Published online 29 November 2021

https://doi.org/10.1093/nargab/lqab105

RNA-Scoop: interactive visualization of transcripts in single-cell transcriptomes

Maria Stephenson[1,2,†], Ka Ming Nip[1,3,*,†], Saber HafezQorani[1,3], Kristina K. Gagalova[1,3], Chen Yang[1,3], René L. Warren[1] and Inanc Birol[1,4,*]

[1]Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC V5Z 4S6, Canada, [2]Computer Science Co-op Program, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, [3]Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC V5Z 4S6, Canada and [4]Department of Medical Genetics, University of British Columbia, Vancouver, BC V6H 3N1, Canada

Received May 01, 2021; Revised August 21, 2021; Editorial Decision October 11, 2021; Accepted November 26, 2021

ABSTRACT

Recent advances in single-cell RNA sequencing technologies have made detection of transcripts in single cells possible. The level of resolution provided by these technologies can be used to study changes in transcript usage across cell populations and help investigate new biology. Here, we introduce RNA-Scoop, an interactive cell cluster and transcriptome visualization tool to analyze transcript usage across cell categories and clusters. The tool allows users to examine differential transcript expression across clusters and investigate how usage of specific transcript expression mechanisms varies across cell groups.

INTRODUCTION

In eukaryotic genomes, alternative splicing (AS), alternative transcriptional initiation (ATI) and alternative cleavage (AC) enable genes to express multiple transcript isoforms. Resulting isoforms differ in structure and may also have different cell functions. For instance, ATI can act as a regulator of translation (1–3), and some AS events may play important roles in cell differentiation (3–6), development (3, 7, 8) and disease (3, 9–14). Further, while its precise biological roles are still under investigation, differential isoform expression has been found to occur across cell types (15), tissue types (15, 16), bodily regions (15) and even individuals (17).

In recent years, single-cell RNA sequencing (scRNA-seq) technologies have advanced greatly, enabling accurate detection of transcript isoforms in single cells (18–20). In parallel to these advances, a number of scRNA-seq visualization tools have been developed to guide analyses at the gene level, but most of them do not support transcript-level analysis (21). Exceptions are VALERIE (22) and Millefy (23) – two tools that display percent spliced-in (PSI) values and read coverage across small genomic regions. However, the effective identification of patterns in transcript usage across thousands of genes remains an open problem. Unsupervised dimensionality reduction techniques, such as t-SNE (24) and UMAP (25), are typically used to generate two-dimensional embeddings for the visualization of scRNA-seq expression data, allowing researchers to identify clusters of cells with similar expression profiles. These techniques, however, do not present information regarding the differentially expressed genes/transcripts between cell clusters.

Here, we introduce RNA-Scoop, an interactive tool that visualizes transcript usage across single-cell transcriptomes. RNA-Scoop enables easy identification of differentially expressed transcripts across cell groups and transcripts with specific expression patterns, such as isoform switching, co-expression, and category/cluster specific expression. It also displays basic transcript structure, allowing users to examine how usage of specific transcript expression mechanisms, such as AS, ATI and AC vary across different groups of cells. RNA-Scoop is designed to work with data produced from single-cell protocols that support transcript isoform level analysis. In other words, 3′ end capture protocols (such as 10x Genomics or Drop-Seq) designed for measuring gene expression in single cells are not supported.

