Investigating RNA splicing as a source of cellular diversity using a binomial mixture model

Keren Isaev
ki2255@columbia.edu
Columbia University
New York Genome Center

David A Knowles
dak2173@columbia.edu
Columbia University
New York Genome Center

Abstract

Alternative splicing (AS) contributes significantly to RNA and protein variability yet its role in defining cellular diversity is not fully understood. While Smart-seq2 offers enhanced coverage across transcripts compared to 10X single cell RNA-sequencing (scRNA-seq), current computational methods often miss the full complexity of AS. Most approaches for single cell based differential splicing analysis focus on simple AS events such as exon skipping, and rely on predefined cell type labels or low-dimensional gene expression representations. This limits their ability to detect more complex AS events and makes them dependent on prior knowledge of cell classifications. Here, we present Leaflet, a splice junction centric approach inspired by Leafcutter, our tool for quantifying RNA splicing variation with bulk RNA-seq. Leaflet is a probabilistic mixture model designed to infer AS-driven cell states without the need for cell type labels. We detail Leaflet's generative model, inference methodology, and its efficiency in detecting differentially spliced junctions. By applying Leaflet to the Tabula Muris brain cell dataset, we highlight cell-state specific splicing patterns, offering a deeper insight into cellular diversity beyond that captured by gene expression alone.

1 Introduction

Alternative splicing (AS) is a crucial cellular process that generates multiple RNA isoforms, leading to the protein and regulatory RNA diversity necessary for the proper function of complex organisms [1, 2]. The mammalian nervous system is a prime example where AS is prevalent and responsible for increased proteomic and functional complexity [3, 4, 5] . However, the extent to which AS contributes to cell-type level specific variation remains incompletely understood. Addressing this requires an unbiased analysis of AS in relation to cellular states yet quantifying it using common single-cell RNA sequencing techniques such as 10X is challenging due to its pronounced 3' bias in coverage. Smart-seq2 however, while less high-throughput than droplet based approaches, offers superior transcript coverage, making it a promising experimental approach [6]. Importantly, several single-cell atlases encompassing both 10X and Smart-seq2 data across various cell types and tissues in mouse and human have been published [7, 8]. These offer potential insights into the AS profiles of individual cells and cell types.

Despite the challenges posed by current single-cell technologies, a number of methods have been developed to capture differential splicing in single cells. BRIE2 [9] uses Bayesian regression to model the association of specific AS events with cell-level phenotypes. However, it is mainly tailored for exon-skipping events and requires predefined binary cell type labels such as treatment status. Another approach, scQuint [10], integrates a method where introns sharing 3' splice sites converge to form alternative intron groups. It then maps the genome-wide intron usage onto a low-dimensional latent space with a Dirichlet-Multinomial VAE, followed by a differential splicing test across predefined cell types or clusters, which emerge from the splicing latent space. However, the obtained splicing latent

space with this approach did not reveal any novel cell sub population in Tabula Muris. Meanwhile, Psix[11] employs a probabilistic model to detect splicing variation across a spectrum of single cells. Inspired by autocorrelation, Psix seeks patterns of AS that correspond to continuous cell states, such as developmental stages, without the need for explicit cell clustering, labeling, or trajectory inference. However, Psix's approach relies heavily on a gene expression-based low-dimensional representation. Thus, it is inherently limited to identifying splicing events that are directly associated with gene expression changes. Furthermore, a significant limitation of Psix is its reliance on K-nearest neighbors (KNN) for determining cell similarity, which makes the method time-consuming, especially with large single cell datasets. Overall, the focus of recent approaches has largely remained on identifying variable cassette exons, likely resulting in the oversight of more complex splicing events.

To address this gap in our understanding of splicing events in single cells, we developed Leaflet, building on the splicing quantification representation employed in Leafcutter [12] for bulk RNA-seq. Our binomial mixture model uses junction counts from split reads (described in **Section 2**) to model latent cell states and the splicing patterns that define them **(Figure 1)**.

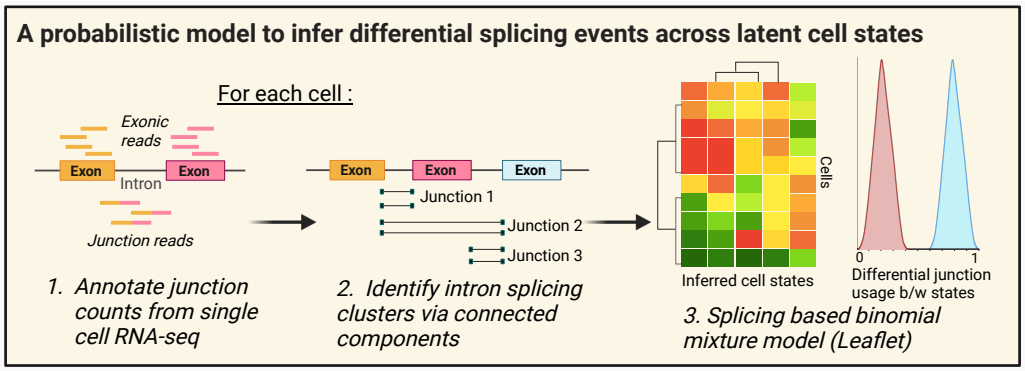

Figure 1: Overview of single cell AS quantification. Junction coordinates and reads are first obtained for each cell, followed by intron clustering of overlapping junctions with common splice sites. Three exemplary junctions are shown for one intron cluster event, exon skipping. Read counts for junctions and intron clusters are used as input for Leaflet's binomial mixture model.

2 Splicing quantification with Smart-seq2 reads

To generate an input dataset for our model, we processed reads mapping to splice junctions in individual cells. We used Regtools [13] to obtain a matrix of junction coordinates and read counts while keeping track of cell barcodes. Leaflet follows Leafcutter's approach to defining local splicing variation via "intron clusters" but provides additional functionality for single-cell applications. For example, in our analysis of Tabula Muris, we required each junction to be observed in at least five cells in each cell type with at least one read each. These parameters are user adjustable and the full list of them is described in the package https://github.com/daklab/Leaflet. Leaflet then combines junctions into intron clusters based on overlapping introns and splice sites. An example of an intron cluster composed of three overlapping junctions is shown in **Figure 1**. Intron clusters with a minimum of two junctions are kept for downstream analysis. We define the junction usage ratio (or in a slight abuse of notation, the percent spliced in or PSI) as the splice read count for the junction divided by the total count for the cluster.

