Hypatia: Comparative Isoform Profiling Across Cell Populations from Long-Read Single-Cell Transcriptomes

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

Email addresses

Timothy Pan, timothy.pan@northwestern.edu

Cheng-Kai Shiau, chengkai.shiau@northwestern.edu

Lina Lu, lina.lu@northwestern.edu

Changhu Wang, wangch156@g.ucla.edu

Minhua Wang, minhua.wang@northwestern.edu

Yueying He, yueying.he@northwestern.edu

Arvind Bhimaraj, ABhimaraj@houstonmethodist.org

Daniel Brat, daniel.brat@northwestern.edu

Jason Huse, jhuse@mdanderson.org

Jingyi Jessica Li, lijy03@fredhutch.org

Ruli Gao, ruli.gao@northwestern.edu

Abstract

High-throughput long-read single-cell RNA-sequencing enables isoform-level study across single cells, yet methods for systematically assessing 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.

Keywords

isoform usage shift, Tsallis isoform diversity, Chi-square correction

Background

Alternative splicing is an essential post-transcriptional mechanism that expands the coding capacity of eukaryotic genomes(1). Despite its critical roles in health and disease, cell-type- and cell-state-specific RNA isoform complexity remains poorly understood due to the limitations of bulk and short-read sequencing technologies. Long-read single-cell RNA-sequencing (LR-scRNAseq)(2-7) achieves high-throughput and single-cell resolution of isoform quantification by coupling single-cell barcoding with long-read sequencing platforms (i.e., ONT and PacBio). However, existing analytical strategies(8-10) lack the desired levels of granularity in calling cell population-specific events relevant to isoform-centric regulatory mechanisms in cells. To address this knowledge gap, we develop Hypatia, a computational platform for the systematic quantification and comparison of population-specific isoform profiles from LR-scRNAseq data.

Results and Discussion

The workflow of Hypatia inputs single cell isoform expression data enabled by LR-scRNAseq to perform three layers of isoform complexity quantification and subsequently compare them across cell populations, including 1) isoform diversity analysis, which classifies genes according to isoform species heterogeneity and identifies differential isoform diversity events (DIVs) between cell populations, 2) isoform usage shift analysis, which identifies differential isoform usage shifts (DIUs) between cell populations, and 3) isoform expression analysis, which detects differentially expressed isoforms (DEIs) across cells (**Fig. 1a**). Each analyses offers unique insights into cell-specific isoforms that are essential for understanding cellular function and disease mechanisms, integrating statistical methods for isoform species diversity, relative usage shifts, and expression abundance in a single platform. In addition, Hypatia leverages object-oriented R programming in the *SingleCellExperiment* class(11), which structures count data, cell-level and feature-level metadata in a synchronized manner, supporting seamless multi-scale data analysis.

We established statistical measurements via rigorous tests in simulation, where isoform counts are modeled in single cells for two populations by applying a multinomial probabilistic approach on simulated gene expression, with a subset of genes engineered to bear ground-truth isoform usage variations (**Fig. 1b**, **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 (**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,