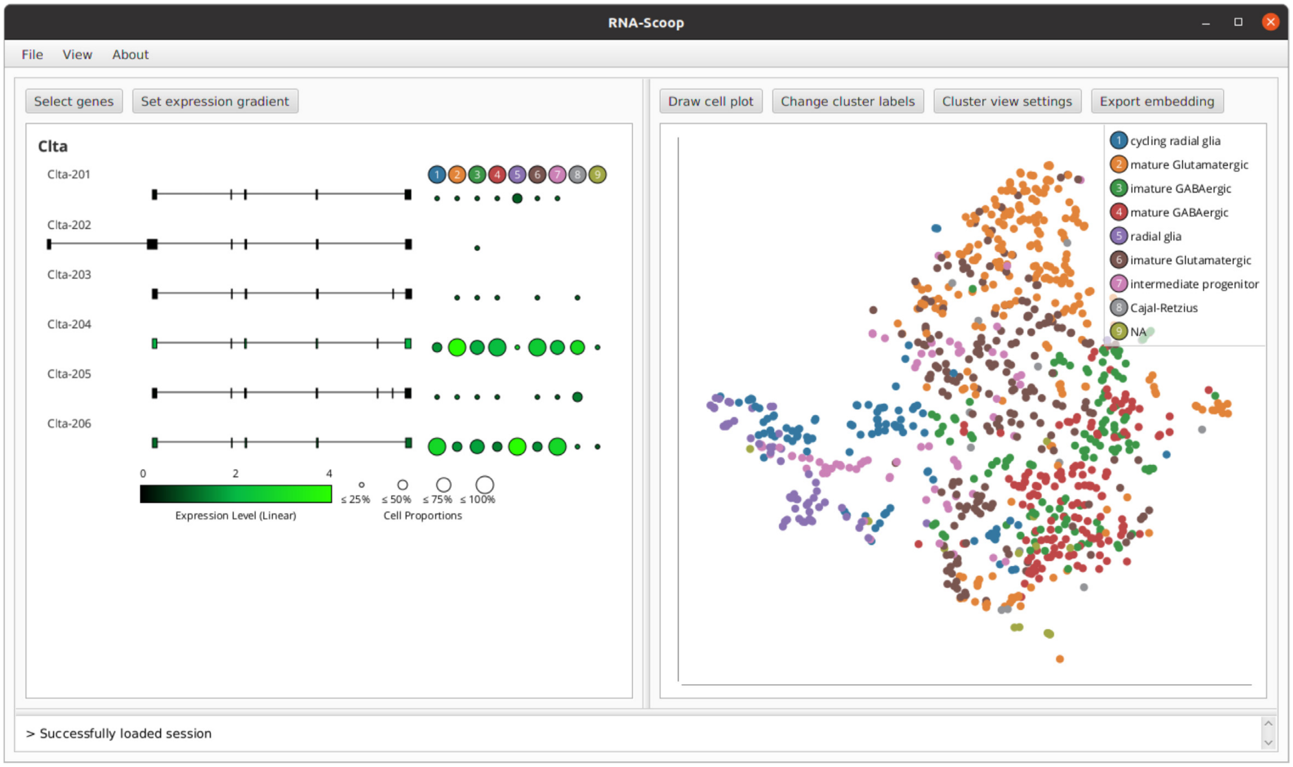
MATERIALS AND METHODS

We utilized two mouse single-cell transcriptomic datasets to illustrate RNA-Scoop's features in visualizing transcripts in scRNA-seq data.

ScNaUmi-seq dataset

Input files for RNA-Scoop were prepared using the transcript expression results from a single-cell Nanopore-UMI

[*]To whom correspondence should be addressed. Tel: +1 604 707 5900 x 675448; Fax: +1 604 876 3561; Email: ibirol@bcgsc.ca
Correspondence may also be addressed to Ka Ming Nip. Email: kmnip@bcgsc.ca
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors



> Successfully loaded session

Figure 1. RNA-Scoop visualization of *Clta* isoform expression across nine cell category labels from the ScNaUmi-seq dataset. The transcript view on the left shows the isoform structures of selected genes and its integrated dot plot shows the proportion and magnitude of transcript expression in each cell category. Transcripts and dots are colored according to their non-zero median expression levels in all cells and cells of their particular category, respectively. Dots are not drawn for transcripts not expressed in any cells of a given category. The cell cluster plot on the right shows the clustering of cells based on transcript expression.

sequencing (ScNaUmi-seq) sample from (20) (Gene Expression Omnibus accession: GSE130708), which consists of 951 brain cells from an E18 mouse. Cell label information was obtained from the study's authors through personal communications. Ensembl release 104 mouse annotation GTF was used to match the transcript IDs within the expression data.

Smart-seq2 dataset

A 6912-cell dataset was compiled using the plate-based Smart-seq2 raw data for 18 mouse tissues from *Tabula Muris* (GEO accession: GSE109774) (26). For each tissue type, 384 cells with at least 500 000 read pairs were arbitrarily selected. Raw reads were trimmed and filtered for adaptor sequences with fastp v0.20.0 (27).

Single-cell transcripts were assembled from the trimmed reads with RNA-Bloom v1.3.1 (28), where cells from the same tissue type were in the same pooled assembly. Assemblies for all tissues were merged with BBTools v38.86 (29). The merged assembly was aligned to the mouse reference genome GRCm38 with minimap2 v2.17 (30), and a GTF file was generated from the alignment results with our script 'make_gtf.py'. Gene labels from Ensembl release 99 were added to the GTF using our script 'annotate_gtf.py'. This GTF file was combined with the Ensembl annotation and then filtered with gffread v0.12.3 (31).

