Robust and annotation-free analysis of isoform variation using short-read scRNA-seq data

Gonzalo Benegas[1], Jonathan Fischer[2], and Yun S. Song[*3,4,5]

[1]Graduate Group in Computational Biology, University of California, Berkeley, CA 94720
[2]Department of Biostatistics, University of Florida, Gainesville, FL 32611
[3]Department of Statistics, University of California, Berkeley, CA 94720
[4]Computer Science Division, University of California, Berkeley, CA 94720
[5]Chan Zuckerberg Biohub, San Francisco, CA 94158

April 27, 2021

Abstract

Although isoform diversity is acknowledged as a fundamental and pervasive aspect of gene expression in higher eukaryotes, it is often omitted from single-cell studies due to quantification challenges inherent to commonly used short-read sequencing technologies. To address this issue, we have developed a suite of computational tools to investigate isoform variation by focusing on splice junction usage patterns, which can often be well characterized in spite of technical difficulties. Our method, which we name scQuint (single-cell quantification of introns), can perform accurate quantification, dimensionality reduction, and differential splicing analysis using short-read, full-length single-cell RNA-seq data. Notably, scQuint does not require transcriptome annotations and is robust to technical artifacts. In applications across diverse mouse tissues from *Tabula Muris* and the primary motor cortex from the BRAIN Initiative Cell Census Network, we find evidence of strong cell-type-specific isoform variation, complementary to total gene expression, and also identify a large volume of previously unannotated splice junctions. As a community resource, we provide ways to interactively visualize and explore these results, accessible at https://github.com/songlab-cal/scquint-analysis/.

[*]To whom correspondence should be addressed: yss@berkeley.edu

Introduction

Advances in single-cell genomics over the past decade have enabled the data-driven characterization of a wide variety of distinct cell populations. Isoform-level variation in gene expression has often been ignored, despite affecting more than 90% of human genes [1], as its quantification is difficult to carry out using data from popular short-read sequencing technologies such as 10x Genomics Chromium and Smart-seq2 [2]. Long-read single-cell technologies, which greatly simplify isoform quantification, are improving [3, 4, 5, 6, 7], but remain more costly and lower-throughput than their short-read counterparts. Due to experimental considerations, the analysis of isoform variation in 10x Chromium data is mostly restricted to the 3' end of genes; in contrast, Smart-seq2 and other full-length, short-read protocols theoretically enable characterization of isoform variation along the whole gene. Nevertheless, numerous challenges impede such analyses in practice. For example, low transcript capture efficiency introduces additional technical noise into isoform quantification [8, 9, 10], and incomplete transcriptome annotations result in discarded reads and reduced sensitivity to cross-cell differences [9]. Some authors have even recommended avoiding the analysis of alternative splicing in single-cell RNA sequencing (scRNA-seq) data until such obstacles can be suitably overcome [9].

Despite these difficulties, several methods (summarized in Table 1) have sought to analyze isoform variation in short-read, full-length scRNA-seq. Most methods, including kallisto [11], Census [12], BRIE [13], SCATS [14], Quantas [15] and VALERIE [16], depend on transcript annotations and consequently cannot easily identify unannotated alternative splicing events, which may comprise a sizable fraction of events. Annotation-free approaches are also available, though they possess their own limitations. For example, ODEGR-NMF [17] examines disparities in read coverage in fixed-size genomic windows but struggles to detect transcript variation involving small exons. Meanwhile, Expedition [18] and ASCOT [19] only consider binary alternative splicing events such as skipped or mutually exclusive exons, ignoring other highly significant events such as alternate transcription start sites (TSS). SingleSplice [20] and RNA-Bloom [21] propose *de novo* transcript assembly, but do not provide a statistical test for differential transcript usage across conditions. More importantly, each of these methods assumes that read coverage is uniform along a transcript, an assumption we have found often fails to hold for Smart-seq2 data.

In this paper, we introduce scQuint (single-cell quantification of introns), a new suite of tools to facilitate the analysis of isoform variation using short-read full-length scRNA-seq data. Our quantification approach, based on *de novo* junction assembly, captures a wide range of alternative splicing events and accurately identifies novel isoforms. Further, it is robust to coverage biases, a heretofore largely unappreciated factor which we have encountered across several datasets. Lastly, we accompany our quantification approach with efficient methods for dimensionality reduction and statistical differential splicing analysis.

To demonstrate the effectiveness of our approach, we apply scQuint to characterize novel transcript variation in large scRNA-seq data sets—*Tabula Muris* data [22] and primary motor cortex data from the BRAIN Initiative Cell Census Network (BICCN) [23]—spanning a broad range of mouse tissues. We find strong signals of cell-type-specific splicing events complementary to what can be identified by considering total gene expression. Our splicing results, which we make available for visualization via interactive browsers, should be a useful resource for the genomics community.

Method	Full-transcript quantification	Annotation-free	Differential transcript usage	Robust to 3' coverage decay
Quantas			✓	
SingleSplice	✓	✓	*	
kallisto	✓		*	
Census	✓		*	
BRIE			✓	
Expedition		✓		
ODEGR-NMF		✓	*	
SCATS	✓		✓	
VALERIE			✓	
RNA-Bloom	✓	✓		
ASCOT		✓		
scQuint		✓	✓	✓

Table 1: **Summary of methods available to analyze transcript variation in short-read full-length scRNA-seq.** *Full-transcript quantification*: Does the method quantify all forms of transcript variation (excluding RNA editing) or only a subset? *Annotation-free*: Does quantification require an accurate transcriptome reference? *Differential transcript usage*: Does the method provide a two-sample test for differences in transcript proportions? Some methods, denoted by (*), provide other statistical tests. SingleSplice tests for alternative splicing within a single sample. kallisto and ODEGR-NMF test for differential transcript expression, i.e., changes in absolute transcript expression rather than their proportions. Census tests for differential transcript usage along a pseudotime trajectory. *Robust to 3' coverage decay*: Do isoform proportions avoid distortion by decaying read coverage away from the 3' ends of genes?

Results

Robust, annotation-free quantification based on alternative introns

Whereas short reads generally cannot be associated with specific transcript isoforms, nor even exons, if they partially overlap, split reads uniquely associate with a particular intron. Consequently, intron-based quantification does not depend on annotated transcriptome references and allows for novel isoform discovery, a fact exploited by both bulk [24, 25, 26] and single-cell methods [18, 19].