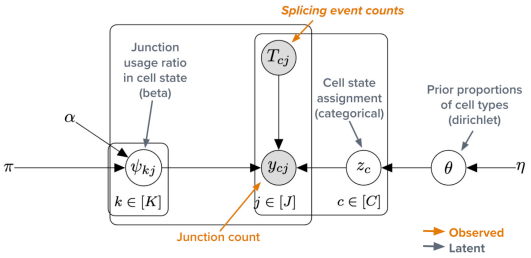

Figure 2: A graphical model for Leaflet's mixture model. Variables described in **Table 1**.

3 Leaflet probabilistic model

3.1 Generative model

Table 1: Summary of Variables in Leaflet

Variable	Description
C	Total number of cells in the dataset.
J	Total number of junctions considered.
K	Total number of cell states or cell types.
$\psi_{k:}$	Vector of junction usage ratios for each state k.
θ	Proportion of each cell type or cell state.
z_c	Cell assignment for cell c.
y_{cj}	Observed junction counts for junction j in cell c.
T_{cj}	Observed intron cluster counts for junction j in cell c.

The graphical model for Leaflet is shown in **(Figure 2)** with C cells, J junctions and K states. Leaflet posits the existence of K cell states, each defined by a vector of junction usage ratios $\psi_{k:}$. In the context of a collection of single cells, we model the prior proportion of each cell type or cell state in the sample as θ. Cell assignments z_c are sampled from a categorical distribution parameterized by θ. z_c alongside ψ_{kj} model the observed junction counts y_{cj}. T_{cj} are the observed intron cluster counts for a given junction j in a given cell c.

Our likelihood model is

$$y_{cj}|z_c, \psi \sim \text{Binomial}(T_{cj}, \psi_{z_c,j}).$$

The proportion of counts attributed to each junction within an intron cluster is specified by ψ_{kj}, the latent success rate of junction j in whichever state k the cell is in. The junction usage ratios are given a Beta prior distribution with fixed hyperparameters a and b,

$$\psi_{kj} \sim \text{Beta}(a, b).$$

The latent cell assignments z_c, specifying the state of cell c, are modeled as a categorical given θ,

$$z_c|\theta \sim \text{Categorical}(\theta).$$

Of note, within a single cell c, the model assumes that the observations at different junctions are conditionally independent, given the cluster assignment and the associated ψ. Meanwhile, θ, denotes the prior on cell assignments and is modeled using a Dirichlet distribution with fixed prior η/k,

$$\theta \sim \text{Dirichlet}\left(\frac{\eta}{K}, \cdots, \frac{\eta}{K}\right).$$

The true number of cell states is typically unknown. Our approach is to set the parameter vector of the Dirichlet to $[\eta/K, \cdots, \eta/K]$ which approximates a Dirichlet Process (DP) as $K \to \infty$. When used for clustering, the approximate DP with large K provides a way to infer the number of clusters from the data, instead of having to pre-specify it. Thus, we recommend setting K to a larger number of cell states than expected to infer the optimal value.

4 Variational inference

We approximate the posterior distribution using mean-field coordinate ascent variational inference (CAVI) [14]. To learn the parameters, we optimize the Evidence Lower Bound (ELBO, 9.1). Our variational posterior is

$$q(\psi, z, \theta) = q(\theta) \prod_c q(z_c) \prod_{kj} q(\psi_{kj}). \tag{1}$$

where

$$q(\psi_{kj}) = \text{Beta}(\alpha_{jk}, \beta_{jk}) \tag{2}$$
$$q(z_c) = \text{Cat}(\phi_c) \tag{3}$$
$$q(\theta) = \text{Dir}(\gamma) \tag{4}$$

Here $\alpha \in \mathbb{R}^{K\times J}, \beta \in \mathbb{R}^{K\times J}, \phi \in \mathbb{R}^{K\times N}$, and $\gamma \in \mathbb{R}^K$ are the variational parameters to be optimized. Given the variational parameters α and β, the expected log-likelihood under the variational approximation of observing y_{ck} under component k is

$$l_{ck} = y_{ck} \cdot \mathbb{E}[\log(\psi_k)] + (T_{ck} - y_{ck}) \cdot \mathbb{E}[\log(1 - \psi_k)], \tag{5}$$

where $\mathbb{E}[\log(\psi_k)] = \text{digamma}(\alpha_k) - \text{digamma}(\alpha_k + \beta_k)$ and $\mathbb{E}[\log(1 - \psi_k)] = \text{digamma}(\beta_k) - \text{digamma}(\alpha_k + \beta_k)$.

The CAVI updates are

$$\phi_{ck} \propto \exp(\mathbb{E}[\log \theta_k] + l_{ck}) \tag{6}$$

$$\gamma_k = \frac{\eta}{K} + \sum_c \phi_{kc} \tag{7}$$

$$\alpha_{kj} = a + \sum_c \phi_{kc} y_{cj} \tag{8}$$

$$\beta_{kj} = b + \sum_c \phi_{kc} \tilde{y}_{cj}, \tag{9}$$

where $\tilde{y}_m = T_m - y_m$ and $\mathbb{E}[\log \theta_k] = \text{digamma}(\gamma_k) - \text{digamma}(\sum_{k'} \gamma_{k'})$. The total cluster count T_{cj} is 0 for most (cell-junction) pairs across any (current) single cell sequencing platform. It is thus highly beneficial to represent y and $\tilde{y}$ as sparse matrices. The updates for α and β are then given in matrix notation as,

$$\alpha = a + \phi y \tag{10}$$

$$\beta = b + \phi \tilde{y}. \tag{11}$$

Our implementation uses `pytorch` allowing the CAVI updates to be easily run on a GPU for highly efficient model fitting.

