Genome analysis

**ScisorWiz: visualizing differential isoform expression in single-cell long-read data**

Alexander N. Stein 1,2, Anoushka Joglekar 1,2, Chi-Lam Poon 1,2 and Hagen U. Tilgner 1,2,*

1Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY 10065, USA and 2Center for Neurogenetics, Weill Cornell Medicine, New York, NY 10065, USA

*To whom correspondence should be addressed.

Associate Editor: Peter Robinson

Received on December 23, 2021; revised on April 11, 2022; editorial decision on May 9, 2022; accepted on May 18, 2022

Abstract

Summary: RNA isoforms contribute to the diverse functionality of the proteins they encode within the cell. Visualizing how isoform expression differs across cell types and brain regions can inform our understanding of disease and gain or loss of functionality caused by alternative splicing with potential negative impacts. However, the extent to which this occurs in specific cell types and brain regions is largely unknown. This is the kind of information that ScisorWiz plots can provide in an informative and easily communicable manner. ScisorWiz affords its user the opportunity to visualize specific genes across any number of cell types, and provides various sorting options for the user to gain different ways to understand their data. ScisorWiz provides a clear picture of differential isoform expression through various clustering methods and highlights features such as alternative exons and single-nucleotide variants. Tools like ScisorWiz are key for interpreting single-cell isoform sequencing data. This tool applies to any single-cell long-read RNA sequencing data in any cell type, tissue or species.

Availability and implementation: Source code is available at http://github.com/ans4013/ScisorWiz. No new data were generated for this publication. Data used to generate figures was sourced from GEO accession token GSE158450 and available on GitHub as example data.

Contact: hut2006@med.cornell.edu

1 Introduction

Differential isoform expression between cell types and across conditions plays a major role in the diversification of the proteome (Nilsen and Graveley, 2010) and functionality of transcripts in the cell (Yang et al., 2016). Long-read sequencing has become widely used to address this problem (Au et al., 2013; Bolsery et al., 2015; Koren et al., 2012; Leung et al., 2021; Oikonomopoulou et al., 2016; Ruiz-Reche et al., 2019; Schulz et al., 2021; Sharon et al., 2013; Tilgner et al., 2015), and with applications to single-cell isoform sequencing studies (Arzalluz-Luque et al., 2022; Gupta et al., 2018; Hardwick et al., 2022; Joglekar et al., 2021; Volden and Vollmers, 2022). These approaches have been reviewed in Hardwick et al. (2019). Such data require informative visualizations for single genes, so that the impact of alternative exons, exon combinations, as well as those of transcription start site (TSS) and PolyA sites can be easily appreciated. Here, we present ScisorWiz, a streamlined tool to visualize isoform expression differences across single-cell clusters in an informative and easily communicable manner. ScisorWiz achieves this with an easy, fast and reliable method of visualizing differential isoform expression data across multiple clusters and is executable from the command line with the R language (R Core Team, 2018).

2 Usage

ScisorWiz visualizes pre-processed single-cell long-read RNA sequencing data. For a user-specified gene, reads for any number of cell types can be visualized and are clustered by chain of introns (the ordered list of a read’s introns), TSS and/or PolyA site for each cell type. We have used such plots in our long-read (Sharon et al., 2013; Tilgner et al., 2014, 2015) and single-cell long-read publications (Gupta et al., 2018; Hardwick et al., 2022; Joglekar et al., 2021). However, customizing such a plot for publication standards includes read mapping, shrinking of introns and recalculation of coordinates, calculation of alternative exons, adjusting plot area depending on number of reads and cell types, as well as plotting single-nucleotide variants (SNVs), insertions and deletions. This process was previously not automated and was only intended to be used for publication purposes. Now, ScisorWiz does this with a single command in R, allowing for many user-specified options including exploratory, interactive outputs and multiple ways to sort isoforms within each cell type: namely by intron chain, TSS, PolyA site, as well as all three combined.

ScisorWiz can be run on output generated by scisorseqr (Joglekar et al., 2021) or a similarly formatted dataset, which, in turn, can be...
based on diverse mappers including STAR (Dobin et al., 2013) and 
minimap2 (Li, 2018). The first approach uses GFF-files for mappings and 
read-to-gene assignment files that are generated automatically by 
scisorseq. However, the user is free to generate these standardized 
files by other means. The second method uses more specific files that 
are assigned to scisorseq—the file in question already contains an 
assigned gene, TSS, PolyA sites and the intron and exon-mappings for 
each read. Thus, this gene plotting library communicates intimately 
with scisorseq. Additionally, through theMismatchFinder function, 
the dataset in question can be compared against the reference genome 
to determine the locations of SNVs, insertions and deletions to be 
visualized in the plot.