Single-cell transcript expression levels were quantified and merged with Salmon v1.3.0 (32). An expression level matrix of TPM values was generated with our script 'make_matrix.py'. The GTF file was further filtered using our script 'filter_gtf.py' to only contain transcripts in the generated expression level matrix. Our scripts are included within the RNA-Scoop package, and the results presented are generated using release version 1.0.0. See Supplemental Methods S1 for exact commands and runtime parameters. After removing transcripts with low expression levels, the final dataset contained 14 444 genes and 90 195 transcripts.

Software overview

RNA-Scoop is implemented as a Java graphical user interface using the JavaFX package and the following external packages: T-SNE-Java, Java UMAP, JFreeChart - Future State Edition, JSON-Java and ControlsFX (see version numbers in Supplemental Methods S4). The graphical user interface of RNA-Scoop consists of two interactive main panels: a transcript view, which includes a dot plot representation of transcript expression levels per cell category, and a cell cluster plot (Figure 1). Users can save the current state of an RNA-Scoop session to a file, for future use or to share with other users.

The input for RNA-Scoop consists of four files: (i) a Gene Transfer Format (GTF) file containing the transcript annotation, (ii) a cell-by-transcript expression level matrix,

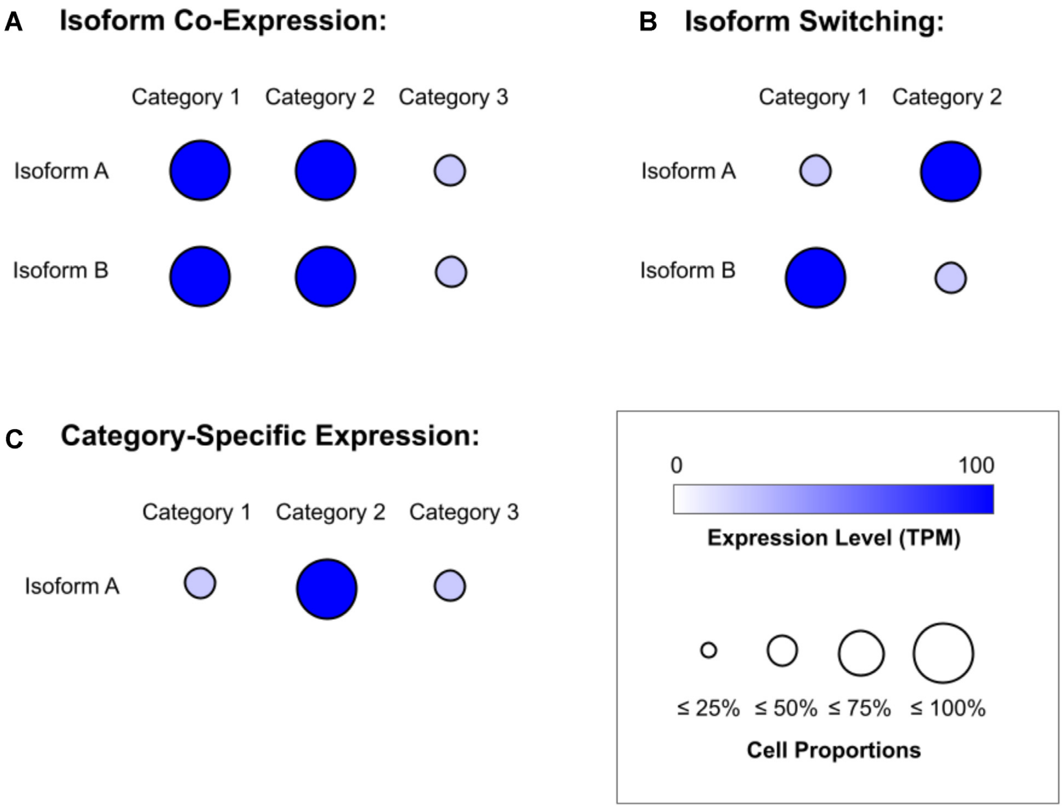


Figure 2. Dot plot enables the visualization of isoform co-expression, isoform switching, and category-specific expression. (**A**) Isoform co-expression refers to cases where two isoforms are both highly or lowly expressed across cell categories. (**B**) Isoform switching between two isoforms can be recognized based on the mutual exclusivity of isoform expression in two or more cell categories. When one isoform is highly expressed, the other isoform is lowly expressed, and vice versa. (**C**) Category-specific isoform expression occurs when the isoform expression is observed in only one (or very few) cell category (or categories).

(iii) a file containing the transcript labels for the columns of the matrix and (iv) file(s) containing cell category labels (e.g. tissue type) for the rows of the matrix. These files can be either selected individually or summarized in a JavaScript Object Notation (JSON) file, where its format specification is described in Supplemental Methods S5. All input files are in popular plain-text file formats that can accommodate data generated from a wide variety of analysis pipelines and sequencing platforms. For example, the GTF file can originate from publicly available transcript annotations, RNA-seq assemblies, or a combination of the two.

Transcript visualization

The transcript view (Figure 1) displays all isoforms of selected genes, allowing users to analyze expression at the transcript level. Since transcripts are visualized as separate entities, users can examine co-expression of isoforms from the same gene, a function that, to the best of our knowledge, is not available in other tools (21–23). Additionally, genes can be filtered specifically for isoform switching events, differential isoform expression across cell categories, and category-specific isoform expression (Supplemental Methods S3). Genes can also be sorted by maximum fold change across all cell categories, allowing transcripts with the highest magnitude of differential expression to be easily identified. As new category labels can be assigned via cell selection, this enables users to easily identify key changes in transcript expression across both cell categories and clusters.

Transcript representation is similar to that used in UCSC Genome Browser (33) and Integrative Genomics Viewer (34), where boxes represent exons and lines represent spliced introns; however, no distinction is made between translated and untranslated regions. By default, transcripts are drawn with respect to their exons' chromosomal coordinates. For datasets produced through strand-specific sequencing protocols, and in cases where gene orientation is known, transcripts on the negative strand can be reverse-complemented and displayed in their 5′-3′orientation. Transcripts are colored according to their median (default) or average expression levels across the cells displayed within the cell cluster plot.

Cell category expression visualization

To compare transcript expression across cell categories, a dot plot (Figure 1) is used to visualize the median or average transcript expression levels (as indicated by the color of the dot) and the proportion of cells which express the transcript (as indicated by dot size) within each cell category. This allows users to easily identify transcript expression patterns, such as co-expression, isoform switching, differential

