Bioinformatics, 2021, 1–3
doi: 10.1093/bioinformatics/btab103
Advance Access Publication Date: 15 February 2021
Applications Note

Data and Text Mining

IsoTV: processing and visualizing functional features of translated transcript isoforms

Siddharth Annaldasula[1,2], Martyna Gajos [1,2] and Andreas Mayer [1,*]

[1]Otto-Warburg-Laboratory, Max Planck Institute for Molecular Genetics, Berlin 14195, Germany and [2]Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin 14195, Germany

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 18, 2020; revised on February 6, 2021; accepted on February 11, 2021editorial decision on February 10, 2021;

Abstract

Summary: Despite the continuous discovery of new transcript isoforms, fueled by the recent increase in accessibility and accuracy of long-read RNA sequencing data, functional differences between isoforms originating from the same gene often remain obscure. To address this issue and enable researchers to assess potential functional consequences of transcript isoform variation on the proteome, we developed IsoTV. IsoTV is a versatile pipeline to process, predict and visualize the functional features of translated transcript isoforms. Attributes such as gene and isoform expression, transcript composition and functional features are summarized in an easy-to-interpret visualization. IsoTV is able to analyze a variety of data types from all eukaryotic organisms, including short- and long-read RNA-seq data. Using Oxford Nanopore long read data, we demonstrate that IsoTV facilitates the understanding of potential protein isoform function in different cancer cell types.

Availability and implementation: IsoTV is available at https://github.molgen.mpg.de/MayerGroup/IsoTV, with the corresponding documentation at https://isotv.readthedocs.io/.

Contact: mayer@molgen.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Alternative splicing diversifies transcriptomes of metazoans by generating transcript isoforms (Ule and Blencowe, 2019). Coding transcripts are then translated to protein isoforms, contributing to proteome diversity. The development of long-read sequencing technologies has enabled the characterization of full length transcripts and identification of non-canonical isoforms (Sakamoto *et al.*, 2020). However, there is a lack of dedicated computational tools for protein isoform analysis and visualization, especially regarding novel isoforms.

Here, we introduce IsoTV (Isoform Transcript Visualizer), a versatile Snakemake (Koster and Rahmann, 2012) pipeline to analyze and visualize the functional features of translated transcript isoforms. IsoTV incorporates various tools to predict protein domains, secondary structure, disordered regions and post-translational modification sites. The visualization facilitates comparison of the set of transcript isoforms expressed across different conditions and helps to explore functional consequences of isoform differences. IsoTV supports a range of transcriptome sequencing technologies, including short-reads, and Oxford Nanopore (ONT) and Pacific Biosciences long-reads. In addition, the pipeline is able to process raw ONT long-reads to *de novo* assemble a transcriptome and quantify isoform expression. We demonstrate the functionality of IsoTV on cancer cell lines sequenced using ONT long-reads (The Singapore Nanopore Expression Consortium, 2020).

2 Materials and methods

IsoTV's modular architecture allows visualization on a range of inputs containing at least a transcriptome FASTA file and the corresponding annotation GTF file, and a list of genes. If a file containing isoform expression is provided, gene and isoform expression can be compared across conditions.

IsoTV can process raw or basecalled ONT reads to *de novo* assemble the transcriptome. This sub-workflow is inspired by ONT's long-read processing pipeline (Nanopore Technologies, 2019), with the following changes. If raw ONT signals are provided, the reads are basecalled with Guppy, and low quality reads are discarded using Filtlong (Wick, 2017). The transcriptome is then comprehensively defined for all samples.

Finally, Gffcompare (Pertea and Pertea, 2020) generates consensus sequences for all isoforms and removes redundant transcripts. After assembling the transcriptome, filtered reads are mapped (Li, 2018), quantified and normalized (Love *et al.*, 2014). Supplementary Figure S1 illustrates the scheme of the pipeline.

An additional utility of IsoTV is its isoform translation approach. One obstacle of directly translating mRNA sequences is the presence of upstream open reading frames (uORFs), which are non-coding ORFs located upstream of the coding region (Zhang *et al.*, 2019). Existing tools do not account for uORFs and rely on ribosome profiling data. We developed an algorithm that addresses both issues. Potential start codon sequences are scored using the position weight matrix of the

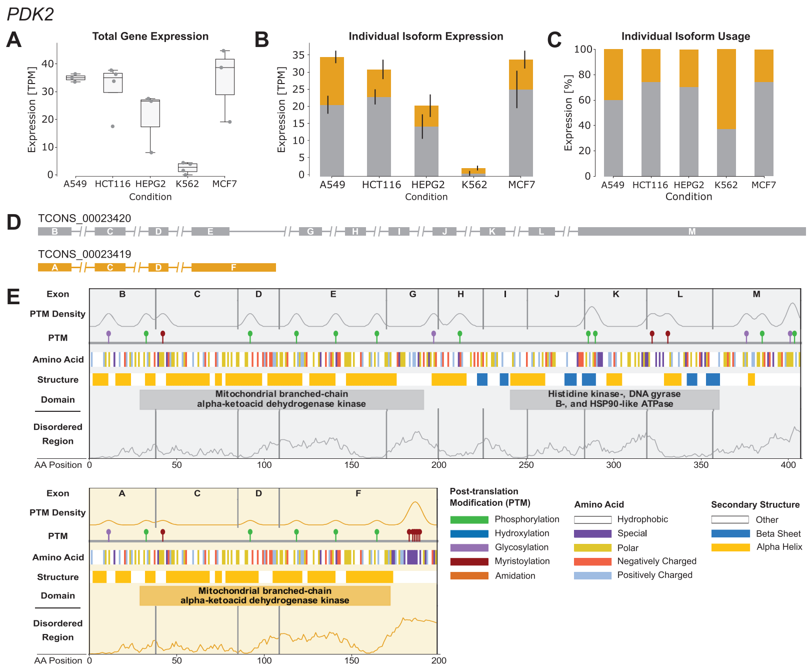

Fig. 1. IsoTV visualization of translated transcript isoforms for *PDK2* in five human cancer cell lines using ONT data from the Singapore Nanopore Expression Project. For all sections, each isoform is consistently color-coded. (**A**) Total gene expression with each individual replicate plotted. (**B**) Individual isoform expression. (**C**) Isoform usage normalized to total gene expression. (**D**) Transcript composition. The lines represent introns, while the rectangles designate exons. Exons with identical 5′ and 3′ boundaries are annotated with the same identifier. (**E**) Feature plots of translated isoforms

motif for initiation of protein translation (Kozak, 1987), and the amino acid chain length is stored. The start sites' score and their respective amino acid chain length are compared, and the most probable ORF determined from these criteria is translated.

The IsoTV visualization consists of three major sections depicting various facets of the analysis (Fig. 1). The first section illustrates the gene and isoform expression levels across all conditions (Fig. 1A–C). The second visualizes the intron-exon structure of all isoforms (Fig. 1D). The third summarizes the features of all translated isoforms, with each isoform having its own dedicated panel consisting of subsections that can be individually included (Fig. 1E). These sections and tools (de Castro *et al.*, 2006; Cock *et al.*, 2013; Erdős and Dosztányi, 2020; Torrisi*et al.*, 2019) are further described in the IsoTV documentation.

3 Results

In order to evaluate the pipeline, we chose *ADAM15* because of its known increased expression in breast and lung cancers (Schütz*et al.*, 2005) and well characterized alternative splicing events generating distinct transcript isoforms (Kleino*et al.*, 2007). A detailed description of this case study is given in the Supplementary Material. Briefly, using basecalled ONT direct cDNA data for five human cancer cell lines (The Singapore Nanopore Expression Consortium, 2020), the annotation and expression plots from the visualization showed increased expression and usage for Isoform 3 that skips Exon W in breast and lung cancer cells (Supplementary Fig. S2A–D). Moreover, the feature plot for Isoform 3 lacks a proline-rich sequence and a disordered region compared with other isoforms (Supplementary Fig. S2E). This case study demonstrates that IsoTV is able to process ONT long-reads, identify transcript isoforms and characterize functional features of translated isoforms with intuitive visualizations.

Acknowledgements

The authors thank Julia di Iulio and Annkatrin Bressin for comments on the manuscript, and Olga Jasnovidova for advice. They also thank Jonathan Goeke for the generous permission of using the Singapore Nanopore Expression Project data prior to publication.

Funding

This work was supported by the Max Planck Society (to A.M.). S.A. was supported by a Fulbright Fellowship Grant.

Conflict of Interest: none declared.

References

de Castro, E. *et al.* (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.*, **34**, W362–5.

Cock, P.J.A. *et al.* (2013) Galaxy tools and workflows for sequence analysis with applications in molecular plant pathology. *PeerJ*, **1**, e167.

Erdős, G. and Dosztányi, Z. (2020) Analyzing protein disorder with IUPred2A. *Curr. Protoc. Bioinf.*, **70**, e99.

Kleino, I. *et al.* (2007) *ADAM15* gene structure and differential alternative exon use in human tissues. *BMC Mol. Biol.*, **8**, 90.

Koster,J., and Rahmann,S. (2012) Snakemake–a scalable bioinformatics workflow engine. Bioinformatics, **28**, 2520–2522.

Kozak, M. (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.*, **15**, 8125–8148.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

Love, M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Nanopore Technologies. (2019) ont-tutorial-pinfish. https://github.com/nanoporetech/ont_tutorial_pinfish, (29 November 2019, date last accessed).

Pertea, G. and Pertea, M. (2020) GFF utilities: GffRead and GffCompare. *F1000Res*, **9**, 304.

Sakamoto, Y. *et al.* (2020) A new era of long-read sequencing for cancer genomics. *J. Hum. Genet.*, **65**, 3–10.

Schütz, A. *et al.* (2005) Expression of *ADAM15* in lung carcinomas. *Virchows Arch.*, **446**, 421–429.

The Singapore Nanopore Expression Consortium (2020) The Singapore Nanopore Expression Project (SG-NEx) data pre-release v0.1. Zenodo. 10.5281/zenodo.4159715.

Torrisi, M. *et al.* (2019) Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. *Sci. Rep.*, **9**, 12374.

Ule, J. and Blencowe, B.J. (2019) Alternative splicing regulatory networks: functions, mechanisms, and evolution. *Mol. Cell*, **76**, 329–345.

Wick, R. (2017) Filtlong. https://github.com/rrwick/Filtlong, (29 November 2019, date last accessed).

Zhang, H. *et al.* (2019) Function and evolution of upstream ORFs in eukaryotes. *Trends Biochem. Sci.*, **44**, 782–794.