3 Approach

to visualize exons separated by up to ~100-fold larger introns, each 
purely intronic region is shrunk to 100 bases, while sequences that have 
annotated or novel exons are displayed with their real size. A drawback 
of this approach is that short introns (<1 kb) that are fully retained in a 
long read will be drawn to scale. However, very large introns (>10 kb), 
for which long reads are unlikely to represent the retained form will be 
shrunk to 100 bases. By default, the package clusters read according to 
intron chains. Reads with identical intron chains are thus displayed to-
gether to form exonic blocks. Alternatively, clustering can take into ac-
count any combination of TSS, PolyA sites and intron chains when 
using scisorseq-generated files as input. In this situation, only reads 
with an assigned TSS and/or PolyA site are plotted.

ScisorWiz provides a clear picture of differential isoform expres-
sion of genes in any dataset by clustering reads. This reveals differ-
ential patterns more clearly, such as alternate exon expression across 
and within cell types.

4 Output

ScisorWiz’s output visualizes isoforms read-by-read for any number of 
cell types for any user-specified gene. Figure 1 shows Snap25 gene 
isoforms across six cell types. Colored boxes are exons per read. For 
each cell type, reads are ordered by intron chain. Orange exons indi-
cate alternatively spliced exons, defined as being included in at least 
5% and at most 95% of overlapping reads taken from the entire data-
set irrespective of cell type—this range is also user-specified. Consistent with previous observations (Joglekar et al., 2021; 
Johansson et al., 2008), we find that two neighboring alternative 
exons in Snap25 are mutually exclusive. Importantly, we observe this 
mutual exclusivity to be present in multiple cell types. For higher error 
rates such as currently in Oxford Nanopore, 20% and 80% cutoffs 
provide a clearer picture of alternative exons. There are multicolored 
dots among the cell types representing the locations of SNVs, inser-
tions and deletions. By default, only SNVs, insertions and deletions 
present in at least 5% and at most 95% of overlapping reads are high-
lighted in order to avoid plotting random sequencing errors. However, 
these cutoffs can be adjusted as options by the user allowing 
the visualization of every single-nucleotide disagreeing with the refer-
ence genome, should this be of interest. This course of action may be 
useful in low error-rate sequencing such as Pacific Biosciences (Eid et al., 2009). Similarly, any mismatches present within the first or last 
20 bases of an alignment are not shown in order to avoid alignment 
artifacts at alignment ends. The bottom section is the GENCODE an-
notation covered by long reads. ScisorWiz also generates a file for all 
single-cell long reads that can be uploaded and inspected on the 
UCSC Genome Browser (Kent, 2002).

Acknowledgements

We thank the Weill Cornell Medicine Scientific Computing Unit (SCU) for 
use of their computational resources.

Funding

This work was supported by the NIGMS [grant number 1R01GM135247-01].

Conflict of Interest: none declared.

References

Arazallu-Luque, A. et al. (2022) ACORDE unravels functionally interpretable 
networks of isoform co-usage from single cell data. Nat. Commun., 13, 1828.

Au,K.F. et al. (2013) Characterization of the human ESC transcriptome by hy-
brid sequencing. Proc. Natl. Acad. Sci. USA, 110, E4821–E4830.

Bolisetty, M.T. et al. (2015) Determining exon connectivity in complex 
mRNAs by nanopore sequencing. Genome Biol., 16, 204.

Dobin,A. et al. (2013) Star: ultrafast universal RNA-seq aligner. Bioinformatics, 
29, 15–21.

Edel, J. et al. (2009) Real-time DNA sequencing from single polymerase mole-
cules. Science, 323, 133–138.

Gupta,I. et al. (2018) Single-cell isoform RNA sequencing characterizes iso-
forms in thousands of cerebellar cells. Nat. Biotechnol., 36, 1197–1202.

Hardwick,S.A. et al. (2019) Getting the entire message: progress in isoform 
sequencing. Front. Genet., 10, 709.

Hardwick,S.A. et al. (2022) Single-molecule isoform RNA sequencing unlocks 
barcoded exon connectivity in frozen brain tissue. Nat. Biotechnol., [Epub 
ahead of print].