Most methods rely on the assumption that coverage across a transcript is essentially uniform (e.g., *Akr1r1*, Figure S1a). We instead found that Smart-seq2 data [2] frequently contain sizable fractions of genes with coverage that decays as the distance from the 3' ends of transcripts increases. For example, in mammary gland basal cells from the *Tabula Muris* dataset [22], *Ctnbb1* shows a gradual drop in coverage (Figure S1b) while *Pdpn* displays an abrupt reduction halfway through the 3' UTR (Figure S1c). That the magnitude of these effects varies across technical replicates (plates) suggests they could be artifacts, possibly related to degradation or interrupted reverse transcription. Such coverage biases affect gene expression quantification, and in some cases these batch effects are sufficient to comprise a significant proportion of the observed variation in

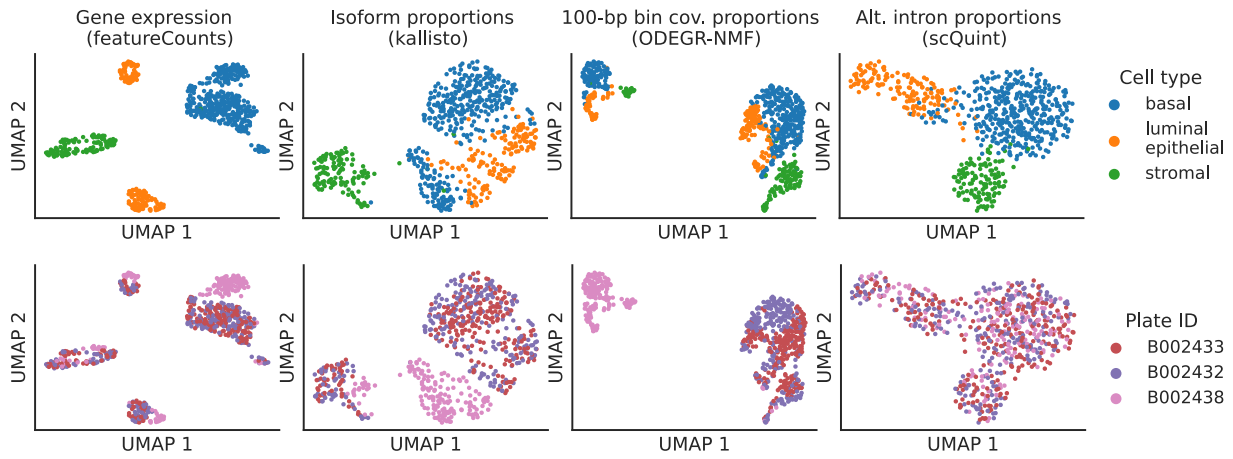

Figure 1: **Clustering patterns by cell type and plate in the mammary gland from a three month-old female mouse in Tabula Muris.** Cell embeddings based on different features were obtained by running PCA followed by UMAP and subsequently colored by cell type (top row) and the plate in which they were processed (bottom row). *First column:* Gene expression, quantified using featureCounts (log-transformed normalized counts), showing clustering primarily by cell type and secondarily by plate. *Second column:* Isoform proportions. Isoform expression was estimated with kallisto and divided by the total expression of the corresponding gene to obtain isoform proportions. There is clustering by both cell type and plate. *Third column:* Coverage proportions of 100 base-pair bins along the gene, as proposed by ODEGR-NMF. There is stronger clustering by plate than by cell type. *Fourth column:* Alternative intron proportions (for introns sharing a 3' acceptor site) as quantified by scQuint, showing clustering by cell type but not plate.

expression levels. For this data set, a low-dimensional embedding of cells based on gene expression reveals that cell type clusters exhibit internal stratification by plate (Figure 1, bottom row). A subsequent test of differential gene expression between plate B002438 and all other plates returns 2,870 significant hits after correction for multiple hypothesis testing, and all manually inspected differentially expressed genes exhibit these types of coverage biases. Perhaps unsurprisingly, quantification at the transcript level is apt to be even more sensitive to these artifacts than gene-level quantification. The UMAP embedding of isoform proportions (as estimated by kallisto [11]) or 100-bp bin coverage proportions (as in ODEGR-NMF [17]) depicts a stronger plate clustering pattern which supersedes the anticipated cell type clusters (Figure 1, second and third columns). Similar coverage bias artifacts are also apparent in the BICCN primary motor cortex data [23] (Figure S2).

With these considerations in mind, we sought to quantify isoform variation in an annotation-free fashion based on alternative intron groups à la MAJIQ [24], JUM [25], and LeafCutter [26]. Critically, we modified their approach by restricting to intron groups that share a common 3' acceptor site (Figure 2). As a result, alternative events are equidistant from polyadenylation sites and will not be affected by the coverage biases we have described above. Indeed, an embedding of cells based on our quantification approach does not suffer from clustering by plate, instead producing cell type clusters with a well-mixed arrangement of cells according to plate (Figure 1, rightmost column). Although alternative intron usage patterns can reveal a wide range of simple and complex alternative splicing events, we note that it cannot generally detect other forms of isoform variation such as intron retention and alternative polyadenylation or transcription start sites (when they do not involve alternative introns).

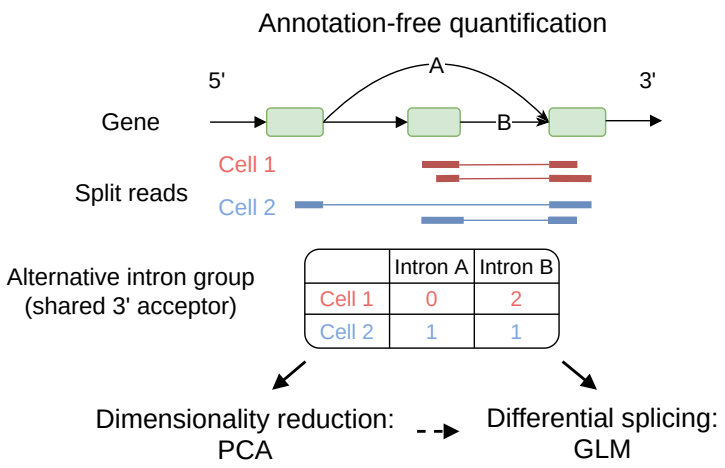

Figure 2: **Overview of scQuint.** Intron usage is quantified from split reads in each cell, with introns sharing 3' splice sites forming alternative intron groups. Genome-wide intron usage is mapped into a low dimensional latent space using PCA after smoothing and transformation to intron proportions. Visualization of the latent space is done via UMAP. A Dirichlet-Multinomial GLM tests for differential splicing across conditions such as predefined cell types or clusters identified from the splicing latent space.