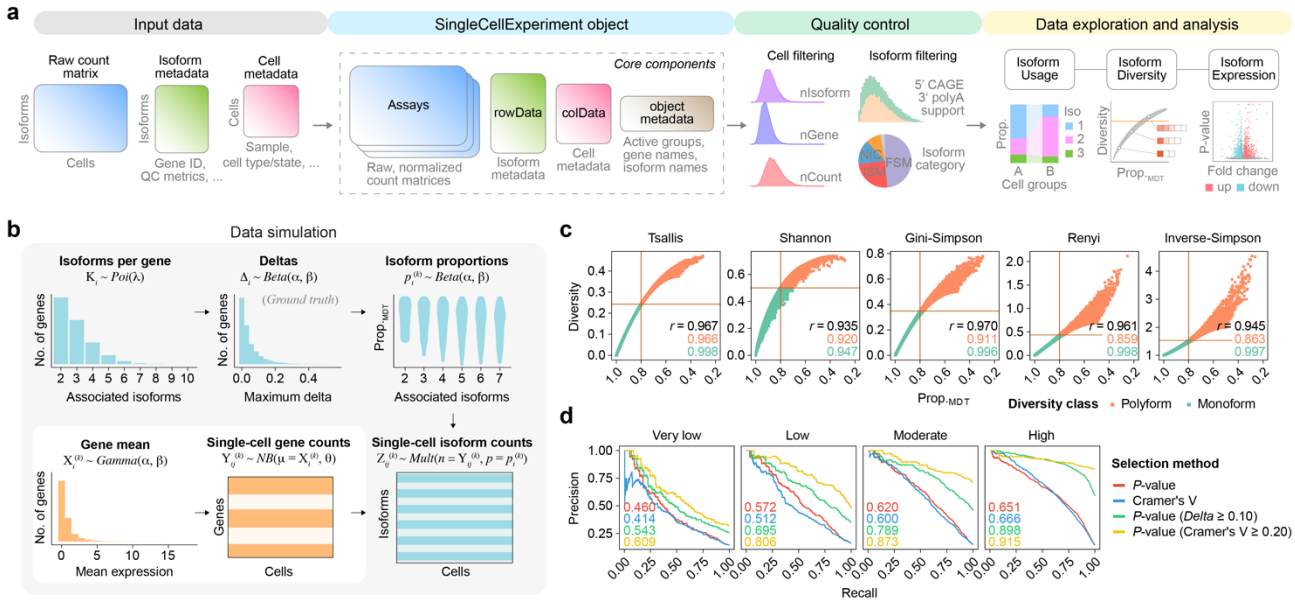

Figure 1. Hypatia workflow and simulation-guided assessment. a, Illustration of the main functional modules and workflow of Hypatia R package. **b,** Diagram outlining the steps in single-cell isoform count simulation. **d,** Relationship between isoform diversity and P_{MDT} of all genes. Pearson correlation 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.

Tsallis entropy facilitates cross-gene comparisons, being independent of isoform species number compared to other diversity methods (e.g., Shannon, Renyi), minimizing the impact of lowly expressed isoforms. Together, these data establish Tsallis entropy as a unified metric for quantifying isoform species diversity for comparative analyses.

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

To test applications in profiling cell-type-specific isoforms, we applied Hypatia on three LR-scRNAseq datasets, including an in-house glioblastoma (GBM) dataset and previously published datasets derived from renal cell carcinoma (RCC)(2) and normal heart(12) tissues (**Table S1, Methods**). Data showed marked expression of unique isoforms associated with cell-type-specific genes in all samples (**Fig. 2a-c**). Deeper analysis with Hypatia revealed a bimodal distribution of Tsallis entropy in all cell-types, corresponding to 'monoform' and 'polyform' genes (**Fig. 2d**).

Notably, we observed a distinct increase in isoform diversity of RCC tumor cells and heart cardiomyocytes, suggesting that these cells undergo more active splicing programs.