Joglekar,A. et al. (2021) A spatially resolved brain region- and cell type-specific 
isoform atlas of the postnatal mouse brain. Nat. Commun., 12, 463.

Johansson,J.U. et al. (2008) An ancient duplication of exon 5 in the Snap25 
gene is required for complex neuronal development/function. PLoS Genet., 
4, e1000278.

Kent,W.J. et al. (2002) The human genome browser at UCSC. Genome Res., 
12, 996–1006.

Koren,S. et al. (2012) Hybrid error correction and de novo assembly of 
single-molecule sequencing reads. Nat. Biotechnol., 30, 693–700.

Leung,S.K. et al. (2021) Full-length transcript sequencing of human and mouse 
cerebral cortex identifies widespread isoform diversity and alternative splic-
ing. Cell Rep., 37, 110022.

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

Nilsen,T.W. and Grønlie,B.R. (2015) Expansion of the eukaryotic proteome 
by alternative splicing. Nature, 463, 457–463.

Oikonomopoulos,S. et al. (2016) Benchmarking of the Oxford Nanopore min-
ion sequencing for quantitative and qualitative assessment of cDNA popula-
tions. Sci. Rep., 6, 31602.

Fig. 1. The isoforms of the Snap25 gene present in each read of a specific cell type 
are displayed one above the other to form a consistent picture of the gene expression 
of each cell type. The orange-colored exon represents an exon which is considered 
alternative as a result of a Y value of 5–95% inclusion irrespective of cell type. The 
multicolored dots on the plot represent SNVs (blue), insertions (green) and deletions 
(red). All SNVs, insertions and deletions included are in at least 5% and at most 
95% of overlapping reads. The reads at the bottom (black) represent the part of the 
GENCODE annotation for Snap25 (A color version of this figure appears in the 
online version of this article.)

4 Output

ScisorWiz's output visualizes isoforms read-by-read for any number of 
cell types for any user-specified gene. Figure 1 shows Snap25 gene 
isoforms across six cell types. Colored boxes are exons per read. For 
each cell type, reads are ordered by intron chain. Orange exons indi-
cate alternatively spliced exons, defined as being included in at least 
5% and at most 95% of overlapping reads taken from the entire data-
set irrespective of cell type—this range is also user-specified. Consistent with previous observations (Joglekar et al., 2021; 
Johansson et al., 2008), we find that two neighboring alternative 
exons in Snap25 are mutually exclusive. Importantly, we observe this 
mutual exclusivity to be present in multiple cell types. For higher error 
rates such as currently in Oxford Nanopore, 20% and 80% cutoffs 
provide a clearer picture of alternative exons. There are multicolored 
dots among the cell types representing the locations of SNVs, inser-
tions and deletions. By default, only SNVs, insertions and deletions 
present in at least 5% and at most 95% of overlapping reads are high-
lighted in order to avoid plotting random sequencing errors. However, 
these cutoffs can be adjusted as options by the user allowing 
the visualization of every single-nucleotide disagreeing with the refer-
ence genome, should this be of interest. This course of action may be 
useful in low error-rate sequencing such as Pacific Biosciences (Eid et al., 2009). Similarly, any mismatches present within the first or last 
20 bases of an alignment are not shown in order to avoid alignment 
artifacts at alignment ends. The bottom section is the GENCODE an-
notation covered by long reads. ScisorWiz also generates a file for all 
single-cell long reads that can be uploaded and inspected on the 
UCSC Genome Browser (Kent, 2002).
R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruiz-Reche, A. *et al.* (2019) ReorientExpress: reference-free orientation of nanopore cDNA reads with deep learning. *Genome Biol.*, **20**, 260.

Schulz, L. *et al.* (2021) Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts. *Genome Biol.*, **22**, 190.

Sharon, D. *et al.* (2013) A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.*, **31**, 1009–1014.

Tilgner, H. *et al.* (2014) Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. USA*, **111**, 9869–9874.

Tilgner, H. *et al.* (2015) Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.*, **33**, 736–742.

Volden, R. and Vollmers, C. (2022) Single-cell isoform analysis in human immune cells. *Genome Biol.*, **23**, 47.

Yang, X. *et al.* (2016) Widespread expansion of protein interaction capabilities by alternative splicing. *Cell*, **164**, 805–817.