Hypatia: a set of quantitative methods for profiling isoform across cell populations

Timothy Pan[1, 2, 3], Cheng-Kai Shiau[1, 2], Lina Lu[1, 2], Changhu Wang[4, 5], Minhua Wang[1, 2], Yueying He[1, 2], Arvind Bhimaraj[6], Daniel Brat[7], Jason Huse[8], Jingyi Jessica Li[4, 5], Ruli Gao[1, 2, 3, *]

Affiliations

[1] Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA 60611

[2] Center for Cancer Genomics, Robert H. Lurie Cancer Center, Northwestern University, Chicago, IL, USA 60611

[3] Driskill Graduate Program, Northwestern University, Chicago, IL, USA 60611

[4] Biostatistics Program, Public Health Sciences Division, Fred Hutch Cancer Center, Seattle, WA, USA 98109

[5] Department of Biostatistics, University of Washington, Seattle, WA, USA 98109

[6] Department of Cardiology, Methodist DeBakey Cardiology Associates, Houston Methodist Hospital, Houston, TX, USA-77030

[7] Department of Pathology, Northwestern University – Feinberg School of Medicine, Chicago, IL, USA 60611

[8] Department of Pathology and Translational Molecular Pathology, University of Texas MD Anderson Cancer Center, Houston, TX, USA 77030

* Corresponding author: Ruli Gao, ruli.gao@northwestern.edu

Abstract

Long-read single-cell RNA-sequencing enables isoform-level study across single cells, yet methods to systematically assess cell-to-cell variations remain limited. Here, we develop Hypatia, a comprehensive platform for dissecting isoform complexities across cell populations, devising Tsallis entropy and Cramer's V to facilitate robust comparative profiling. Hypatia revealed prominent isoform species variations and usage shifts across cell-types in glioblastoma, renal cell carcinoma, and heart, highlighting clinically relevant applications for studying isoform-derived, cell-specific functions.

33 **Main**

34 Alternative splicing is an essential post-transcriptional mechanism that expands the coding
35 capacity of eukaryotic genomes[1]. Despite its critical roles in health and disease, cell-type and cell-
36 state-specific RNA isoform complexity remains poorly understood due to the limitations of bulk
37 and short-read sequencing technologies. Long-read single-cell RNA-sequencing (LR-
38 scRNAseq)[2-7] achieves high-throughput and single-cell resolution of isoform quantification by
39 coupling single-cell barcoding with long-read sequencing platforms (i.e., ONT and PacBio).
40 However, existing analytical strategies[8-10] are suboptimal, being tailored to bulk sequencing and
41 lacking cell-type specificity and optimization required for single-cell profiling.

42 We present Hypatia, a statistical platform for the systematic investigation of population-
43 specific isoform complexity from LR-scRNAseq data, starting from quality control and feature
44 annotation, followed by three modes of differential analyses: 1) isoform diversity, which classifies
45 genes according to isoform species heterogeneity and identifies differential isoform diversity
46 events (DIVs) between cell populations, 2) isoform usage, which identifies differential isoform
47 usage shifts (DIUs) between cell populations, and 3) isoform expression, which detects
48 differentially expressed isoforms (DEIs) across cells (**Fig. 1a**). Each analyses offers unique
49 insights into cell-specific isoforms that are essential for understanding cellular function and
50 disease mechanisms, integrating statistical methods for isoform species diversity, relative usage
51 shifts, and expression abundance in a single platform. In addition, Hypatia leverages object-
52 oriented R programming in the *SingleCellExperiment* class[11], which structures count data, cell-
53 level and feature-level metadata in a synchronized manner, supporting seamless multi-scale data
54 analysis.