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

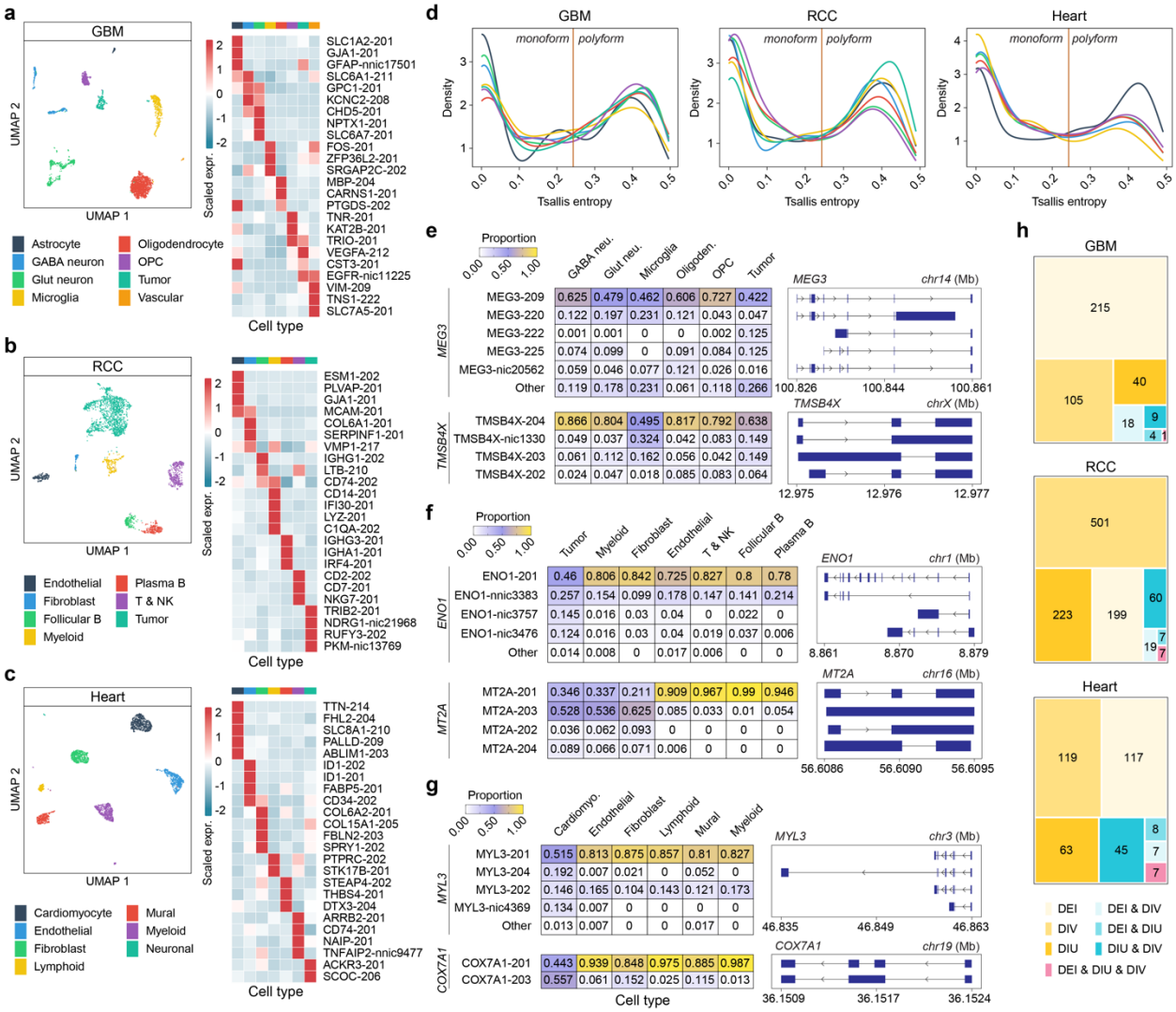

Figure 2. Hypatia characterizes isoform-level variations across cell-types in LR-scRNAseq. a-c, (left) UMAP determined by gene expression analysis and (right) heatmap of DEIs of each cell-type. **d,** Distribution of cell-type-specific isoform diversity of expressed genes in each sample, color coded 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.

125 abundance of dominant isoform MEG3-209 (ENST00000451743) varied broadly across multiple
126 cell-types, whereas MEG3-220 (ENST00000524035) and novel isoform MEG3-nic20562 had
127 distinct expression in neoplastic tumor cells and oligodendrocytes, respectively. Similarly, the
128 inflammation-related gene, *TMSB4X,* exhibited a marked decrease in protein-coding isoform
129 TMSB4X-204 (ENST00000451311) specific to microglial cells, accompanied by an increase in
130 novel isoform TMSB4X-nic1330, which harbors a retained intron, indicating cell-type-specific
131 splicing regulation. In RCC, significant DIU events were detected in glycolysis gene *ENO1* and
132 metallothionein-encoding gene *MT2A* (**Fig. 2f**). Tumor cells showed a distinct polyform
133 heterogeneity of *ENO1*, while a striking dichotomy in the dominant isoform of *MT2A* emerged
134 across cell-types. In cardiomyocytes of the heart, we observed unique splicing patterns in *MYL3*
135 and *COX7A1*, which are key regulators in sarcomere contraction and energy production,
136 respectively (**Fig. 2g**). These data signified that splicing regulation involved key cell-type-specific
137 functional genes across various tissue types.
138
139 While DIU analysis captures changes in relative isoform usage with consideration of the
140 ranked order of dominant isoforms, DIVs reflect differences in evenness independent of ranking.
141 Indeed, this was well-reflected in comparative findings from DIU and DIV analysis. In GBM, DIU
142 and DIV analyses identified isoform complexity changes in 54 and 133 genes, respectively, with
143 9 overlapped genes (**Fig. 2h**). In RCC, 60 among 297 DIU and 587 DIV events were jointly
144 detected. In heart, 45 genes were jointly detected among 123 DIU and 96 DIV events. In contrast,
145 DEI analysis identified fewer genes linked to cell-type-specific isoform profiles (N= 23, 33, and 22
146 genes in GBM, RCC, and heart, respectively), suggesting that DEI analysis is largely confounded
147 by overall gene expression levels. Taken together, our results show that DIU and DIV analyses
148 offer a complementary approach for discovering isoform complexity changes across diverse cell
149 populations.