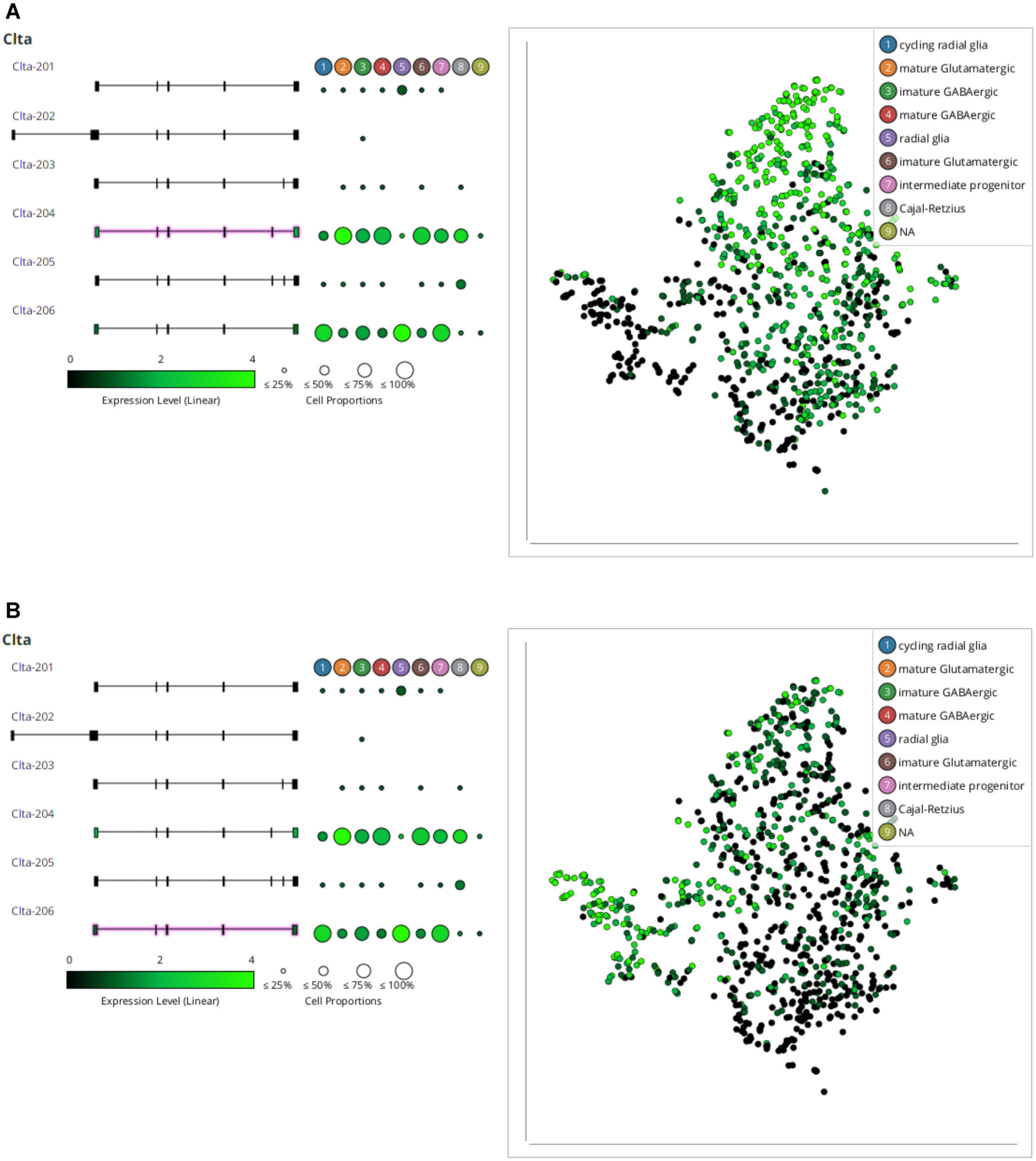


Figure 3. Isoform switching between *Clta-204* and *Clta-206* in the ScNaUmi-seq dataset visualized in RNA-Scoop. Cells in the cell cluster plot are colored according to the selected (highlighted in magenta) isoform's expression level in the cell instead of cell category label colors. (**A**) *Clta-204* isoform expression visualization. (**B**) *Clta-206* isoform expression visualization. The dot plots in A and B are identical, except their coloring.

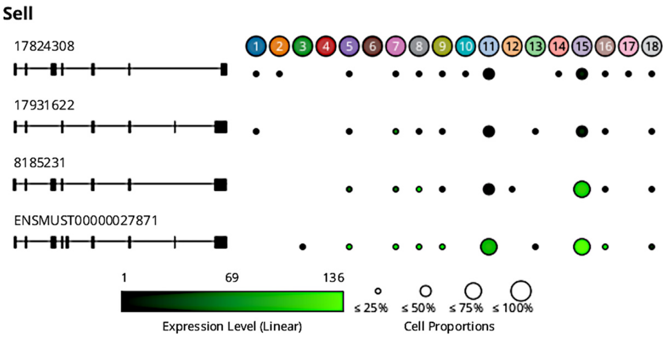


Figure 4. Expression of *Sell* isoforms in 18 mouse tissues. Spleen and marrow are columns 11 and 15 in the dot plot.

expression, and category/cluster specific expression, across all groups of cells. Figure 2 shows examples of how these events are visualized in the dot plot.

Cell cluster visualization

The cell cluster plot (Figure 1) presents two-dimensional embeddings of the expression level matrix. UMAP and t-SNE embeddings can be generated in RNA-Scoop, with tunable parameters for each algorithm. Alternatively, a wider variety of embedding or dimensionality reduction representations can be imported from a user-specified file. The plot is interactive, allowing users to zoom in and out of different areas, and pan the plot freely. Cells are colored based on their category labels, and they can be selected via free-hand lasso selection or by category. Additionally, users can load multiple cell category label sets and create custom label sets in RNA-Scoop via cell selection. This facilitates the examination of transcript expression across clusters and different sets of cell categories, such as tissue or cell type.

When cells are selected, the transcripts are colored according to their expression levels in the selected cells only. As cells can be selected either by free-hand lasso selection or by category, this allows transcript expression to be compared across any groups of cells. When transcripts are selected, the cells in which they are expressed are highlighted.

Cells can be colored based on their expression of the selected transcript instead of cell category (example shown in Figure 3). This feature allows users to examine cell heterogeneity in transcript expression. The expression for the selected transcript (highlighted in magenta in the isoform view) is visualized at three levels: the entire dataset, cell categories, and individual cells. First, exons of a given transcript in the transcript view are colored based on the transcript's median (default) or average expression level over all cells. Second, the dot plot displays the median or average transcript expression level and the proportion of cells that express the transcript, for each cell category. Third, each cell in the cell cluster plot is colored based on their level of expression of the selected transcript (instead of cell category label).

RESULTS AND DISCUSSION

Recapitulating results from a published study

Isoform switching between *Clta-204* and *Clta-206* was previously reported in the ScNaUmi-seq dataset. We recapitulate this finding in RNA-Scoop and the resulting visualizations are shown in Figure 3.