55 We established statistical measurements via rigorous tests in simulation, where isoform
56 counts are modeled in single cells for two populations by applying a multinomial probabilistic
57 approach on simulated gene expression, with a subset of genes engineered to bear ground-truth

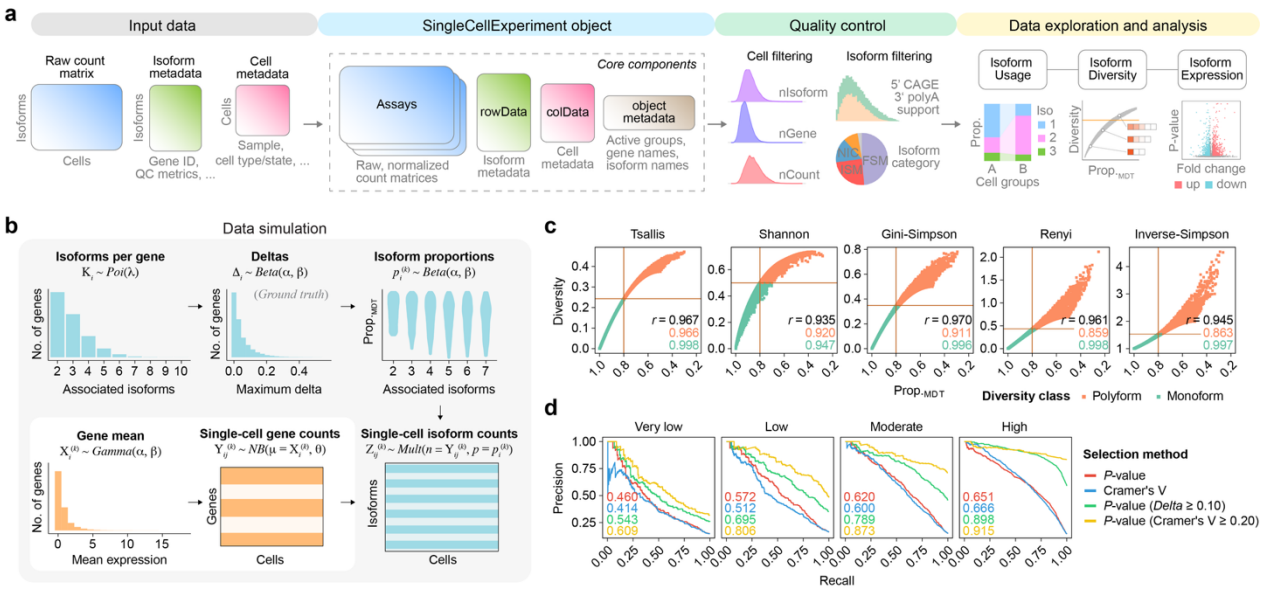

Figure 1. Hypatia workflow and simulation-guided assessment. a, Illustration of the main functional modules and workflow of Hypatia R package. **b,** Diagram outlining the steps in single-cell isoform count simulation. **d,** Relationship between isoform diversity and P_{MDT} of all genes. Pearson correlation coefficients are shown for all data (black), only polyform (orange), and only monoform (cyan) genes. **d,** Precision-recall curve of DIU identification across multiple gene coverage groups with AUCPR shown, colored by selection criteria.

isoform usage variations (**Fig. 1b**, **Online Methods**). To support fair comparisons across cells and genes, we began by assessing isoform species diversity, which represents both the number of isoform species and their evenness within individual genes. A robust diversity quantification is expected to mitigate bias from gene-to-gene variation in isoform species number and from the heavy tail of lowly expressed isoforms. Previous bulk studies have applied entropy-based methods, yet no unified metric capable of effectively capturing differences across genes and cell-types has been established. To this end, we evaluated a range of entropic measures (**Online Methods**) for their reliability and sensitivity in classifying isoform heterogeneity (e.g., 'monoform' versus 'polyform') and the relative prevalence of the most dominant transcript (MDT) per gene. Our analysis revealed that Tallis entropy (q=3) demonstrated the highest association with MDT proportions in both 'monoform' and 'polyform' genes, outperforming the widely used Shannon entropy[2, 12] (**Fig. 1d**). Of note, Tsallis entropy facilitates cross-gene comparisons, being independent of isoform species number compared to other diversity methods (e.g., Shannon,

71 Renyi), minimizing the impact of lowly expressed isoforms. Together, these data establish Tsallis
72 entropy as a unified metric for quantifying isoform species diversity for comparative analyses.