5 Leaflet differential splicing analysis

We test whether junction j is differentially spliced across cell states by calculating an approximate log Bayes factor (ALBF) between hypothesis H_0 that $\psi_{jk} = \psi_j$ for all k and H_1 that the ψ_{jk} are different across states. Given that the variational posterior on ψ is a Beta distribution, we can compute an approximate marginal likelihood under H_0 as

$$P_j(H_0) = \int \prod_k \text{Beta}(\psi_j | \alpha_{kj}, \beta_{kj}) d\psi_j = \frac{B(\sum_k \alpha_{kj}, \sum_k \beta_{kj})}{\prod_k B(\alpha_{kj}, \beta_{kj})} \tag{12}$$

where $\text{B}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$. The approximate marginal likelihood for H_1 is

$$P_j(H_1) = \prod_k \int \text{Beta}(\psi_{kj} | \alpha_{kj}, \beta_{kj}) d\psi_{kj} = 1.$$

Thus, the ALBF for junction j is

$$\text{ALBF}_j = \log \frac{P_j(H_1)}{P_j(H_0)} = \sum_k \log \text{B}(\alpha_{kj}, \beta_{kj}) - \log \text{B}\left(\sum_k \alpha_{kj}, \sum_k \beta_{kj}\right) \tag{13}$$

6 Evaluation of performance on simulated data

We compared Leaflet, Psix and Kruskal-Wallis in their abilities to detect simulated differentially spliced events. Here we show results simulating discrete cell-types, but we also compare the methods on a continuous trajectory dataset simulated by Psix in Appendix 9.2, Table 3.

6.1 Splicing differences between discrete cell-types

Our simulation approach adheres closely to Leaflet's generative model (**Figure 2**), but uses the observed cluster counts matrix (detailed in **Section 2**) and predefined brain cell types from Tabula Muris to ensure the overall sparsity of read coverage is maintained. To allow for a fair comparison with the type of splicing events captured by Psix (exon skipping), we filtered the Leaflet intron clusters to exclusively encompass exon inclusion/exclusion events, corresponding to three junctions per intron cluster.

Method	AUROC	AUPR
Leaflet-junction	**0.9696**	**0.9782**
Leaflet-cluster	**0.9832**	**0.9875**
Psix	0.7937	0.8557
KW-junction	0.9499	0.9649
KW-cluster	0.4717	0.4753

Table 2: Comparison of methods on discrete counts generated from Leaflet generative model simulation. Data corresponds to 5,968 cells and 5,676 junctions from 1,892 intron clusters.

In our simulation, every intron cluster is randomly assigned a positive (has differential splicing across cell states) or negative label (no differential splicing). Junctions are sorted into the two corresponding to exon inclusion (**J1** and **J2**) and the one corresponding to exon exclusion (**J3**). For negative intron clusters we sample one shared ψ^{J3} from Beta$(0.5, 0.5)$ (hyperparameters chosen to emulate the known bimodality of splicing patterns [15]) for all cell states. For positive intron clusters, we sample a different $\psi_k^{J3} \sim$ Beta$(0.5, 0.5)$ for each cell state k. We then set $\psi_k^{J1} = \psi_k^{J2} = (1 - \psi_k^{J3})/2$ for each k. Junction counts are then sampled from binomial distributions with these ψ values and total count from the observed ("real") Tabula Muris data.

6.2 Identifying cell-state associated splicing events

For our first test, we focused on the two most prevalent cell types in Tabula Muris brain tissue: microglia and oligodendrocytes (5,499 cells total). From 1,892 intron clusters, we sampled split read counts for 5,676 junctions based on these specified cell types. The ability of Leaflet, Psix, and the Kruskal-Wallis test to detect the simulated differentially spliced events was then assessed.

Leaflet's mixture model was fit on the simulated counts with 100 CAVI iterations across 50 independent trials with K set to two. These trials represent random initializations of the variational parameters. The trial resulting in the highest ELBO was used for downstream analysis. For each junction cell-type pair, the posterior mean was $\hat{\psi}_{kj} = \frac{\alpha_{kj}}{\alpha_{kj}+\beta_{kj}}$. The Spearman correlation between the inferred and simulated $\Delta\psi$ between the two cell states was 0.92.

For comparison, Psix (version 0.11.0) was executed using its default parameters, and the first two principal components were determined based on the simulated intron cluster counts. In this setting, an intron cluster corresponded to a "gene". Psix requires several input files including a gene expression-based low-dimensional representation, a junction counts matrix and cassette exon annotations to compute Psix scores. Empirical *p*-value estimation with its kNN-based method extends its run time (21.34 seconds for Leaflet vs 5 minutes for Psix for the same dataset).

The Kruskal-Wallis (KW) test was run using the two vectors of simulated junction usage ratios (KW-junction) or intron cluster based PSI corresponding to J_3 (KW-cluster) only for the two cell types. The results are shown in **Table 2** where Leaflet consistently outperforms other methods. Specifically, Leaflet-cluster, which evaluates the overall intron cluster Leaflet score (by summing scores across the three junctions in a given cluster), achieves the highest Area Under the Receiver Operating Characteristic (AUROC) and Area Under the Precision-Recall (AUPR) scores.

6.3 Inference of cell state identity

Leaflet simultaneously infers junctions that are differentially spliced across cell states as well as z_c, the cell assignments. We evaluated Leaflet on a larger simulated dataset using eight brain cell types from Tabula Muris, fixing cell intron cluster counts and true cell-type assignments while simulating junction ψ.

This data consists of 7,816 cells and 3,613 intron clusters corresponding to cassette exon events. We applied Leaflet's mixture model to the simulated counts, running 100 CAVI iterations from 50 random initializations with K set at 50. The variational parameter ϕ_c determined the cell state assignments for each individual cell. To evaluate reproducibility, we analyzed the consistency of cell-cell co-assignments across random initializations **Figure 3**. This revealed a robust structure: cells originating from the same cell types were largely co-assigned across trials (Adjusted Rand Score = 1.0). Further, even with K set at 50, Leaflet identified the correct number of predefined cell types as eight, specifically where $\theta \geq 0.05$. This analysis reflects Leaflet's ability to detect inherent cell types by leveraging AS patterns.

