6
166 bp
EXON7
171 bp
SA2
SA2-R
B
VisoQLR detected isoforms
Isoform ID | size | % of reads
Iso1 | 776bp | 53.14%
Iso2 | 575bp | 17.41%
Iso3 | 818bp | 6.28%
Iso4 | 645bp | 5.01%
Iso5 | 263bp | 4.04%
Iso6 | 617bp | 3.69%
Iso7 | 892bp | 2.51%
Iso8 | 409bp | 1.72%
Iso9 | 560bp | 1.31%
StringTie2 detected isoforms
Transcript ID | size | % of reads
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%
Exon coordinates
579-698
579-1414
607-698
1284-1414
2206-2247
2342-2356
2342-2557
3262-3405
3262-3427
3955-4070
5726-5896
5726-5908
5741-5908
C
0
170
340
510
680
850
3900
2600
1300
0
Iso1
772
Iso2
569
Iso5
*260
Iso8
*404
Iso9
552
Iso6
612
Iso4
642
Iso3
816
Iso7
892