73

74 Isoform usage shifts provide another critical aspect for examining differential isoform
75 profiles across cells, while isoform diversity overlooks the relative ranks among isoforms and can
76 remain unchanged in reciprocal switching events. For this purpose, we evaluated the frequently
77 adapted Chi-square test, which is highly dependent on sample-size and whose feasibility in
78 analyzing sparse single-cell data has not been established. Our assessment revealed that Chi-
79 square P-value alone provided moderate performance (AUCPR 0.460-0.651) in calling true DIU
80 events across coverage groups (**Fig. 1d**). Coupling with an empirical criterion, namely *Delta* (i.e.,
81 the greatest difference in isoform proportions per gene between groups) partially rescued the
82 performance, which however remained low with poor data coverage. Intriguingly, we found that
83 coupling with effect size (e.g., Cramer's V)[13] (**Online Methods**) considerably mitigated the
84 spurious results from sample-size variation and yielded the strongest performance (AUCPR
85 0.609-0.915) across all coverage groups (**Fig. 1d**). To summarize, while the combined use of Chi-
86 square statistics and *Delta* provided a solid baseline for DIU selection, integration with Cramer's
87 V emerged as a superior approach and consistently outperformed all other selection methods.

88 To test applications in profiling cell-type-specific isoforms, we applied Hypatia on three LR-
89 scRNAseq datasets, including an in-house glioblastoma (GBM) dataset and previously published
90 datasets derived from renal cell carcinoma (RCC)[2] and normal heart[12] tissues (**Table S1, Online**
91 **Methods**). Data showed marked expression of unique isoforms associated with cell-type-specific
92 genes in all samples (**Fig. 2a-c**). Deeper analysis with Hypatia revealed a bimodal distribution of
93 Tsallis entropy in all cell-types, corresponding to 'monoform' and 'polyform' genes (**Fig. 2d**).
94 Notably, we observed a distinct increase in isoform diversity of RCC tumor cells and heart
95 cardiomyocytes, suggesting that these cells undergo more active splicing programs.