Isoform switching was visualized between individual cells using the cell cluster plot. Cells that were green when *Clta-204* was selected (Figure 3A) have become black when *Clta-206* was selected (Figure 3B). Similarly, cells that were black when *Clta-204* was selected have become green when *Clta-206* was selected. Isoform switching was also visualized between cell categories using the dot plot. Indicated by larger and brighter green dots, *Clta-204* was expressed higher or in a larger proportion in cell categories 2, 4, 6 and 8, while *Clta-206* was expressed higher or in a larger proportion in cell categories 1, 5 and 7.

Exploring the mouse single-cell transcriptomic atlas

We also explore *Tabula Muris*, a mouse single-cell transcriptomic atlas, where analyses in the original study were only done based on gene level expression. Filtering within RNA-Scoop (see Supplemental Methods S2 for details) finds a total of 251 genes that exhibit differential isoform expression between spleen and marrow. An interesting example is the *Sell* gene, where four isoforms were shown to be co-expressed in both spleen and marrow (Figure 4). Although the isoform structure of *Sell* transcripts differed due to alternative transcription mechanisms, such as AS and ATI, RNA-Scoop was able to clearly show that all four isoforms followed similar expression patterns across different tissues.

In the cell cluster plot, each tissue and cell type formed several distinct clusters based on transcript expression levels. An excellent example is the microglia cells, which formed three clusters, separate from the other tissues (Figure 5). Several genes exhibited differential isoform expression between the clusters, including *Tmem119,* in which two isoforms were shown to undergo isoform switching. Isoforms *2284771* and *2628717* underwent switching between microglia clusters 1 and 3. Isoforms *2184694* and *28292321* appear to be co-expressed in all three microglia clusters and lowly expressed in the non-microglia cluster. These results cannot be examined through use of other transcriptome visualization tools, as they do not support the identification of cell subpopulations, nor identification of the transcripts differentially expressed across them.

While other scRNA-seq visualization tools, such as VALERIE and Millefy, require users to identify genes of interest through other methods, RNA-Scoop can help users identify subsets of genes exhibiting differential isoform expression across tissue and cell types. With RNA-Scoop, users can identify and examine transcripts undergoing specific expression patterns, such as co-expression, isoform switching, differential expression, and category-specific expression. Additionally, RNA-Scoop visualizes basic transcript structures, allowing examination of AS, ATI, and AC usage across cell groups. Through these easy-to-use features, RNA-Scoop is an effective tool for interactive anal-