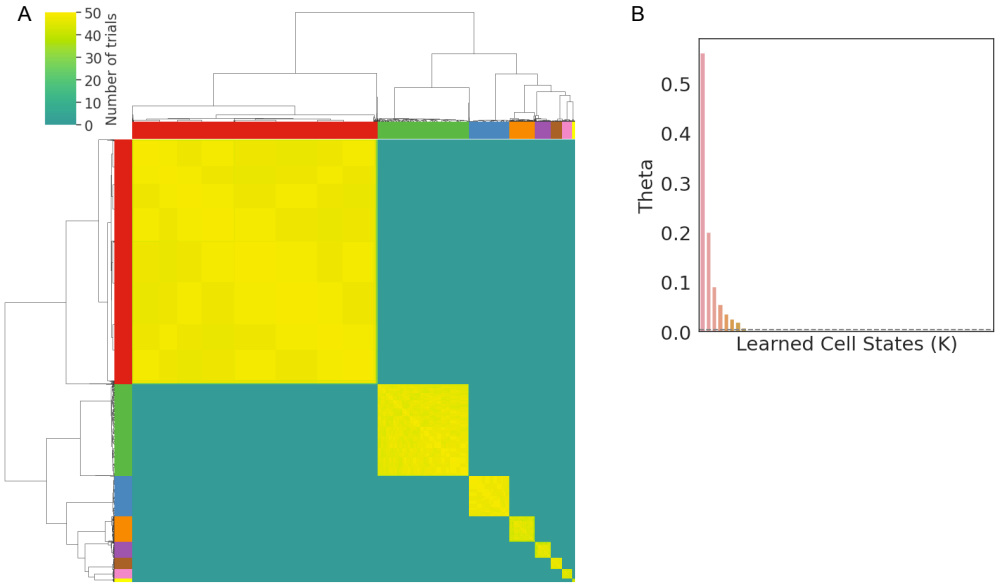

Figure 3: (A) Hierarchical clustering of cell-cell co-assignments across 50 independent trials on simulated data. Color bars indicate pre-assigned cell types. (B) Learned values for θ across 50 cell states. Dashed grey line indicates 0.05%.

7 Application to Tabula Muris Brain data

7.1 Model consistency in cell assignments and association with known cell types

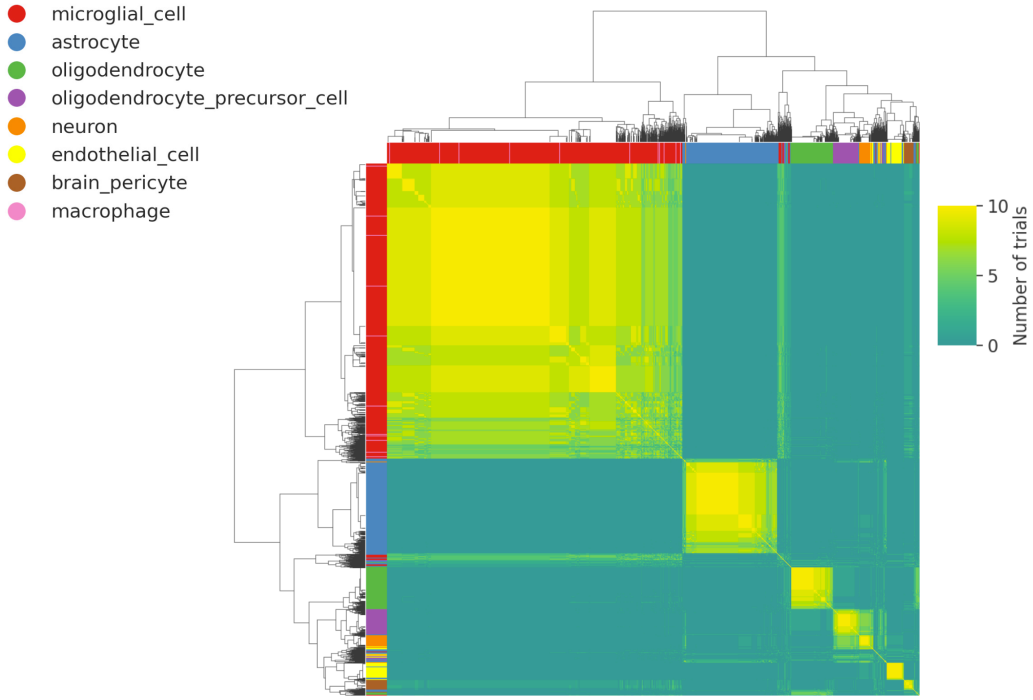

Figure 4: Hierarchical clustering of cell pair co-assignments across ten independent trials (n=5,200 randomly sampled cells are shown).

Smart-seq2 data for individual cells across eight brain cell types in Tabula Muris was processed as described in **Section 2**. Inton clusters were then annotated with M10-PLUS genes. In this reference-aware mode, we detected 16,191 junctions across 5,433 intron clusters and 4,135 genes.

We initially assessed Leaflet's clustering consistency across ten random initializations by calculating how often each pair of cells were assigned to the same cell state (to account for label switching between runs). We set K to 10 for this assessment. The results revealed a strong tendency for the same cell type pairs to be co-assigned, highlighting clear cluster structure in the data (Adjusted Rand Score = 0.80, **Figure 4**). We then selected the trial with the highest ELBO from the ten runs. Inspecting the posterior mean $\hat{theta} = \gamma / \sum_k \gamma_k$, we found that all K (10) cell states appeared in at least 5% of cells (**Figure 8A,B**). Notably, most cell states predominantly corresponded to one or two closely related cell types (**Figure 5A,B**). For example, Cell States 1 and 5 were predominantly enriched in macrophages and microglia (**Figure 5D**). These results suggest that Leaflet is able to recover biologically meaningful distinct cell-types using only splicing patterns.