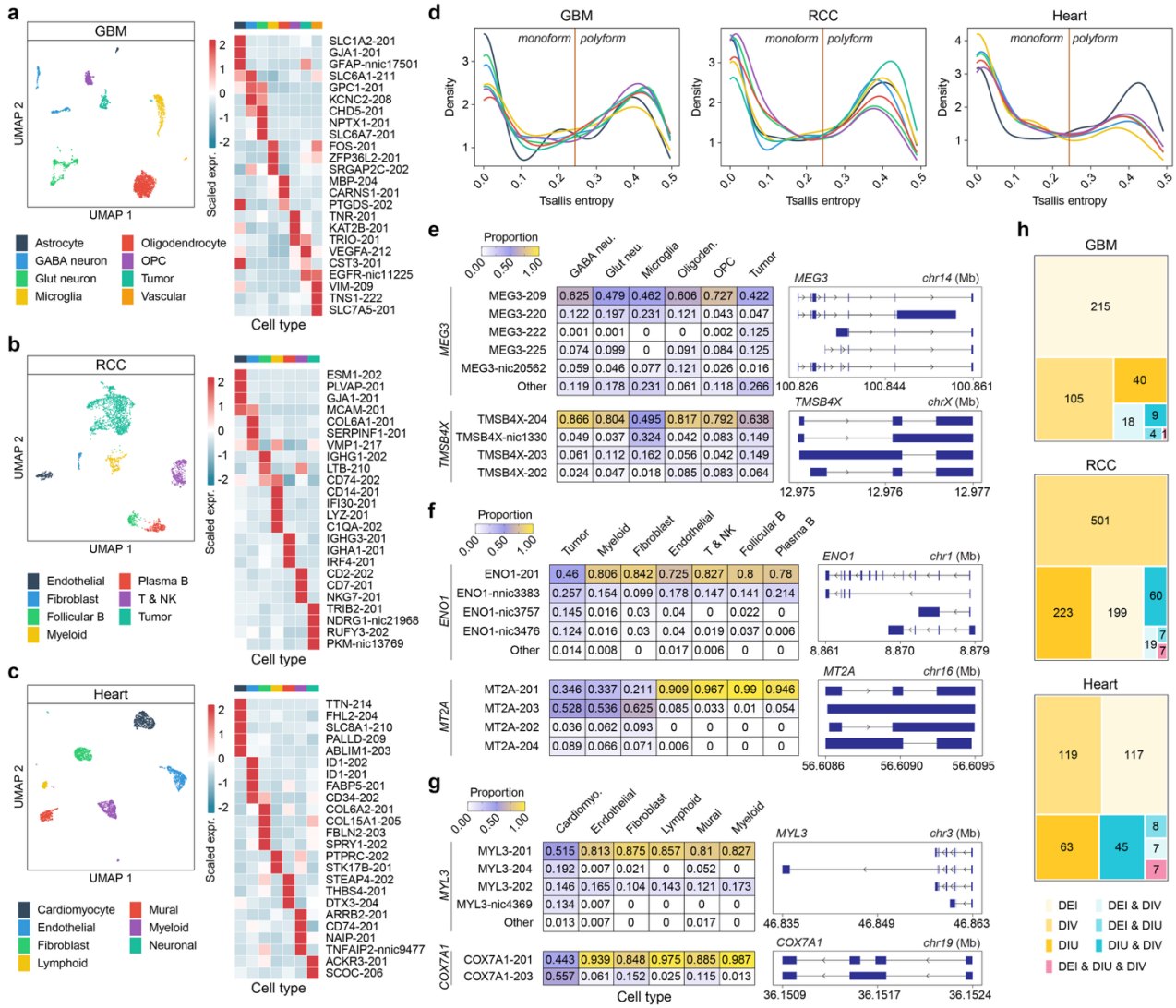

Figure 2. Hypatia characterizes isoform-level variations across cell-types in LR-scRNAseq. a-c, (left) UMAP determined by gene expression analysis and (right) heatmap of DEIs of each cell-type. **d,** Distribution of cell-type-specific isoform diversity of expressed genes in each sample, color coded as in panels **a-c**. **e-g,** (left) Heatmap showing cell-type-specific isoform proportions of DIU events in each sample. (right) Transcript structures of isoforms corresponding to heatmap rows. **h,** Treemaps showing intersections between three analyses, shaded by overlapping genes in each sample.

Furthermore, Hypatia identified extensive isoform usage shifts across cell-types that may reflect underlying regulatory mechanisms of cell-type-specific functions. In GBM, multiple DIU events were identified in the tumor-suppressor lncRNA gene *MEG3* (**Fig. 2e**). The relative abundance of dominant isoform MEG3-209 (ENST00000451743) varied broadly across multiple cell-types, whereas MEG3-220 (ENST00000524035) and novel isoform MEG3-nic20562 had

101 distinct expression in neoplastic tumor cells and oligodendrocytes, respectively. Similarly, the
102 inflammation-related gene, *TMSB4X,* exhibited a marked decrease in protein-coding isoform
103 TMSB4X-204 (ENST00000451311) specific to microglial cells, accompanied by an increase in
104 novel isoform TMSB4X-nic1330, which harbors a retained intron, indicating cell-type-specific
105 splicing regulation. In RCC, significant DIU events were detected in glycolysis gene *ENO1* and
106 metallothionein-encoding gene *MT2A* (**Fig. 2f**). Tumor cells showed a distinct polyform
107 heterogeneity of *ENO1*, while a striking dichotomy in the dominant isoform of *MT2A* emerged
108 across cell-types. In cardiomyocytes of the heart, we observed unique splicing patterns in *MYL3*
109 and *COX7A1*, which are key regulators in sarcomere contraction and energy production,
110 respectively (**Fig. 2g**). These data signified that splicing regulation involved key cell-type-specific
111 functional genes across various tissue types.