Collectively, our results highlight the utility of Hypatia for isoform-resolved analysis of single-cell transcriptomes. We show that Tsallis entropy robustly quantifies isoform species heterogeneity and Cramer's V effectively curates Chi-square statistics from long-read single-cell isoform data. 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.

Conclusions

In this study, we present Hypatia, a novel analytical framework for in-depth comparisons of cell-to-cell variations in isoform complexity from long-read single-cell transcriptomic data. Hypatia mitigates challenges of data noise and sparsity through structured and systematic approaches, which greatly extends the profiling capacity of single-cell transcriptomics beyond overall gene expression levels and provides comprehensive details at isoform-level.

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

176 performance evaluation, gene coverage groups were categorized according to total counts per
177 gene: ≤ 100 counts ('very low'), 100 to 200 ('low'), 200 to 500 ('moderate'), and ≥ 500 ('high').
178
179 **Sample acquisition and LR-scRNA sequencing**
180 Fresh frozen tissue from a grade 4 glioblastoma of the frontal lobe was collected from the NBST
181 tissue bank at Northwestern University and subjected to snNanoRNAseq(2). The single-nuclei
182 suspension was prepared using the Chromium Nuclei Isolation Kit (10X Genomics) following the
183 manufacturer's protocol. Nuclei were then loaded onto a 10X Genomics Chromium Controller (iX)
184 with Chip J with an expected recovery of 3,000 nuclei. Full-length barcoded cDNAs were amplified
185 and enriched for long fragments. The final library was prepared using the SQK-LSK114 ligation
186 sequencing kit (Oxford Nanopore Technologies) and sequenced on an in-house PromethION with
187 an R10.4M flow cell (Oxford Nanopore Technologies).
188
189 **LR-scRNAseq data preprocessing**
190 All processing of raw data was performed using scNanoGPS(2) (version 2.0). Isoform
191 quantification was performed with IsoQuant(15) (version 3.4.2) using the 'unique only' transcript
192 quantification setting, and reference genome GRCh38 (GENCODE v44/Ensembl110 annotations),
193 excluding reads mapping to *MALAT1* and *NEAT1* due to computational restraints. SQANTI3(16)
194 (version 5.2.2) was used for transcript annotation. Transcripts categorized as full-splice match
195 were retained. Novel transcripts required support of 5' CAGE peaks in refTSS(17) (version 4.1)
196 and 3' polyA sites in PolyASite(18) (version 3.0). All transcripts were required to be detected in at
197 least 3 cells in the sample and free of predicted RT-switching artifacts.
198
199 **Single-cell gene expression analysis**
200 Single-cell gene expression analysis was performed using Seurat(19) (version 5.0.1). Cells with
201 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).

Isoform diversity analysis

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

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

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

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

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

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

223 where n denotes the total number of expressed isoforms of the gene, p is the isoform proportion,
224 and q and α represent the order of the Tsallis and Renyi entropies, respectively. The Shannon
225 and Renyi entropies utilized only the top 2 isoform abundances to normalize comparisons across
226 genes. Cutoffs for monoform and polyform classifications were guided by a P_{MDT} = 0.80, which
227 corresponded to entropies at 0.243 (Tsallis), 0.500 (Shannon), 0.348 (Gini-Simpson), 0.435
228 (Renyi), and 1.533 (inverse-Simpson). DIVs were defined as genes whose diversity classifications
229 diverged across cell groups. Only genes with expression in at least 5% of cells of each group of
230 the comparison were considered.
231

232 **Isoform expression analysis**

233 Raw isoform count data were first normalized using 'NormalizeCounts' in Hypatia. Counts were
234 normalized by the total counts per cell, multiplied by a scale factor of 10,000, and subsequently
235 Log1p-transformed. The average isoform expression across cell-types were calculated using
236 'RunDEI', which applies the Wilcoxon Rank Sum Test with Bonferroni corrected *P*-values.
237 Significant DEIs were those that had |Log2 fold change| ≥ 0.60, adjusted *P*-value ≤ 0.05, and
238 detected in at least 5% of cells in comparisons.
239

240 **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.

Declarations

Ethics approval and consent to participate

The IRB protocol for studying the multi-omics profiles of human tissues were approved by Northwestern University (STU00219936). Consent was waved because this study only involves the secondary analysis of the banked tissues, which has no direct interaction with subjects.

Consent for publication

Not applicable.

Availability of data and materials

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

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Institute of General Medical Sciences (NIH R35GM142539) and National Cancer Institute SPORE for Translational Approaches to Brain Cancer (P50CA221747).

Authors' contributions

TP led the development of Hypatia, performed data analysis, and wrote the manuscript. CKS processed raw sequencing data and reviewed results. MW performed the sequencing experiment and edited the manuscript. LL and YH reviewed results. QW and JJL participated in data analysis, reviewed results, and edited the manuscript. JH, DB and AB assisted with sample collection and edited the manuscript. RG supervised the study, conceived concepts, designed the computational tool, and wrote the manuscript.

Acknowledgements

We thank all our funding supports and patients who consented for research usages of their tissue samples.

References

1. Jiang W, Chen L. Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. Comput Struct Biotechnol J. 2021;19:183-95.
2. Shiau CK, Lu L, Kieser R, Fukumura K, Pan T, Lin HY, et al. High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. Nat Commun. 2023;14(1):4124.
3. Al'Khafaji AM, Smith JT, Garimella KV, Babadi M, Popic V, Sade-Feldman M, et al. High-throughput RNA isoform sequencing using programmed cDNA concatenation. Nat Biotechnol. 2024;42(4):582-6.

292 4. Fan X, Tang D, Liao Y, Li P, Zhang Y, Wang M, et al. Single-cell RNA-seq analysis of
293 mouse preimplantation embryos by third-generation sequencing. PLoS Biol.
294 2020;18(12):e3001017.
295 5. Lebrigand K, Magnone V, Barbry P, Waldmann R. High throughput error corrected
296 Nanopore single cell transcriptome sequencing. Nat Commun. 2020;11(1):4025.
297 6. Philpott M, Watson J, Thakurta A, Brown T, Jr., Brown T, Sr., Oppermann U, et al.
298 Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq. Nat Biotechnol.
299 2021;39(12):1517-20.
300 7. Wang Q, Boenigk S, Boehm V, Gehring NH, Altmueller J, Dieterich C. Single cell
301 transcriptome sequencing on the Nanopore platform with ScNapBar. RNA. 2021.
302 8. Guo W, Calixto CPG, Brown JWS, Zhang R. TSIS: an R package to infer alternative
303 splicing isoform switches for time-series data. Bioinformatics. 2017;33(20):3308-10.
304 9. de la Fuente L, Arzalluz-Luque A, Tardaguila M, Del Risco H, Marti C, Tarazona S, et al.
305 tappAS: a comprehensive computational framework for the analysis of the functional impact of
306 differential splicing. Genome Biol. 2020;21(1):119.
307 10. Wu S, Schmitz U. ScIsoX: a multidimensional framework for measuring isoform-level
308 transcriptomic complexity in single cells. Genome Biol. 2025;26(1):289.
309 11. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating
310 single-cell analysis with Bioconductor. Nature Methods. 2020;17(2):137-45.
311 12. Pan T, Lu L, Youker K, Shiau CK, Wang M, Lin HY, et al. Single-Cell Splicing Isoform Atlas
312 of the Adult Human Heart and Heart Failure. Circulation. 2025;152(21):1501-14.
313 13. Cramér, Harald. 1946. Mathematical Methods of Statistics. Princeton: Princeton University
314 Press, page 282 (Chapter 21. The two-dimensional case). ISBN 0-691-08004-6 (table of content
315 Archived 2016-08-16 at the Wayback Machine).
316 14. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data.
317 Genome Biology. 2017;18(1):174.

318 15. Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, et al. Accurate
319 isoform discovery with IsoQuant using long reads. Nature Biotechnology. 2023;41(7):915-8.
320 16. Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín
321 R, et al. SQANTI3: curation of long-read transcriptomes for accurate identification of known and
322 novel isoforms. Nat Methods. 2024;21(5):793-7.
323 17. Abugessaisa I, Noguchi S, Hasegawa A, Kondo A, Kawaji H, Carninci P, et al. refTSS: A
324 Reference Data Set for Human and Mouse Transcription Start Sites. Journal of Molecular Biology.
325 2019;431(13):2407-22.
326 18. Moon Y, Herrmann Christina J, Mironov A, Zavolan M. PolyASite v3.0: a multi-species
327 atlas of polyadenylation sites inferred from single-cell RNA-sequencing data. Nucleic Acids
328 Research. 2024;53(D1):D197-D204.
329 19. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic
330 data across different conditions, technologies, and species. Nature Biotechnology.
331 2018;36(5):411-20.
332 20. Couturier CP, Ayyadhury S, Le PU, Nadaf J, Monlong J, Riva G, et al. Single-cell RNA-
333 seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. Nature
334 Communications. 2020;11(1):3406.
335