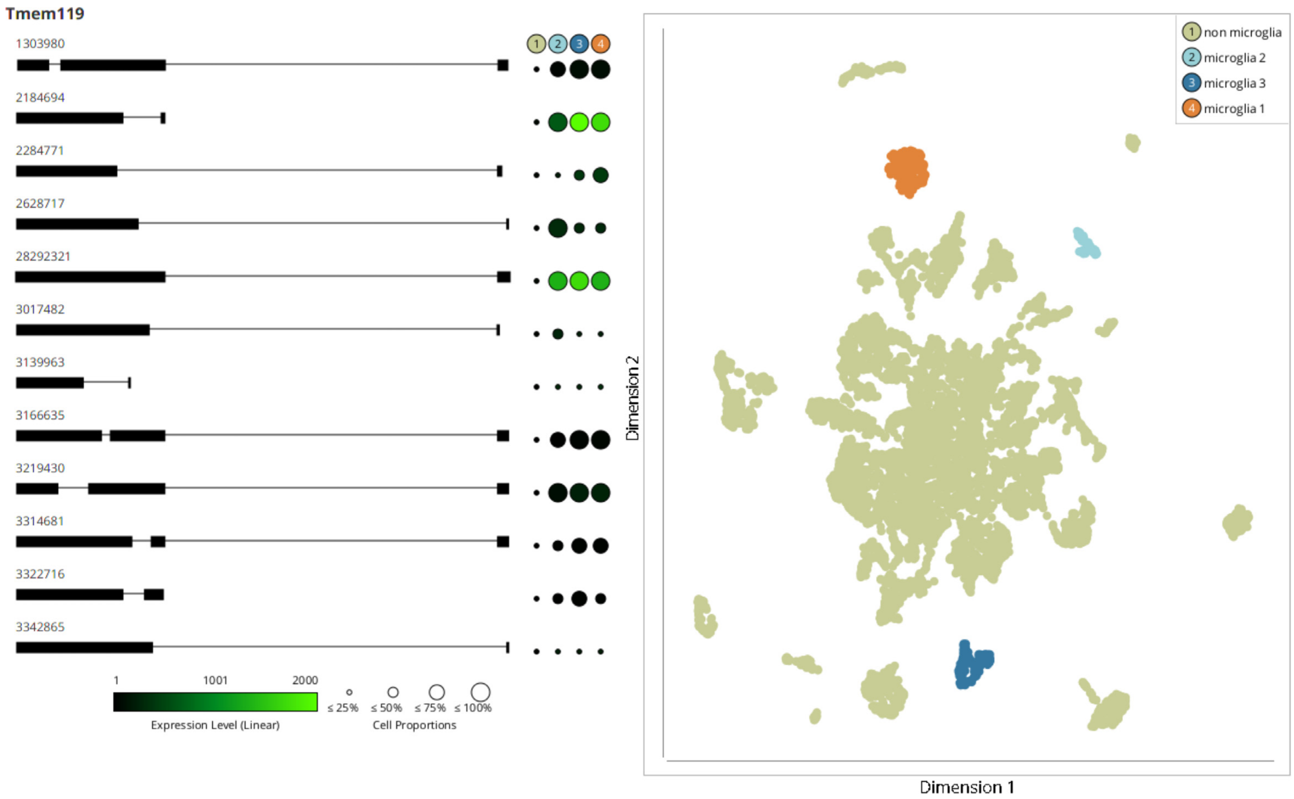


Figure 5. RNA-Scoop visualization of *Tmem119* isoform expression across three microglia cell clusters. Isoforms are colored according to their average expression levels across all cells, and dots are colored according to their average non-zero expression levels in the cells of their category.

ysis of transcript expression across and within cell populations.

DATA AVAILABILITY

The release version 1.0.1 of RNA-Scoop can run on any operating system that has Java Runtime Environment 8. The source and compiled Java archive of RNA-Scoop are available at https://github.com/bcgsc/RNA-Scoop under open source license GPL-3.0. The input data files for the ScNaUmi-seq and Smart-Seq2 datasets used in this manuscript are publicly available at: https://github.com/bcgsc/RNA-Scoop/blob/master/test_data/GSM3748089.zip and https://www.bcgsc.ca/downloads/supplementary/rnascoop/rnascoop_test_data.tar.gz, respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

Author contributions: K.M.N., I.B. and M.S. designed the method. M.S. implemented the software under the guidance of K.M.N. S.H., C.Y. and K.K.G. tested the software prototype and provided feedback for improvements. All authors analyzed the results and wrote the manuscript. The content of this work is solely the responsibility of the authors, and does not necessarily represent the official views of the National Institutes of Health or other funding organizations.
The authors would like to thank Dr. Veronique LeBlanc and Dr. Elizabeth Chun at Canada's Michael Smith Genome Sciences Centre for their insightful discussions on our method. The authors would like to thank Dr. Kevin Lebrigand for providing the cell label information of the ScNaUmi-seq dataset.

FUNDING

Genome Canada/Genome British Columbia [281ANV]; National Human Genome Research Institute of the National Institutes of Health [R01HG007182].
Conflict of interest statement. None declared

REFERENCES

1. Wang,X., Hou,J., Quedenau,C. and Chen,W. (2016) Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. *Mol. Syst. Biol.*, **12**, 875.
2. Kurihara,Y., Makita,Y., Kawashima,M., Fujita,T., Iwasaki,S. and Matsui,M. (2018) Transcripts from downstream alternative transcription start sites evade uORF-mediated inhibition of gene expression in arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 7831–7836.
3. de la Fuente,L., Arzalluz-Luque,Á., Tardáguila,M., Del Risco,H., Martí,C., Tarazona,S., Salguero,P., Scott,R., Lerma,A., Alastrue-Agudo,A. *et al.* (2020) tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. *Genome Biol.*, **21**, 119.

4. Furlanis,E. and Scheiffele,P. (2018) Regulation of neuronal differentiation, function, and plasticity by alternative splicing. *Annu. Rev. Cell Dev. Biol.*, **34**, 451–469.
5. Li,H., Cheng,Y., Wu,W., Liu,Y., Wei,N., Feng,X., Xie,Z. and Feng,Y. (2014) SRSF10 regulates alternative splicing and is required for adipocyte differentiation. *Mol. Cell. Biol.*, **34**, 2198–2207.
6. Sen,S., Jumaa,H. and Webster,N.J.G. (2013) Splicing factor SRSF3 is crucial for hepatocyte differentiation and metabolic function. *Nat. Commun.*, **4**, 1336.
7. Baralle,F.E. and Giudice,J. (2017) Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.*, **18**, 437–451.
8. Weyn-Vanhentenryck,S.M., Feng,H., Ustianenko,D., Duffié,R., Yan,Q., Jacko,M., Martinez,J.C., Goodwin,M., Zhang,X., Hengst,U. *et al.* (2018) Precise temporal regulation of alternative splicing during neural development. *Nat. Commun.*, **9**, 2189.
9. Paronetto,M.P., Passacantilli,I. and Sette,C. (2016) Alternative splicing and cell survival: from tissue homeostasis to disease. *Cell Death Differ.*, **23**, 1919–1929.
10. Scotti,M.M. and Swanson,M.S. (2016) RNA mis-splicing in disease. *Nat. Rev. Genet.*, **17**, 19–32.
11. Daguenet,E., Dujardin,G. and Valcárcel,J. (2015) The pathogenicity of splicing defects: mechanistic insights into pre-mRNA processing inform novel therapeutic approaches. *EMBO Rep.*, **16**, 1640–1655.
12. Fu,R.-H., Liu,S.-P., Huang,S.-J., Chen,H.-J., Chen,P.-R., Lin,Y.-H., Ho,Y.-C., Chang,W.-L., Tsai,C.-H., Shyu,W.-C. *et al.* (2013) Aberrant alternative splicing events in parkinson's disease. *Cell Transplant.*, **22**, 653–661.
13. Anthony,K. and Gallo,J.-M. (2010) Aberrant RNA processing events in neurological disorders. *Brain Res.*, **1338**, 67–77.
14. Cieply,B. and Carstens,R.P. (2015) Functional roles of alternative splicing factors in human disease. *Wiley Interdiscipl. Rev.: RNA*, **6**, 311–326.
15. Joglekar,A., Prjibelski,A., Mahfouz,A., Collier,P., Lin,S., Schlusche,A.K., Marrocco,J., Williams,S.R., Haase,B., Hayes,A. *et al.* (2021) A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nat. Commun.*, **12**, 463.
16. Noh,S.-J., Lee,K., Paik,H. and Hur,C.-G. (2006) TISA: tissue-specific alternative splicing in human and mouse genes. *DNA Res.*, **13**, 229–243.
17. Kwan,T., Benovoy,D., Dias,C., Gurd,S., Serre,D., Zuzan,H., Clark,T.A., Schweitzer,A., Staples,M.K., Wang,H. *et al.* (2007) Heritability of alternative splicing in the human genome. *Genome Res.*, **17**, 1210–1218.
18. Hagemann-Jensen,M., Ziegenhain,C., Chen,P., Ramsköld,D., Hendriks,G.-J., Larsson,A.J.M., Faridani,O.R. and Sandberg,R. (2020) Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.*, **38**, 708–714.
19. Gupta,I., Collier,P.G., Haase,B., Mahfouz,A., Joglekar,A., Floyd,T., Koopmans,F., Barres,B., Smit,A.B., Sloan,S.A. *et al.* (2018) Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.*, **36**, 1197–1202.
20. Lebrigand,K., Magnone,V., Barbry,P. and Waldmann,R. (2020) High throughput error corrected nanopore single cell transcriptome sequencing. *Nat. Commun.*, **11**, 4025.
21. Cakir,B., Prete,M., Huang,N., van Dongen,S., Pir,P. and Kiselev,V.Y. (2020) Comparison of visualization tools for single-cell RNAseq data. *NAR Genom Bioinform*, **2**, lqaa052.
22. Wen,W.X., Mead,A.J. and Thongjuea,S. (2020) VALERIE: visual-based inspection of alternative splicing events at single-cell resolution. *PLoS Comput. Biol.*, **16**, e1008195.
23. Ozaki,H., Hayashi,T., Umeda,M. and Nikaido,I. (2020) Millefy: visualizing cell-to-cell heterogeneity in read coverage of single-cell RNA sequencing datasets. *BMC Genomics*, **21**, 177.
24. van der Maaten,L. (2014) Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.*, **15**, 3221–3245.
25. McInnes,L., Healy,J. and Melville,J. (2018) UMAP: uniform manifold approximation and projection for dimension reduction. *JOSS* , **3**, 861.
26. Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group and Principal investigators (2018) Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, **562**, 367–372.
27. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
28. Nip,K.M., Chiu,R., Yang,C., Chu,J., Mohamadi,H., Warren,R.L. and Birol,I. (2020) RNA-Bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes. *Genome Res.*, **30**, 1191–1200.
29. Bushnell,B. (2021) *BBTools*. https://jgi.doe.gov/data-and-tools/bbtools, Last accessed: 14 Oct 2021.
30. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
31. Pertea,G. and Pertea,M. (2020) GFF utilities: gffread and gffcompare. *F1000Res.*, **9**, 304.
32. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.
33. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,a. D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
34. Robinson,J.T., Thorvaldsdóttir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.