112 While DIU analysis captures changes in relative isoform usage with consideration of the
113 ranked order of dominant isoforms, DIVs reflect differences in evenness independent of ranking.
114 Indeed, this was well-reflected in comparative findings from DIU and DIV analysis. In GBM, DIU
115 and DIV analyses identified isoform complexity changes in 54 and 133 genes, respectively, with
116 9 overlapped genes (**Fig. 2h**). In RCC, 60 among 297 DIU and 587 DIV events were jointly
117 detected. In heart, 45 genes were jointly detected among 123 DIU and 96 DIV events. In contrast,
118 DEI analysis identified fewer genes linked to cell-type-specific isoform profiles (N= 23, 33, and 22
119 genes in GBM, RCC, and heart, respectively), suggesting that DEI analysis is largely confounded
120 by overall gene expression levels. Taken together, our results show that DIU and DIV analyses
121 offer a complementary approach for discovering isoform complexity changes across diverse cell
122 populations.

123 In closing, we present Hypatia, a novel statistical framework for the systematic analysis of
124 cell-to-cell variations in isoform complexity. Our analysis reveals that Tsallis entropy robustly
125 quantifies isoform species heterogeneity, and Cramer's V effectively curates Chi-square statistics.

Critically, we show that DIU and DIV tests are complementary in dissecting differential isoform patterns across cell populations. Results from clinically relevant datasets show the prevalence of isoform-level variations across cell-types and exemplify the important applications of Hypatia in revealing comprehensive isoform spectrums of human tissues.

Data availability

Both raw and processed data generated in this study have been deposited in the Gene Expression Omnibus (GEO) database under accession GSE310974. Previously published sequencing data were downloaded from SRA under runs SRR21492157 (RCC) and SRR32154432 (heart). Simulation data have been deposited in Zenodo (doi: 10.5281/zenodo.17781226).

Code availability

All source code and documentation of Hypatia are released to the public and freely available on Github at https://github.com/gaolabtools/Hypatia.

Acknowledgements

We thank research funding support provided to R.G. by the National Institute of General Medical Sciences (NIH R35GM142539) and National Cancer Institute SPORE for Translational Approaches to Brain Cancer (P50CA221747).

Author contributions

T.P. led the development of Hypatia, performed data analysis, and wrote the manuscript. C.K.S. processed raw sequencing data and reviewed results. M.W. performed the sequencing experiment and edited the manuscript. L.L. and Y.H. reviewed results. Q.W. and J.J.L. participated in data analysis, reviewed results, and edited the manuscript. J.H., D.B. and A.B. assisted with sample collection and edited the manuscript. R.G. supervised the study, conceived concepts, designed the computational tool, and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Online Methods

Simulation of single-cell isoform expression data

Isoform data were simulated for 30,000 genes and 1,000 cells of two populations, each with 500 cells. Generation of single-cell gene expression counts for each group was adapted from Splatter[14] and utilized the gamma distribution for mean gene expression (k = 0.2, θ = 2) and the negative binomial distribution with a dispersion parameter of 0.10. The number of isoforms per gene were drawn from a Poisson distribution (λ = 2) with a minimum of 2 and maximum of 10 possible isoforms. Isoform-level *Deltas* for each gene and corresponding isoform proportions for both groups were sampled from beta distributions. Genes with at least one *Delta* of $\geq$ 0.10 across groups were considered ground-truth DIUs. Isoform counts were then generated from the multinomial distribution using group-specific isoform proportions and gene counts per cell. In performance evaluation, gene coverage groups were categorized according to total counts per gene: $\leq$ 100 counts (‘very low’), 100 to 200 (‘low’), 200 to 500 (‘moderate’), and $\geq$ 500 (‘high’).