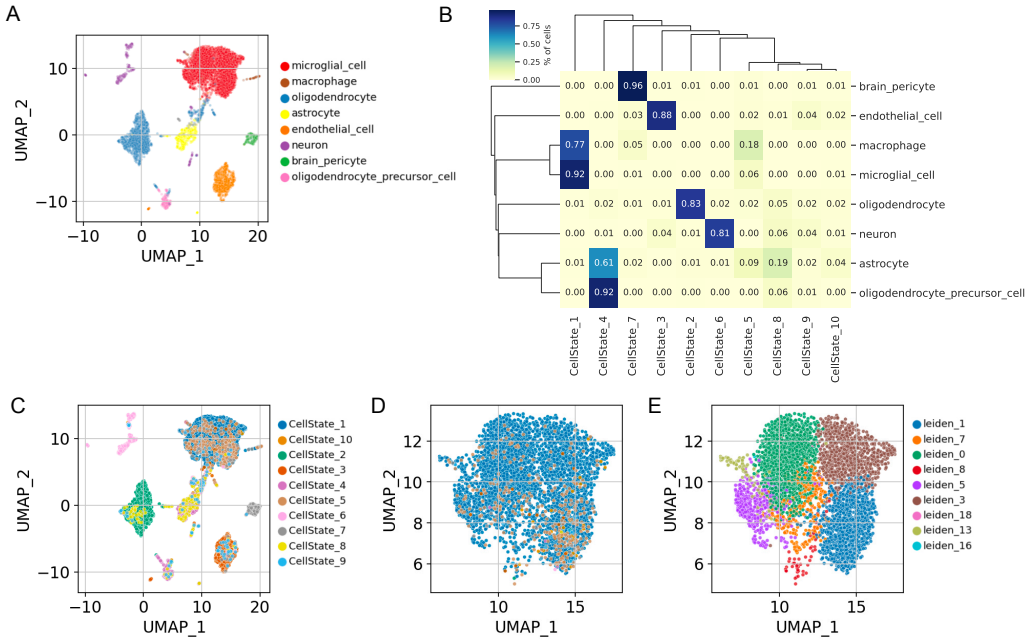

Figure 5: (A) UMAP based on PCA analysis of gene expression overlaid with known cell types. (B) Percentages of cell types belonging to each Leaflet cell state.(C) Leaflet cell state assignments overlaid on UMAP from (A). (D) Subset of UMAP plot is shown for predefined microglial cells (same coour legend as C). Cells are coloured by Leaflet cells states where the first and fifth state make up the majority of this population. (E) Same cells as in panel D but coloured by Leiden clusters defined by total gene expression.

7.2 Cell state specific splicing events shape cellular diversity

Cellular heterogeneity may be shaped by both gene expression and AS patterns, the latter of which is less well understood. We first evaluated the learned $J \times K$ matrix of Leaflet ψs as described above (**Section 6.2**). We compared these learned ψ values to empirical (observed) junction usage ratios (per cell), both of which exhibited peaks at 0 and 1 (**Figure 6A,B**).

We then obtained Leaflet ALBF scores for each junction, where higher scores indicate a junction is more variable across all K cell states (**Figure 6C**). We found that of the top 10% scoring junctions, less than 7% mapped to any marker genes, suggesting this strong splicing variation is mostly independent of expression. The top two scoring junctions were in the same intron cluster (corresponding to an alt 3' splice site) in the *Psap* gene. One splice site is used predominately in pericytes and oligodendrocytes, while the other is used in neurons and astrocytes (and interestingly oligodendrocyte precursors show a mix of both, **Figure6D,E**).

We next calculated Leaflet ALBF scores across only two cell states of interest (cell states 1 and 5), corresponding to microglial cells. One of the top scoring junctions mapped to the gene *App* and was highly used in cell state 5 compared to cell state 1 (**Figure 6F**). This highlights that there may be a significant level of AS heterogeneity even within a well defined cell type such as microglia. We note that these cells are not directly linked to total gene expression based subpopulation of cells as determined by Leiden clustering (**Figure5**E) and may correspond to complementary levels of information.

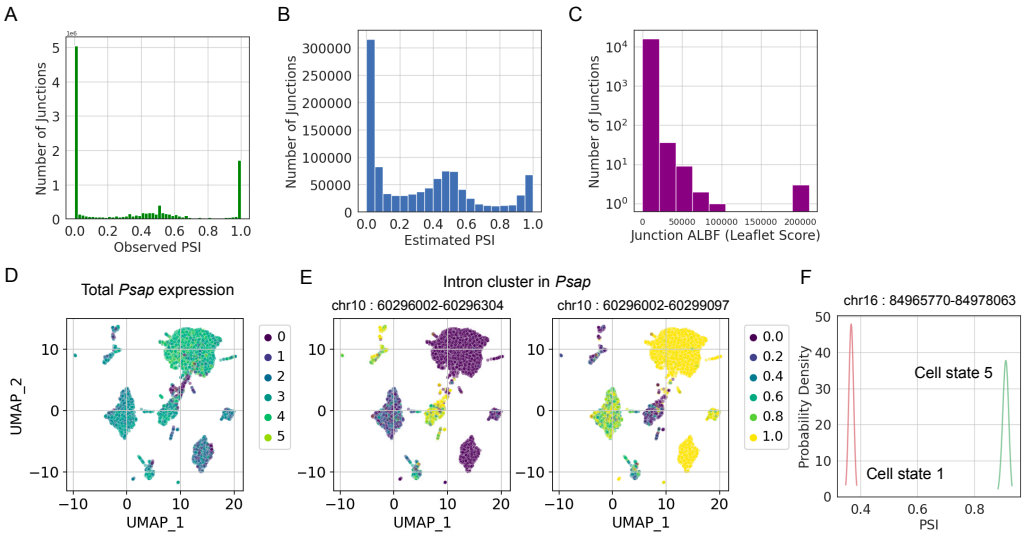

Figure 6: (A) PSI values for observed junction-cell pairs (y_{cj}/T_{cj}). (B) Distribution of Leaflet $\hat{\psi}$s for junction-cell state pairs. (C) Junction ALBF scores across all cell states. PSI_{jk} are shown only for those junctions and cell states in which $\alpha_{jk} + \beta_{jk} \geq 10$ to exclude those with very high uncertainty (B,C).(D) *Psap* normalized gene expression values across cell types suggests it is not a marker gene. (E) An example of an intron cluster in the gene *Psap*, consisting of two junctions with high ALBF scores, showing variable AS across cell types. (F) Posterior distributions $\text{Beta}(\alpha_{kj}, \beta_{kj})$ are shown for a junction mapping to the gene *App* in cell states 1 and 5.