Data set	Cells	Cell types	Genes	Alt. introns	Unannotated
BICCN Cortex	6220	11	26488	39357	29%
Tabula Muris	44518	117	27348	29965	25%

Table 2: **Overview of analyzed data sets.** Number of cells, cell types, detected genes, and detected alternative introns (including the percentage of introns that are not present in the Ensembl reference) for both data sources.

To perform dimensionality reduction using splicing profiles, our method uses Principal Component Analysis (PCA) after smoothing via pseudocounts and transformation of counts to alternative intron proportions. To test for differential splicing across cell types or conditions, we adopt a Dirichlet-Multinomial Generalized Linear Model (GLM) coupled with a likelihood-ratio test (Materials and Methods). This approach has been successfully applied by LeafCutter to bulk RNA-seq, and we find that it also yields calibrated p-values in Smart-seq2 data (Figure S3).

Augmenting cell atlases with splicing information

We applied scQuint to two of the largest available Smart-Seq2 datasets. The first comprehensively surveys the mouse primary motor cortex (*BICCN Cortex*) [23] while the second contains over 100 cell types distributed across 20 mouse organs (*Tabula Muris*) [22] (Table 2). We detect more alternative introns in *BICCN Cortex* neurons than in the entire broad range of cell types present in *Tabula Muris* (which includes neurons but in much smaller number). This observation comports with previous findings that the mammalian brain has exceptionally high levels of alternative splicing [27]. *BICCN Cortex* has already been analyzed at the isoform level [28], while this is only true of microglial cells in *Tabula Muris* [21]. (*Tabula Muris* also contains 10x Chromium data analyzed at the isoform level [29]).

As a community resource, we provide complementary ways to interactively explore splicing patterns present in these data sets (Figure 3), available at https://github.com/songlab-cal/scquint-analysis/. The UCSC Genome Browser [30] permits exploration of alternative splicing events within genomic contexts such as amino acid sequence, conservation score, or protein binding sites, while allowing users to select different length scales for examination. We additionally leverage the cell×gene browser [31] (designed for gene expression analysis) to visualize alternative intron PSI (percent spliced-in, defined as the proportion of reads supporting an intron relative to the total in the intron group) via cell embeddings. Further, one can generate histograms to compare across different groups defined by cell type, gender, or even manually selected groups of cells. These tools remain under active development by the community, and we hope that both the genome- and cell-centric views will soon be integrated into one browser.

Cell-type-specific splicing signal is strong and complementary to gene expression

Primary motor cortex. We first explored the splicing latent space of *BICCN Cortex* cells by comparing it to the usual expression latent space (Figure 4a). Cells in the splicing latent space strongly cluster by cell type (annotated by Yao *et al.* [23] based on gene expression). A similar analysis was recently performed on a different cortex subregion in which most, but not all, neuron subclasses could be distinguished based on splicing profiles (e.g., L6 CT and L6b could not be separated) [33]. However, the authors only considered annotated skipped exons, a subset of the events we quantify, and they used a different dimensionality reduction technique. Figure 4b highlights PSI from some of the top differentially spliced intron groups across cell types, defining "marker introns" as a complement to marker genes (similar to isoform markers defined in [28]). We are able to detect well-known cell-type-specific alternatively spliced genes such as *Nrxn1*, which encodes a key pre-synaptic molecule (Figure 4c) [34]. In this case, we observe an exon (known as splice site 2) exclusively skipped in Vip and Lamp5 neurons. For *Pgm2*, we find a novel exon which is preferentially included in Glutamatergic neurons (Figure 4d). In another example of isoform discovery, we showcase a novel alternative transcription start site (TSS) in *Rbfox1*, a splicing factor known to regulate cell-type-specific alternative splicing in the brain [35] (Figure 4e). This novel TSS (`chr16:5763871-5763913`), which lies in a highly-conserved region, is only (partially) used by L6b neurons.

Marrow B cells. We then focused on developing B cells from the bone marrow in *Tabula Muris*. In the splicing latent space, we found that immature B cells cannot be distinguished from the other B cell subpopulations (Figure 5a). This may suggest they have a less refined splicing program, but it could also result from limitations in transcript capture efficiency. While expression of the transcription factor *Foxp1* peaks in pre-B cells and does not follow a monotonic trend over developmental stages (Figure 5c), we highlight an alternate TSS which is progressively included throughout B cell development (Figure 5b). Combining gene-level expression with TSS usage, which can influence translation rate, provides a more nuanced characterization of *Foxp1* expression patterns.

General patterns across tissues. Because of the breadth of the *Tabula Muris* data set, we can look for general trends across a diverse array of tissues and cell types. Table 3 summarizes differential expression and splicing for some of the cell types and tissues with the largest sample sizes. First, we note the intersection between the top 100 most differentially expressed and top 100 most differentially spliced genes (ranked by p-value) is consistently low. This means that most differentially spliced genes, which might be of critical importance in a biological system, will go unnoticed if a study only considers differential expression. Second, L5 IT neurons have a larger fraction of genes with differential splicing relative to the number of differentially expressed genes.

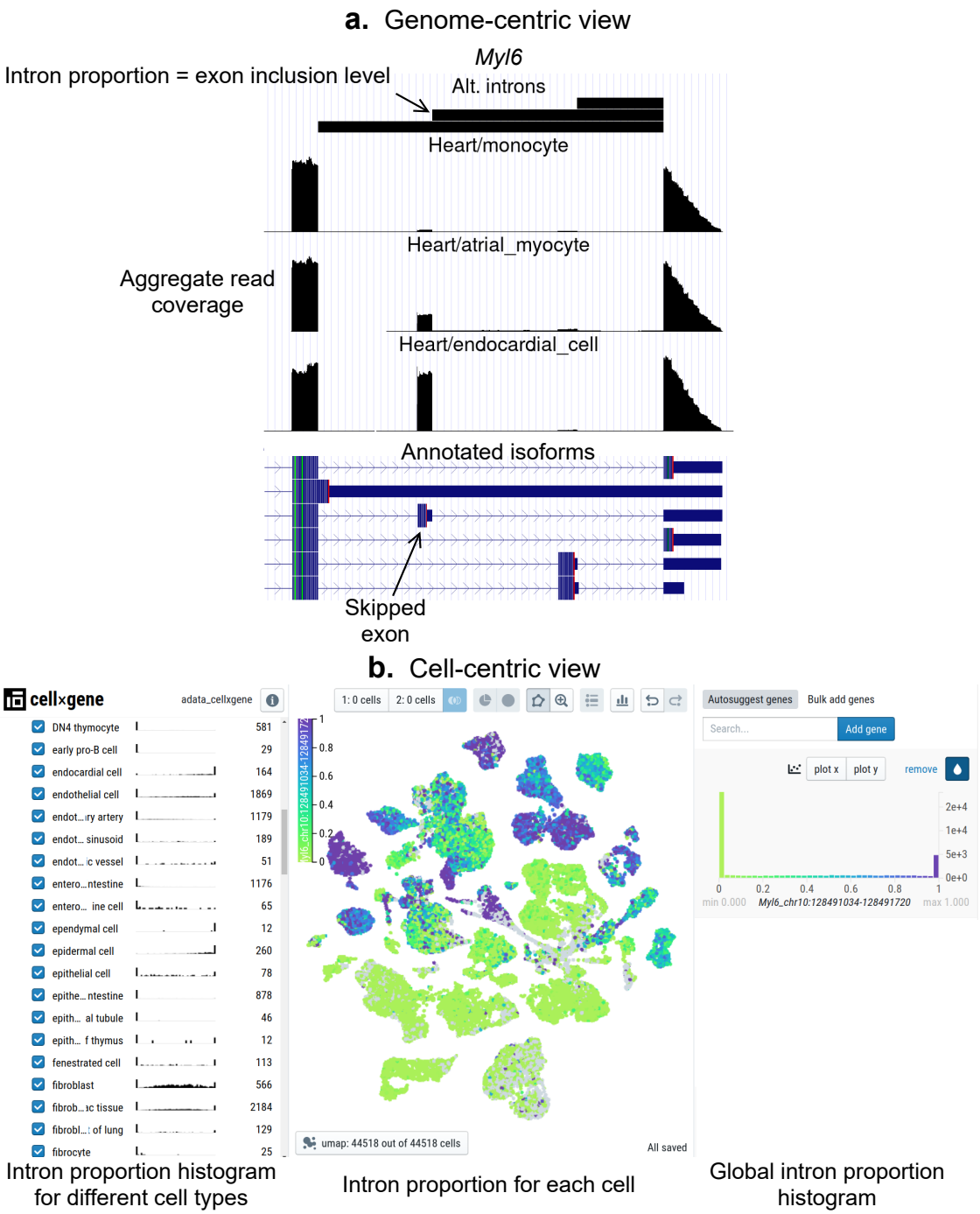

Figure 3: **Interactive visualizations of splicing patterns.** As an example, a skipped exon in *Myl6*. (a) The UCSC Genome browser visualization of this locus. Bottom: annotated isoforms of *Myl6*, including a skipped exon. Center: aggregate read coverage in three cell types with varying inclusion levels of the skipped exon. Top: three alternative introns that share a 3' acceptor site. The identified intron's proportion corresponds to the skipped exon's inclusion level. (b) cell×gene browser visualization of the marked intron's proportions (`Myl6_chr10:128491034-128491720`). Center: intron proportion for each cell in the UMAP expression embedding. Sides: intron proportion histogram for (left) different cell types and (right) all cells.

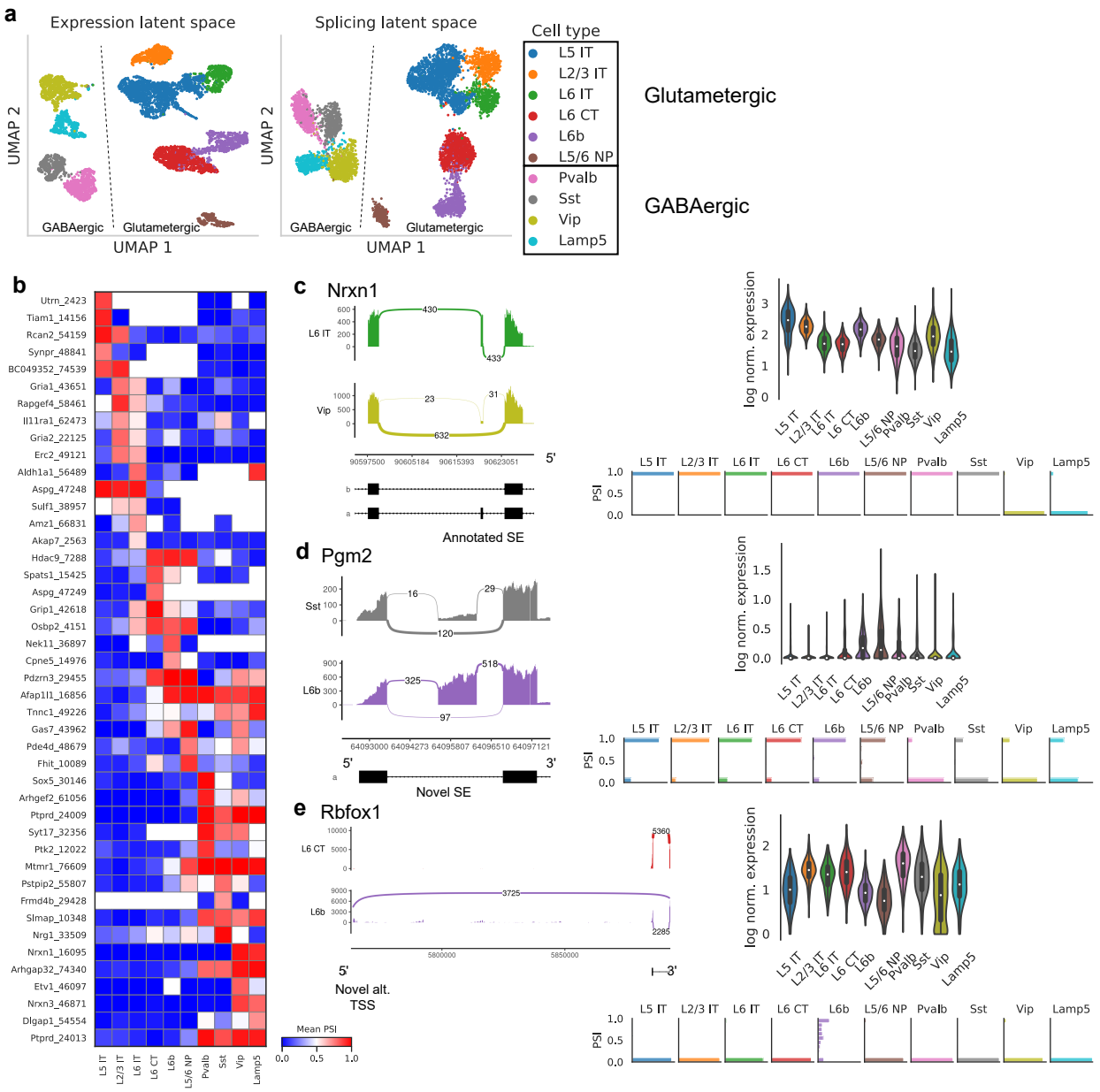

Figure 4: **Splicing in *BICCN Cortex*.** (a) Expression versus splicing latent space, visualized using UMAP. The expression latent space is defined by running PCA on the gene expression matrix, while the splicing latent space is obtained by running PCA on the alternative intron proportions (PSI) matrix. All cell types are well-separated in the splicing latent space. (b) Marker introns' PSIs across cell types, highlighting genes whose splicing patterns distinguish cell types. (c-e) Specific differential splicing events in *Nrxn1*, *Pgm2* and *Rbfox1*, respectively. Left: sashimi plot [32], displaying overall read coverage with arcs indicating usage of different introns. Note: sashimi plots shrink certain introns. Lower right: PSI histogram. Upper right: expression violin plot. (c) An annotated skipped exon (SE) event in *Nrxn1*. The exon is only skipped in Vip and Lamp5 cells. (d) A novel skipped exon event in *Pgm2*. The novel exon is mostly included in Glutamatergic cells and mostly excluded in GABAergic cells. (e) A novel alternative transcription start site (TSS) in *Rbfox1*, only (partially) used in L6b cells.

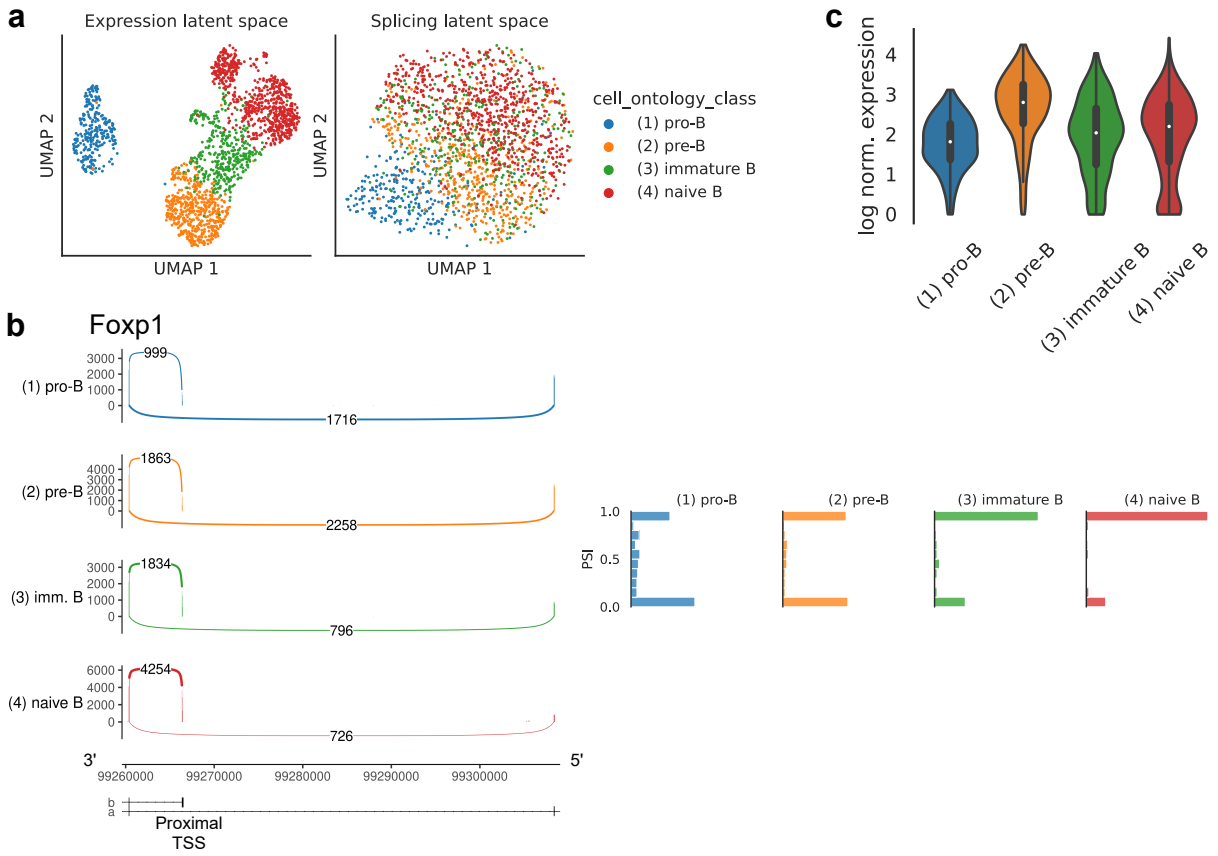

Figure 5: **Splicing in marrow B cells from *Tabula Muris*.** B cell developmental stages include pro-B, pre-B, immature B, and naive B. (a) Expression versus splicing latent space, as defined previously. In the splicing latent space, some cells types (pro-B) are better distinguished than others (immature B). (b) Alternative transcription start site (TSS) in *Foxp1*. Left: sashimi plot, showing read coverage and intron usage. Right: histogram of proportion of the shorter intron (PSI) corresponding to usage of the proximal TSS, increasing as development progresses. (c) Gene expression violin plot of *Foxp1*, showing a non-monotonic trend throughout development.

We found many more cell-type-specific differential splicing events in the cortex than in the marrow, as well as a higher proportion of events involving novel junctions, which can reach 30% (Figure 6a). Most differential splicing events that we detected with alternative introns fall in the coding portion of the gene, with high proportions in the 5' untranslated region (UTR) (Figure 6b). We find an increased proportion of differentially spliced non-coding RNA in the cortex, the majority of which are novel events. To systematically evaluate how well cell types can be distinguished in the expression and splicing latent spaces, we calculated the ROC AUC score for the one-versus-all classification task for each cell type in each tissue using a binary logistic regression model (Figure 6c). Since cell type labels were defined using gene expression values, near-perfect classification is to be expected using the expression latent space. Classification based only on the splicing latent space is very good in general, suggesting that cell-type-specific differential splicing is rather pervasive. A few cell types were more challenging to classify correctly using splicing patterns alone. One such example is immature B cells, a reflection of the lack of clear separation observed in the embedding of Figure 5a.

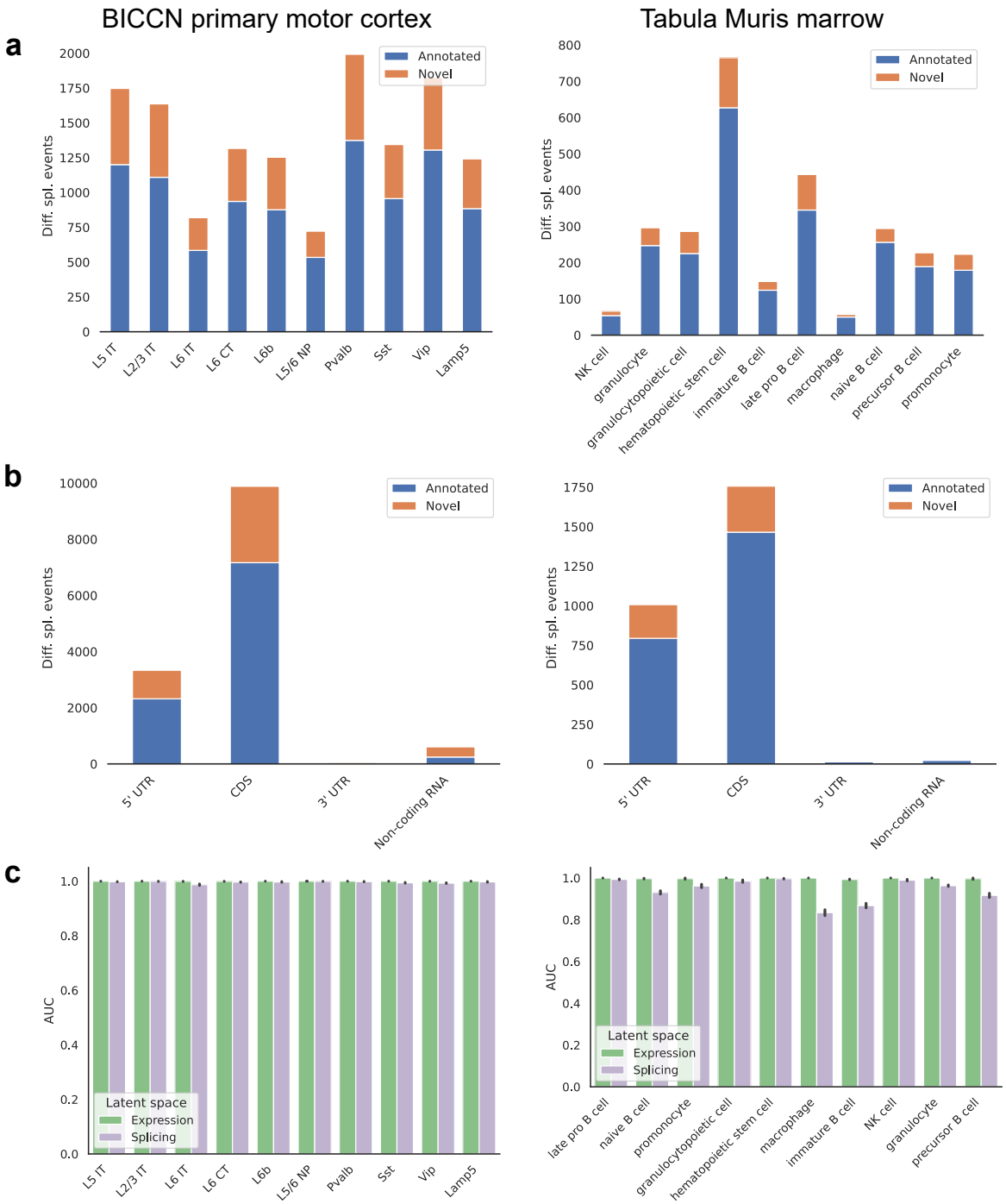

Figure 6: **Patterns across tissues.** (a) Number of differential splicing events detected in each cell type. Cortex cell types have more differential splicing events and larger proportions of novel events (those involving an intron absent from the reference). (b) Number of differential splicing events in different gene regions aggregated over cell types. Cortex cell types have higher proportions of events in coding regions and non-coding RNAs. Note: y-axes are not on the same scale. (c) ROC AUC score for classification of each cell type versus the rest based on either the expression or splicing latent space, using logistic regression. The score for splicing-based classification is near-perfect in most cell types with some exceptions such as immature B cells in the marrow.

Tissue	Total # cells	# cell types	Cell type	# cells	Diff. spl. genes	Diff. exp. genes	Ratio	Top-100 overlap
Brain Non-Myeloid	3049	6	Oligodendrocyte	1390	879	8835	0.10	4
Cortex	6220	10	L5 IT	1571	1442	6402	0.23	1
Heart	4144	6	Endothelial cell of coronary artery	1126	464	7108	0.07	5
Large Intestine	3729	5	Enterocyte of epithelium	1112	585	10786	0.05	2
Marrow	4783	10	Hematopoietic stem cell	1363	692	9909	0.07	2

Table 3: **Summary of differential expression and splicing for select cell types with the largest sample sizes.** The overlap between the top 100 differentially expressed genes and the top 100 differentially spliced genes is low, indicating that splicing provides complementary information. In addition, L5 IT neurons have a higher ratio of differentially spliced genes to differentially expressed genes than the other cell types. *Diff. spl. genes*: number of differentially spliced genes between the cell type and other cell types in the same tissue. *Diff. exp. genes*: number of differentially expressed genes between the cell type and other cell types in the same tissue. See Materials and Methods for details on the tests for differential splicing and expression.

Discussion

In addition to the previously known challenges inherent in the analysis of isoform variation in single cells using short reads, we report here the presence of substantial coverage biases in Smart-seq2 data. Until these biases are better understood, including potential interactions between technical and biological factors, we advocate quantifying transcript variation via alternative intron usage conditioned on sharing the 3' acceptor site. Although this limits the scope of transcript variation one can investigate, we demonstrated that this subset is robust to coverage biases and easily detects novel isoforms. We also successfully adapted existing approaches for dimensionality reduction and differential splicing analysis, illustrating their suitability for single-cell data. As an aside, we note that construction of the splicing latent space using a Variational Auto-Encoder [36] rather than PCA did not result in substantially different clustering patterns.

We successfully applied our methods to data sets consisting of diverse mouse tissues, finding an overall strong cell-type-specific splicing signal. This result, together with the fact that the most strongly differentially expressed and differentially spliced genes do not generally overlap (Table 3), should encourage tissue specialists to further study the contribution of isoform variation to cell identity and its role in cell and tissue homeostasis. Our analyses have not yet produced instances in which clustering in the splicing latent space reveals new cell subpopulations not visible in the expression latent space. This, of course, does not preclude the possibility in other settings where alternative splicing is known to be important, such as in specific developmental transitions or disease conditions. Nevertheless, our current experience leads us to believe that gene expression and splicing proportions are two different projections of the same underlying cell state.

An important follow-up direction is the inference of joint transcriptional and post-transcriptional regulatory networks based on gene- and isoform-level co-expression. While this has been attempted using bulk RNA-seq samples [37], information from single cells provides a natural perturbation system and improves statistical power. Another outstanding question is the functional significance of isoforms. This requires having a more precise understanding of how variation in UTRs affects final protein output as well as the biophysical characteristics of protein isoforms and their roles in different biological systems. Incidentally, RNA Velocity [38] estimates can be distorted by alternative splicing, and Bergen *et al.* [39] discuss incorporating isoform proportions into the

model as a future direction.

Several experimental advances will facilitate the study of isoform variation in single cells. Smart-seq3 [40] allows sequencing of short reads from the entire length of a gene together with unique molecular identifiers. While this protocol improves mRNA capture and allows for the filtering of PCR duplicates, experiments show that less than 40% of reads can be unambiguously assigned to a single (annotated) isoform. Ultimately, long-read scRNA-seq will provide the definitive picture of isoform variation between cells. Until then, there is much biology to be studied using short-read protocols, and variation at the isoform level should not be disregarded. Despite the challenging nature of these data, our method showcases how the analysis of isoform variation in short-read full-length scRNA-seq can be improved with annotation-free methods paying special attention to technical artifacts.

Materials and Methods

Data sets. *Tabula Muris* data have accession code GSE109774. Cells were filtered to those from three month-old mice present in this collection: https://czb-tabula-muris-senis.s3-us-west-2.amazonaws.com/Data-objects/tabula-muris-senis-facs-processed-official-annotations.h5ad (filtering details in [41]). *BICCN Cortex* data were downloaded from http://data.nemoarchive.org/biccn/lab/zeng/transcriptome/scell/SMARTer/raw/MOp/ and filtered as in [28].

Quantification. The bioinformatic pipeline was implemented using Snakemake [42]. Raw reads were trimmed from Smart-Seq2 adapters using Cutadapt [43] before mapping to the GRCm38/mm10 genome reference (https://hgdownload.soe.ucsc.edu/goldenPath/mm10/chromosomes/) and the transcriptome reference from Ensembl release 101 (ftp://ftp.ensembl.org/pub/release-101/gtf/mus_musculus/Mus_musculus.GRCm38.101.gtf.gz). Alignment was done using STAR [44] two-pass mode allowing novel junctions as long as they were supported by reads with at least 20 base pair overhang (30 if they are non-canonical) in at least 30 cells. Also, multimapping and duplicate reads were discarded using the flag `--bamRemoveDuplicatesType UniqueIdentical` (while this can remove duplicates from the second PCR step of Smart-seq, it will not remove duplicates from the first PCR step). Soft-clipped reads were removed as well. Additionally, reads were discarded if they belonged to the ENCODE region blacklist [45] (downloaded from https://github.com/Boyle-Lab/Blacklist/raw/master/lists/mm10-blacklist.v2.bed.gz).

Gene expression was quantified using featureCounts [46], and total-count normalized such that each cell had 10,000 reads (as in the Scanpy [47] tutorial). Intron usage was quantified using split reads with an overhang of at least 6 base pairs. Introns were discarded if observed in fewer than 30 cells in *BICCN Cortex* or 100 cells in *Tabula Muris*. Introns were grouped into alternative intron groups based on shared 3' splice acceptor sites. Introns not belonging to any alternative intron group were discarded. Additionally, we decided to subset our analysis to introns with at least one of their donor or acceptor sites annotated, so we could assign a gene to them and facilitate interpretation for our specific analyses.

Dimensionality reduction. To define the splicing latent space, we worked with alternative intron proportions (PSI) rather than their absolute counts, as the latter would be confounded by gene expression differences. However, given the sparsity of single-cell data, a very high proportion of alternative intron groups will have no reads in a given cell, leaving PSI undefined. More generally, an intron group may contain few reads, resulting in defined but noisy PSI estimates. To navigate this issue, we introduce a form of empirical shrinkage towards a central value. We first define the "global PSI" by aggregating reads from all cells and normalizing. Then, we add this global PSI as

a pseudocount vector to each cell before re-normalizing to obtain each cell's shrunken PSI profile (these are non-uniform pseudocounts adding up to one). We then run standard PCA on the cell-by-intron-smoothed PSI matrix. To verify that we prevent leakage of gene expression information into our splicing profiles, we run this procedure on a perturbed data set where intron counts within each intron group are redistributed with a fixed probability in all cells. This shuffled data set contains expression variability between cells but no splicing differences, and the splicing latent space does not distinguish among cell types (Figure S4).

Differential splicing test. Our differential splicing test across conditions (such as cell types) is based on a Dirichlet-Multinomial Generalized Linear Model which was proposed in LeafCutter [26] for bulk RNA-seq. For each intron group g with L alternative introns:

- $\vec{y}_g$ is a vector of counts for each of the L introns;
- The independent variable, x, equals 0 in one condition and 1 in the other;
- $\vec{a}_g, \vec{b}_g \in \mathbb{R}^{L-1}$ are the intercept and coefficients of the linear model;
- $\alpha_g \in \mathbb{R}$ is a dispersion parameter shared across conditions; and
- the function $\text{softmax} : (z_1, \ldots, z_{L-1}) \mapsto \left(\frac{e^{z_1}}{1+\sum_{i=1}^{L-1} e^{z_i}}, \ldots, \frac{e^{z_{L-1}}}{1+\sum_{i=1}^{L-1} e^{z_i}}, \frac{1}{1+\sum_{i=1}^{L-1} e^{z_i}} \right)$ maps from $\mathbb{R}^{L-1}$ to the $(L-1)$-dimensional probability simplex.

The Dirichlet-Multinomial Generalized Linear Model then proceeds as follows:

1. Obtain the underlying intron proportions:
   $\vec{p}_g = \text{softmax}(\vec{a}_g + \vec{b}_g x)$
2. Sample the intron counts conditioned on the total observed, n_g:
   $\vec{y}_g | n_g \sim \text{DirichletMultinomial}\,(n_g, \alpha_g \vec{p}_g)$

We implemented this model in PyTorch and optimized it using L-BFGS [48].

To test for differential splicing across the two conditions, we compare the following two hypotheses:

$$\text{Null hypothesis } H_0\text{: } \vec{b}_g = \vec{0}$$
$$\text{Alternative hypothesis } H_1\text{: } \vec{b}_g \neq \vec{0}$$

We use the likelihood-ratio test, the test statistic for which is asymptotically distributed as a χ^2 random variable with $L-1$ degrees of freedom under H_0. Finally, we correct p-values for multiple testing using the Benjamini-Hochberg FDR procedure [49].

Analysis. The expression latent space was obtained by running PCA with 40 components on log-transformed and normalized gene expression values. The splicing latent space was obtained by running PCA with 20 components on the smoothed PSI matrix. Both latent spaces were visualized using UMAP [50]. Reported scores for cell type classification within a tissue were obtained by running a binary logistic regression classifier over 30 different random (stratified) splits with 2/3 of the data for training and 1/3 for testing.

For differential splicing testing between a given cell type and the rest of the tissue, we only considered introns expressed in at least 50 cells and intron groups with at least 50 cells from both of the conditions. We called an intron group "differentially spliced" if it was both statistically

significant using a 5% FDR and if it contained an intron with a PSI change greater than 0.05. We considered a differentially spliced intron group as unannotated if it contained an unannotated intron with a PSI change greater than 0.05. Differential expression was performed using the Mann-Whitney test. A gene was considered differentially expressed if it was statistically significant using a 5% FDR and if the fold change was at least 1.5.

Code and data availability. scQuint Python implementation is available at https://github.com/songlab-cal/scquint. Differential splicing results and access to cell and genome browsers, together with code to reproduce results, are available at https://github.com/songlab-cal/scquint-analysis. Processed alternative intron count matrices are provided in the AnnData format (anndata.readthedocs.io) for easy manipulation with Scanpy [47] and other tools.

Acknowledgments

We would like to thank Angela Oliveira Pisco, Spyros Darmanis, and Kif Liakath-Ali for helpful discussions. We are also thankful to Aaron McGeever and the Chan Zuckerberg Biohub for hosting the cell×gene browsers. This research is supported in part by grant number R35-GM134922 from NIH and grant number CZF2019-002449 from the Chan Zuckerberg Initiative Foundation. Y.S.S. is a Chan Zuckerberg Biohub Investigator.

References

[1] Wang ET, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456(7221):470–476.

[2] Picelli S, et al. (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* 9(1):171.

[3] Byrne A, et al. (2017) Nanopore long-read RNA-seq reveals widespread transcriptional variation among the surface receptors of individual b cells. *Nature Communications* 8(1):1–11.

[4] Gupta I, et al. (2018) Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nature Biotechnology* 36(12):1197–1202.

[5] Volden R, Vollmers C (2020) Highly multiplexed single-cell full-length cdna sequencing of human immune cells with 10x genomics and R2C2. *BioRxiv*.

[6] Lebrigand K, Magnone V, Barbry P, Waldmann R (2020) High throughput error corrected nanopore single cell transcriptome sequencing. *Nature Communications* 11(1):1–8.

[7] Joglekar A, et al. (2021) A spatially resolved brain region-and cell type-specific isoform atlas of the postnatal mouse brain. *Nature Communications* 12(1):1–16.

[8] Arzalluz-Luque Á, Conesa A (2018) Single-cell RNAseq for the study of isoforms–how is that possible? *Genome Biology* 19(1):110.

[9] Westoby J, Artemov P, Hemberg M, Ferguson-Smith A (2020) Obstacles to detecting isoforms using full-length scRNA-seq data. *Genome Biology* 21(1):1–19.

[10] Najar CFBA, Yosef N, Lareau LF (2020) Coverage-dependent bias creates the appearance of binary splicing in single cells. *eLife* 9:e54603.

[11] Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* 34(5):525–527.

[12] Qiu X, et al. (2017) Single-cell mrna quantification and differential analysis with census. *Nature Methods* 14(3):309–315.

[13] Huang Y, Sanguinetti G (2017) Brie: transcriptome-wide splicing quantification in single cells. *Genome Biology* 18(1):123.

[14] Hu Y, Wang K, Li M (2020) Detecting differential alternative splicing events in scrna-seq with or without unique molecular identifiers. *PLOS Computational Biology* 16(6):e1007925.

[15] Yan Q, et al. (2015) Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals nmd modulating chromatin regulators. *Proceedings of the National Academy of Sciences* 112(11):3445–3450.

[16] Wen WX, Mead AJ, Thongjuea S (2020) Valerie: Visual-based inspection of alternative splicing events at single-cell resolution. *PLOS Computational Biology* 16(9):e1008195.

[17] Matsumoto H, et al. (2020) An nmf-based approach to discover overlooked differentially expressed gene regions from single-cell RNA-seq data. *NAR Genomics and Bioinformatics* 2(1):lqz020.

[18] Song Y, et al. (2017) Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Molecular Cell* 67(1):148–161.

[19] Ling JP, et al. (2020) ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications* 11(1):1–12.

[20] Welch JD, Hu Y, Prins JF (2016) Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Research* 44(8):e73–e73.

[21] Nip KM, et al. (2020) RNA-Bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes. *Genome Research* 30(8):1191–1200.

[22] Schaum N, et al. (2018) Single-cell transcriptomics of 20 mouse organs creates a tabula muris: The tabula muris consortium. *Nature* 562(7727):367.

[23] Yao Z, et al. (2020) An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. *Biorxiv*.

[24] Vaquero-Garcia J, et al. (2016) A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife* 5:e11752.

[25] Wang Q, Rio DC (2018) JUM is a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. *Proceedings of the National Academy of Sciences* 115(35):E8181–E8190.

[26] Li YI, et al. (2018) Annotation-free quantification of RNA splicing using leafcutter. *Nature Genetics* 50(1):151.

[27] Yeo G, Holste D, Kreiman G, Burge CB (2004) Variation in alternative splicing across human tissues. *Genome Biology* 5(10):1–15.

[28] Booeshaghi A, et al. (2020) Isoform specificity in the mouse primary motor cortex. *bioRxiv*.

[29] Patrick R, et al. (2020) Sierra: Discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *bioRxiv* p. 867309.

[30] Kent WJ, et al. (2002) The human genome browser at ucsc. *Genome Research* 12(6):996–1006.

[31] Megill C, et al. (2021) cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. *bioRxiv*.

[32] Garrido-Martín D, Palumbo E, Guigó R, Breschi A (2018) ggsashimi: Sashimi plot revised for browser-and annotation-independent splicing visualization. *PLoS Computational Biology* 14(8):e1006360.

[33] Feng H, et al. (2021) Complexity and graded regulation of neuronal cell-type–specific alternative splicing revealed by single-cell RNA sequencing. *Proceedings of the National Academy of Sciences* 118(10).

[34] Fuccillo MV, et al. (2015) