Mini-IsoQLR: a pipeline for isoform quantification using long-reads sequencing data for single *locus* analysis

Gonzalo Núñez-Moreno[1,2], Alejandra Tamayo[1,3], 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

Motivation

DNA variants altering the pre-mRNA splicing process represent an underestimated cause of human genetic diseases. Their association to disease traits needs to be confirmed by means of functional assays from patient cell lines or other models to detect the formation of aberrant mRNAs. Long-read sequencing is a suitable technique to identify and quantify splicing isoforms. Available tools for isoform clusterization and/or quantification are not designed for the analysis of a single *locus* without previous information of splice sites.

Results

Mini-IsoQLR identifies and quantifies known and novel isoforms using long-read sequencing data from the mRNA expression and maturation of single-*locus* splicing assays. This pipeline uses GMAP aligner to generate a GFF3 file that is used to extract the consensus splice sites, define, and quantify all isoforms present.

Availability and implementation

Mini-IsoQLR requires GMAP, R (version 3.6), and R packages ggplot, plyr and optparse. It is freely available in GitHub (https://github.com/TBLabFJD/Mini-IsoQLR). This software is provided under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Introduction

Spliceogenic DNA variants are an underestimated cause of genetic diseases (Lord and Baralle, 2021). They disrupt the canonical splicing sites or introduce other cryptic exonic or intronic splicing sites leading to an aberrant mRNA maturation process. The effect of genomic variation in modifying splicing patterns is a challenging task that needs functional characterization studies. Besides global transcriptomics approaches (Mehmood *et al.*, 2020), the usual procedures for Mendelian diseases with a suspected altered expression in a single *locus* are splicing RT-PCR assays applied to patients samples (Anna and Monika, 2018), or minigenes-based exon trapping assays when the damaged primary tissue is not accessible (Cooper, 2005). Whichever approach is used, *in vivo* or *in vitro* splicing assays, identification and quantification of the resultant isoform pattern are crucial to associate these variants to specific phenotypes.

The analytical techniques traditionally used have some important drawbacks to address 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 is unable to determine the exon organization of the full-sized 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).

There are various bioinformatics tools available for isoform quantification and analysis. Most of them were developed for transcriptome analysis and require a reference annotation file (such as GTF with the gene and exon coordinates) (LIQA (Hu *et al.*, 2021), FLAIR (Tang *et al.*, 2020), TALON (Wyman *et al.*, 2019), and Mandalorion (Byrne *et al.*, 2017)), or short-read sequencing data to infer splice sites (FLAIR and IDP (Fu *et al.*, 2018)). Other tools only work with Pacific Biosciences' (PacBio) single-molecule real-time (SMRT) sequencing data (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)). Oxford Nanopore Technologies (ONT) has developed a pipeline for the detection of *de novo* isoforms based on clustering of long transcriptome reads (https://github.com/nanoporetech/pipeline-nanopore-denovo-isoforms). Although useful in genomic context, clustering algorithms are difficult to fine-tune when a high-resolution determination of isoforms is required as in the case of splicing studies of a single gene using minigene assays or directly RT-PCR analysis from patient RNA. StringTie2 (Kovaka et al., 2019) is a reference-guided transcriptome assembler and quantifier that works without a reference annotation file. As it has been designed for whole transcriptome analysis, it lacks the fine-tunning options desirable for single locus sequencing data.

Herein, we present Mini-IsoQLR, a non-clustering-based pipeline for detecting and quantifying the isoform mRNA pattern obtained by the functional characterization of potentially spliceogenic variants derived from LRS data without previous knowledge of splice sites.

Software description

Input files

Mini-IsoQLR needs the sequence of the gene or fragment that is being analyzed in FASTA format. This file will be used to build a custom reference file that is used by the GMAP aligner (Wu and Watanabe, 2005). The second required file is a FASTQ file containing raw reads for each sample that is going to be analyzed.

Algorithm methodology

The general workflow of Mini-IsoQLR is summarized in **Figure 1**. First, Mini-IsoQLR aligns the raw reads (FASTQ files) to the reference using GMAP with the parameter "--format=2" which generates a GFF3 file containing the coordinates of the exons and cDNA from all reads. This is a key step as the rest of the analysis relies on the accuracy of the mapping.

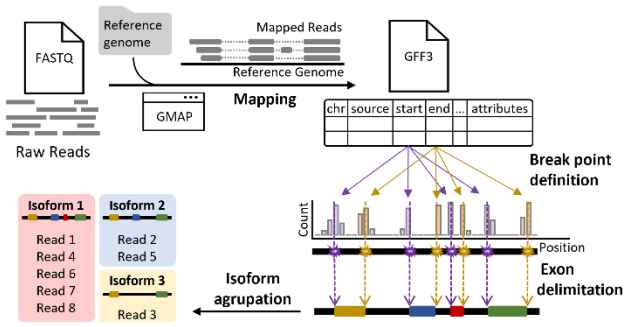

Figure 1. Mini-IsoQLR workflow. GMAP take as input a FASTQ file and a reference genome to perform the alignment generating a GFF3 file. The rest of the pipeline is run by Mini-IsoQLR.R: it defines the consensus splicing breakpoints, delimits the exons, identifies isoforms, and assign the reads to isoforms.

After the mapping, the script Mini-IsoQLR.R reads the GMAP GFF3 together with the standard error and performs the following steps: (1) Defines the consensus breakpoints, start and end of exons which are present in more than 5% of the reads (default value), (2) applies a window flanking region around consensus breakpoints, (3) assigns every read to a consensus breakpoint according to flanking regions to configure the exons, (4) defines isoforms by concatenating the established exons, and (5) quantifies reads and relative abundance of every isoform.

As an optional feature, users can specify the beginning and end coordinates to avoid, for example, exons from the expression vector. If that parameter is chosen, Mini-IsoQLR will trim exons outside delimited regions. It is also possible to specify known splice sites with custom flanking regions as well as specify break points that reads must have to keep in the analysis (e.g. keep reads with first and last break points when a PCR is performed before library preparation to avoid artifacts).

Output

Mini-IsoQLR extracts and displays the isoforms repertoire in single-*locus* RT-PCR amplicons or minigene-based plasmid assay using LRS transcriptomics. It quantifies the number of reads detected for each isoform and reports their relative abundance. The pipeline generates four sets of figures (in bitmaps and vectorial format) including figures showing all detected isoforms, their proportion, and exon coordinates (**Supplementary Fig. S1**); figures showing the distributions of exon/intron breakpoints of all reads (histogram by intervals (**Supplementary Fig. S2**) and by single position (**Supplementary Fig. S3**)); and the aligned combination of these two (**Supplementary Fig. S4**). Mini-IsoQLR also generates 3 tables with information of all consensus exon/intron breakpoints, information of all isoforms and the classification into isoform or category (unmapped reads, only expression vector exons or reads without consensus splice sites) of all reads.

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]; 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). G.N.-M. is supported by a contract of the Comunidad de Madrid [PEJ-2020-AI/BMD-18610]. P.M. and M.C. are supported by a Miguel Servet program contract from ISCIII [CP16/00116, CPII17/00006, respectively].

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

References

Abdel-Ghany,S.E. *et al.* (2016) A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.*, **7**.

Amarasinghe,S.L. *et al.* (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.*, **21**, 1–16.

Anna,A. and Monika,G. (2018) Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet.*, **59**, 253–268.

Byrne,A. *et al.* (2017) Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.*, **8**.

Cooper,T.A. (2005) Use of minigene systems to dissect alternative splicing elements. *Methods*, **37**, 331–340.

Fu,S. *et al.* (2018) IDP-denovo: De novo transcriptome assembly and isoform annotation by hybrid sequencing. *Bioinformatics*, **34**, 2168–2176.

Gonzalez-Garay,M.L. (2016) Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq)., pp. 141–160.

Hu,Y. *et al.* (2021) LIQA: long-read isoform quantification and analysis. *Genome Biol.*, **22**.

Kovaka,S. *et al.* (2019) Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.*, **20**.

Kuang,Z. and Canzar,S. (2018) Tracking alternatively spliced isoforms from long reads by SpliceHunter. *Methods Mol. Biol.*, **1751**, 73–88.

Lord,J. and Baralle,D. (2021) Splicing in the Diagnosis of Rare Disease: Advances and Challenges. *Front. Genet.*, **12**, 1146.

Mehmood,A. *et al.* (2020) Systematic evaluation of differential splicing tools for RNA-seq studies. *Brief. Bioinform.*, **21**, 2052–2065.

Sahlin,K. *et al.* (2018) Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. *Nat. Commun.*, **9**, 4601.

Tang,A.D. *et al.* (2020) Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.*, **11**, 1438.

Tardaguila,M. *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.

Wu,T.D. and Watanabe,C.K. (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Wyman,D. *et al.* (2019) A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv*.