8 Discussion

In this work, we introduce Leaflet, a novel, flexible approach to identify AS-driven cell states, a dimension of cellular diversity that is often unexplored. By focusing on local splicing variation at the single-cell level and comparing the differential splicing observed across learned cell states with those observed in pre-defined cell types, Leaflet reveals AS heterogeneity even within established cell types such as microglia. Our preliminary findings suggest that the landscape of cell-type specific differential splicing is not only a proxy for differential gene expression, but an additional, orthogonal layer of regulatory information.

While the current application to brain cells sequenced with Smart-seq2 has proven valuable, we plan to expand our method to analyze high-throughput datasets from techniques such as SPLiT-seq [16] and EasySci [17]. Although long-read scRNA-seq might ultimately be the definitive approach for deciphering isoform variation, short-read protocols look likely to remain dominant in the short to medium term.

While any biological insights made herein require additional validation in external datasets, we present a new approach and highlight several examples for further study. Leaflet is fast to run and requires minimal input from users.

9 Appendix

9.1 Evidence Lower Bound (ELBO)

The ELBO can be decomposed as

$$\begin{aligned}\text{ELBO} &= \mathbb{E}_q[\log p] - \mathbb{E}_q[\log q] \\ &= \int q \log p - \text{entropy}\end{aligned}$$

where the former term for Leaflet's mixture model is

$$\begin{aligned}\int q \log p = &\sum_{jk} (\alpha - 1) \langle \log \psi_{jk} \rangle + (\beta - 1) \langle \log (1 - \psi_{jk}) \rangle + \sum_k (\frac{\eta}{K} - 1) \langle \log \theta_k \rangle \\ &+ \underbrace{\sum_c \phi_{ck} \langle \log \theta_k \rangle}_{\sum_k \langle \log \theta_k \rangle \sum_c \phi_{ck}} + \sum_{ck} \phi_{ck} \underbrace{\sum_j [y_{cj} \langle \log \psi_{jk} \rangle + \tilde{y}_{cj} \langle \log (1 - \psi_{jk}) \rangle]}_{l_{ck}}\end{aligned}$$

where $\phi_{ck} = \langle z_c = k \rangle$ and $\tilde{y}_{cj} = T_{cj} - y_{cj}$.

9.2 Splicing variation along a simulated continuous trajectory

Here we compare Leaflet and Psix, utilizing data that was simulated to align with continuous cell state trajectories. The simulated dataset consists of 5,000 genes and 1,000 cells that follow a one-dimensional trajectory where 40% of genes have variable splicing. Each gene was simulated to have two isoforms, one with and without a cassette exon. For exons marked as "non-cell state associated", their PSIs were simulated to be the same across all cells (**Figure 7**). This dataset corresponds to the simulated data used in the Psix manuscript [11]. The Kruskal-Wallis test was used as a baseline. Given the junction-centric approach of Leaflet, gene-level Leaflet ALBF scores were derived by summing the ALBF scores for relevant junctions. Importantly, Leaflet was not designed for continuous, trajectory-like cell states yet it obtained accurate estimates even on such data (**Table 3**). Additionally, Psix runtime was 22.2 minutes while Leaflet inference took only 35 seconds with 100 CAVI iterations (all methods tested on CPU).

Method	AUROC	AUPRC
Leaflet-cluster	0.8942	0.9058
Psix	0.9399	0.9438
KW	0.9496	0.9008

Table 3: Performance at detecting differentially spliced genes from data simulated by Psix with cells along a continuous trajectory. KW: Kruskal–Wallis test.

9.3 Low dimensional representation of Tabula Muris brain cells using gene expression

We utilized total gene counts obtained with FeatureCounts [18] for these cells to obtain gene expression based low dimensional representations for downstream analysis. We used the Scanpy pipeline with recommended settings for quality control of cells and genes [19] and total count based normalization. Principal Component Analysis (PCA) was run on 7,560 cells and 3,837 genes. Uniform Manifold Approximation and Projection (UMAP) was run on the first 40 principal components (PCs) to obtain a low dimensional transcriptomic representation of the cells (**Figure 5A**). Marker genes (n=30) were obtained for each of the nine cell types in the data using the built in T-test approach.

9.4 Supplementary Figures

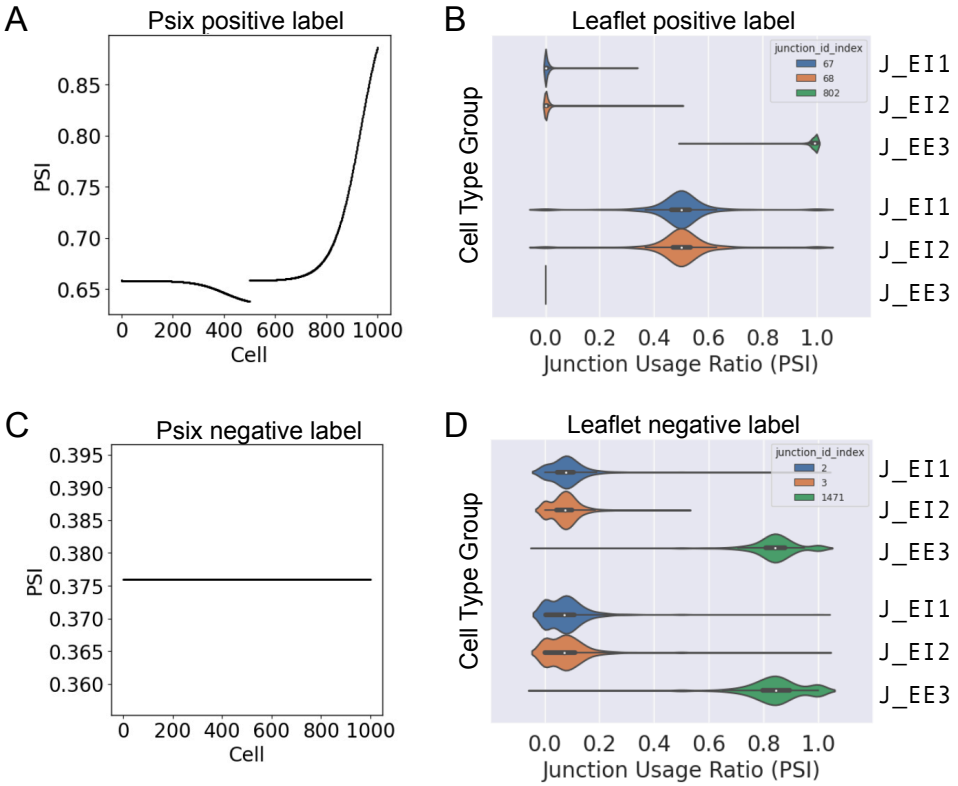

Figure 7: Simulated PSIs from Psix (left) and Leaflet (right). Simulated differential splicing is shown in the top panels for positively labelled events. Psix simulates cells along a trajectory. A cell state associated exon is shown in **A** and a cell type associated intron cluster composed of three junctions is shown in **B**. Bottom panels show instances of simulated PSI values for non-differentially spliced exons (**C**) or intron clusters (**D**). J_{EI}: Exon inclusion junction. J_{EE}: Exon exclusion junction.

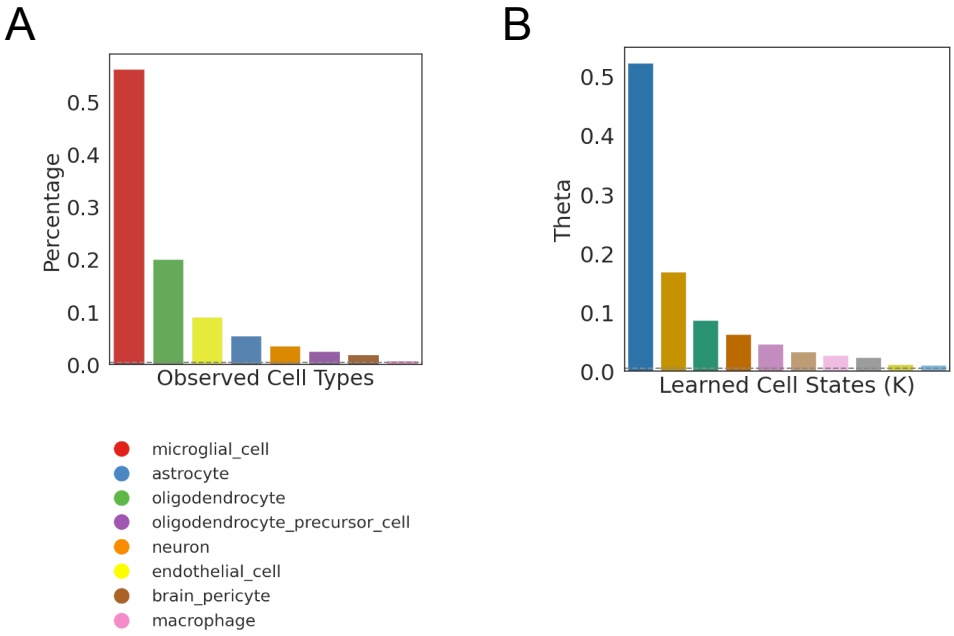

Figure 8: (A) Distribution of predefined cell types in the dataset. (B) Distribution of θ learned for each cell state.

References

[1] Nuno L Barbosa-Morais, Manuel Irimia, Qun Pan, Hui Y Xiong, Serge Gueroussov, Leo J Lee, Valentina Slobodeniuc, Claudia Kutter, Stephen Watt, Recep Colak, Taehyung Kim, Christine M Misquitta-Ali, Michael D Wilson, Philip M Kim, Duncan T Odom, Brendan J Frey, and Benjamin J Blencowe. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593, December 2012.

[2] Pavel V Mazin, Philipp Khaitovich, Margarida Cardoso-Moreira, and Henrik Kaessmann. Alternative splicing during mammalian organ development. *Nat. Genet.*, 53(6):925–934, June 2021.

[3] John A Calarco, Simone Superina, Dave O'Hanlon, Mathieu Gabut, Bushra Raj, Qun Pan, Ursula Skalska, Laura Clarke, Danielle Gelinas, Derek van der Kooy, Mei Zhen, Brian Ciruna, and Benjamin J Blencowe. Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell*, 138(5):898–910, September 2009.

[4] Jonathan P Ling, Christopher Wilks, Rone Charles, Patrick J Leavey, Devlina Ghosh, Lizhi Jiang, Clayton P Santiago, Bo Pang, Anand Venkataraman, Brian S Clark, Abhinav Nellore, Ben Langmead, and Seth Blackshaw. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nat. Commun.*, 11(1):137, January 2020.

[5] Qinghong Yan, Sebastien M Weyn-Vanhentenryck, Jie Wu, Steven A Sloan, Ye Zhang, Kenian Chen, Jia Qian Wu, Ben A Barres, and Chaolin Zhang. Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals NMD modulating chromatin regulators. *Proc. Natl. Acad. Sci. U. S. A.*, 112(11):3445–3450, March 2015.

[6] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.*, 42(14):8845–8860, August 2014.

[7] 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. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, October 2018.

[8] Tabula Sapiens Consortium*, Robert C Jones, Jim Karkanias, Mark A Krasnow, Angela Oliveira Pisco, Stephen R Quake, Julia Salzman, Nir Yosef, Bryan Bulthaup, Phillip Brown, William Harper, Marisa Hemenez, Ravikumar Ponnusamy, Ahmad Salehi, Bhavani A Sanagavarapu, Eileen Spallino, Ksenia A Aaron, Waldo Concepcion, James M Gardner, Burnett Kelly, Nikole Neidlinger, Zifa Wang, Sheela Crasta, Saroja Kolluru, Maurizio Morri, Angela Oliveira Pisco, Serena Y Tan, Kyle J Travaglini, Chenling Xu, Marcela Alcántara-Hernández, Nicole Almanzar, Jane Antony, Benjamin Beyersdorf, Deviana Burhan, Kruti Calcuttawala, Matthew M Carter, Charles K F Chan, Charles A Chang, Stephen Chang, Alex Colville, Sheela Crasta, Rebecca N Culver, Ivana Cvijović, Gaetano D'Amato, Camille Ezran, Francisco X Galdos, Astrid Gillich, William R Goodyer, Yan Hang, Alyssa Hayashi, Sahar Houshdaran, Xianxi Huang, Juan C Irwin, Sori Jang, Julia Vallve Juanico, Aaron M Kershner, Soochi Kim, Bernhard Kiss, Saroja Kolluru, William Kong, Maya E Kumar, Angera H Kuo, Rebecca Leylek, Baoxiang Li, Gabriel B Loeb, Wan-Jin Lu, Sruthi Mantri, Maxim Markovic, Patrick L McAlpine, Antoine de Morree, Maurizio Morri, Karim Mrouj, Shravani Mukherjee, Tyler Muser, Patrick Neuhöfer, Thi D Nguyen, Kimberly Perez, Ragini Phansalkar, Angela Oliveira Pisco, Nazan Puluca, Zhen Qi, Poorvi Rao, Hayley Raquer-McKay, Nicholas Schaum, Bronwyn Scott, Bobak Seddighzadeh, Joe Segal, Sushmita Sen, Shaheen Sikandar, Sean P Spencer, Lea C Steffes, Varun R Subramaniam, Aditi Swarup, Michael Swift, Kyle J Travaglini, Will Van Treuren, Emily Trimm, Stefan Veizades, Sivakamasundari Vijayakumar, Kim Chi Vo, Sevahn K Vorperian, Wanxin Wang, Hannah N W Weinstein, Juliane Winkler, Timothy T H Wu, Jamie Xie, Andrea R Yung, Yue Zhang, Angela M Detweiler, Honey Mekonen, Norma F Neff, Rene V Sit, Michelle Tan, Jia Yan, Gregory R Bean, Vivek Charu, Erna Forgó, Brock A Martin, Michael G Ozawa, Oscar Silva, Serena Y Tan, Angus Toland, Venkata N P Vemuri, Shaked Afik, Kyle Awayan, Olga Borisovna Botvinnik, Ashley Byrne, Michelle Chen, Roozbeh Dehghannasiri, Angela M Detweiler, Adam Gayoso, Alejandro A Granados, Qiqing Li, Gita Mahmoudabadi, Aaron McGeever, Antoine

de Morree, Julia Eve Olivieri, Madeline Park, Angela Oliveira Pisco, Neha Ravikumar, Julia Salzman, Geoff Stanley, Michael Swift, Michelle Tan, Weilun Tan, Alexander J Tarashansky, Rohan Vanheusden, Sevahn K Vorperian, Peter Wang, Sheng Wang, Galen Xing, Chenling Xu, Nir Yosef, Marcela Alcántara-Hernández, Jane Antony, Charles K F Chan, Charles A Chang, Alex Colville, Sheela Crasta, Rebecca Culver, Les Dethlefsen, Camille Ezran, Astrid Gillich, Yan Hang, Po-Yi Ho, Juan C Irwin, Sori Jang, Aaron M Kershner, William Kong, Maya E Kumar, Angera H Kuo, Rebecca Leylek, Shixuan Liu, Gabriel B Loeb, Wan-Jin Lu, Jonathan S Maltzman, Ross J Metzger, Antoine de Morree, Patrick Neuhöfer, Kimberly Perez, Ragini Phansalkar, Zhen Qi, Poorvi Rao, Hayley Raquer-McKay, Koki Sasagawa, Bronwyn Scott, Rahul Sinha, Hanbing Song, Sean P Spencer, Aditi Swarup, Michael Swift, Kyle J Travaglini, Emily Trimm, Stefan Veizades, Sivakamasundari Vijayakumar, Bruce Wang, Wanxin Wang, Juliane Winkler, Jamie Xie, Andrea R Yung, Steven E Artandi, Philip A Beachy, Michael F Clarke, Linda C Giudice, Franklin W Huang, Kerwyn Casey Huang, Juliana Idoyaga, Seung K Kim, Mark Krasnow, Christin S Kuo, Patricia Nguyen, Stephen R Quake, Thomas A Rando, Kristy Red-Horse, Jeremy Reiter, David A Relman, Justin L Sonnenburg, Bruce Wang, Albert Wu, Sean M Wu, and Tony Wyss-Coray. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, May 2022.

[9] Yuanhua Huang and Guido Sanguinetti. BRIE2: computational identification of splicing phenotypes from single-cell transcriptomic experiments. *Genome Biol.*, 22(1):251, August 2021.

[10] Gonzalo Benegas, Jonathan Fischer, and Yun S Song. Robust and annotation-free analysis of alternative splicing across diverse cell types in mice. *Elife*, 11, March 2022.

[11] Carlos F Buen Abad Najar, Prakruthi Burra, Nir Yosef, and Liana F Lareau. Identifying cell state–associated alternative splicing events and their coregulation. *Genome Res.*, July 2022.

[12] Yang I Li, David A Knowles, Jack Humphrey, Alvaro N Barbeira, Scott P Dickinson, Hae Kyung Im, and Jonathan K Pritchard. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.*, 50(1):151–158, January 2018.

[13] Kelsy C Cotto, Yang-Yang Feng, Avinash Ramu, Zachary L Skidmore, Jason Kunisaki, Megan Richters, Sharon Freshour, Yiing Lin, William C Chapman, Ravindra Uppaluri, Ramaswamy Govindan, Obi L Griffith, and Malachi Griffith. RegTools: Integrated analysis of genomic and transcriptomic data for the discovery of splicing variants in cancer. April 2021.

[14] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.*, 112(518):859–877, April 2017.

[15] Carlos F Buen Abad Najar, Nir Yosef, and Liana F Lareau. Coverage-dependent bias creates the appearance of binary splicing in single cells. *Elife*, 9, June 2020.

[16] Alexander B Rosenberg, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T Graybuck, David J Peeler, Sumit Mukherjee, Wei Chen, Suzie H Pun, Drew L Sellers, Bosiljka Tasic, and Georg Seelig. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182, April 2018.

[17] Andras Sziraki, Ziyu Lu, Jasper Lee, Gabor Banyai, Sonya Anderson, Abdulraouf Abdulraouf, Eli Metzner, Andrew Liao, Jason Banfelder, Alexander Epstein, Chloe Schaefer, Zihan Xu, Zehao Zhang, Li Gan, Peter T Nelson, Wei Zhou, and Junyue Cao. A global view of aging and alzheimer's pathogenesis-associated cell population dynamics and molecular signatures in the human and mouse brains. August 2023.

[18] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, April 2014.

[19] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, 19(1):15, February 2018.