170

171 **Sample acquisition and LR-scRNA sequencing**

172 Fresh frozen tissue from a grade 4 glioblastoma of the frontal lobe was collected from the NBST
173 tissue bank at Northwestern University and subjected to snNanoRNAseq[2]. The single-nuclei
174 suspension was prepared using the Chromium Nuclei Isolation Kit (10X Genomics) following the
175 manufacturer's protocol. Nuclei were then loaded onto a 10X Genomics Chromium Controller (iX)
176 with Chip J with an expected recovery of 3,000 nuclei. Full-length barcoded cDNAs were amplified
177 and enriched for long fragments. The final library was prepared using the SQK-LSK114 ligation
178 sequencing kit (Oxford Nanopore Technologies) and sequenced on an in-house PromethION with
179 an R10.4M flow cell (Oxford Nanopore Technologies).

180

181 **LR-scRNAseq data preprocessing**

182 All processing of raw data was performed using scNanoGPS[2] (version 2.0). Isoform quantification
183 was performed with IsoQuant[15] (version 3.4.2) using the 'unique only' transcript quantification
184 setting, and reference genome GRCh38 (GENCODE v44/Ensembl110 annotations), excluding
185 reads mapping to *MALAT1* and *NEAT1* due to computational restraints. SQANTI3[16] (version 5.2.2)
186 was used for transcript annotation. Transcripts categorized as full-splice match were retained.
187 Novel transcripts required support of 5' CAGE peaks in refTSS[17] (version 4.1) and 3' polyA sites
188 in PolyASite[18] (version 3.0). All transcripts were required to be detected in at least 3 cells in the
189 sample and free of predicted RT-switching artifacts.

190

191 **Single-cell gene expression analysis**

Single-cell gene expression analysis was performed using Seurat[19] (version 5.0.1). Cells with expression of more than 200 genes and less than 5% of mitochondrial genes were retained. Clustering utilized the top 15 principal components and cell-types were annotated using established gene markers[20].

Differential isoform usage analysis

To calculate isoform usage shifts as implemented in Hypatia's 'RunDIU' function, a contingency table was constructed for each gene using aggregated isoform counts of each cell group in the comparison. The Chi-square test was applied, and *P*-values were adjusted for multiple testing using the Benjamini-Hochberg method. Each cell-type was compared against all other cell-types for all comparative analyses. Comparisons resulting in an adjusted *P*-value of ≤ 0.05 and Cramer's V ≥ 0.20 were considered significant. Only genes expressed in at least 5% of cells in each group and supported by 15 total UMI counts were considered.

Calculation of effect size

The effect size of Chi-square test was calculated as Cramer's V (**Eq. 1**),

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}, \tag{1}$$

where χ^2 denotes the Chi-square statistic, n denotes the total isoform count of the gene, and k is the number of columns in the contingency table (equal to 2 for two-group comparisons).

212 **Isoform diversity analysis**

213 Isoform diversity was calculated using Tsallis (**Eq. 2**, q = 3), Shannon (**Eq. 3**), Gini-Simpson (**Eq.**
214 **4**), Renyi (**Eq. 5**, α = 2), and Inverse-Simpson (**Eq. 6**) entropies as implemented in the 'RunDIV'
215 function of Hypatia. The entropies were calculated as the following,

$$S = \frac{1}{q-1}\left(1 - \sum_{i}^{n} p_i^q\right), \tag{2}$$

$$H = -\sum_{i=1}^{2} p_i \, log(p_i), \tag{3}$$

$$D = 1 - \sum_{i}^{n} p_i^2, \tag{4}$$

$$H_\alpha = \frac{1}{1-\alpha}\log\left(\sum_{i}^{2} p_i^\alpha\right), \tag{5}$$

$$D^{-1} = \frac{1}{D}, \tag{6}$$

216 where n denotes the total number of expressed isoforms of the gene, p is the isoform proportion,
217 and q and α represent the order of the Tsallis and Renyi entropies, respectively. The Shannon
218 and Renyi entropies utilized only the top 2 isoform abundances to normalize comparisons across
219 genes. Cutoffs for monoform and polyform classifications were guided by a P_{MDT} = 0.80, which
220 corresponded to entropies at 0.243 (Tsallis), 0.500 (Shannon), 0.348 (Gini-Simpson), 0.435
221 (Renyi), and 1.533 (inverse-Simpson). DIVs were defined as genes whose diversity classifications
222 diverged across cell groups. Only genes with expression in at least 5% of cells of each group of
223 the comparison were considered.

224

225 **Isoform expression analysis**

Raw isoform count data were first normalized using 'NormalizeCounts' in Hypatia. Counts were normalized by the total counts per cell, multiplied by a scale factor of 10,000, and subsequently Log1p-transformed. The average isoform expression across cell-types were calculated using 'RunDEI', which applies the Wilcoxon Rank Sum Test with Bonferroni corrected *P*-values. Significant DEIs were those that had |Log2 fold change| $\geq$ 0.60, adjusted *P*-value $\leq$ 0.05, and detected in at least 5% of cells in comparisons.

Intersection of differential isoform events

Genes presenting differential isoform events were enumerated per cell-type comparison, where each cell-type was compared against all others. Genes associated with more than one DEI were only counted once per comparison. Genes that appeared in more than one cell-type comparison were assigned a separate count for each comparison.

References

1. Jiang, W. & Chen, L. Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. *Comput Struct Biotechnol J* **19**, 183-195 (2021).
2. Shiau, C.K. et al. High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. *Nat Commun* **14**, 4124 (2023).
3. Al'Khafaji, A.M. et al. High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nat Biotechnol* **42**, 582-586 (2024).
4. Fan, X. et al. Single-cell RNA-seq analysis of mouse preimplantation embryos by third-generation sequencing. *PLoS Biol* **18**, e3001017 (2020).
5. Lebrigand, K., Magnone, V., Barbry, P. & Waldmann, R. High throughput error corrected Nanopore single cell transcriptome sequencing. *Nat Commun* **11**, 4025 (2020).
6. Philpott, M. et al. Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq. *Nat Biotechnol* **39**, 1517-1520 (2021).
7. Wang, Q. et al. Single cell transcriptome sequencing on the Nanopore platform with ScNapBar. *RNA* (2021).
8. Guo, W., Calixto, C.P.G., Brown, J.W.S. & Zhang, R. TSIS: an R package to infer alternative splicing isoform switches for time-series data. *Bioinformatics* **33**, 3308-3310 (2017).
9. de la Fuente, L. et al. tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. *Genome Biol* **21**, 119 (2020).

10. Wu, S. & Schmitz, U. ScIsoX: a multidimensional framework for measuring isoform-level transcriptomic complexity in single cells. *Genome Biol* **26**, 289 (2025).
11. Amezquita, R.A. et al. Orchestrating single-cell analysis with Bioconductor. *Nature Methods* **17**, 137-145 (2020).
12. Pan, T. et al. Single-Cell Splicing Isoform Atlas of the Adult Human Heart and Heart Failure. *Circulation* **152**, 1501-1514 (2025).
13. Cramér, Harald. 1946. Mathematical Methods of Statistics. Princeton: Princeton University Press, page 282 (Chapter 21. The two-dimensional case). ISBN 0-691-08004-6 (table of content Archived 2016-08-16 at the Wayback Machine).
14. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology* **18**, 174 (2017).
15. Prjibelski, A.D. et al. Accurate isoform discovery with IsoQuant using long reads. *Nature Biotechnology* **41**, 915-918 (2023).
16. Pardo-Palacios, F.J. et al. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nature Methods* **21**, 793-797 (2024).
17. Abugessaisa, I. et al. refTSS: A Reference Data Set for Human and Mouse Transcription Start Sites. *Journal of Molecular Biology* **431**, 2407-2422 (2019).
18. Moon, Y., Herrmann, Christina J., Mironov, A. & Zavolan, M. PolyASite v3.0: a multi-species atlas of polyadenylation sites inferred from single-cell RNA-sequencing data. *Nucleic Acids Research* **53**, D197-D204 (2024).
19. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411-420 (2018).
20. Couturier, C.P. et al. Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nature Communications* **11**, 3406 (2020).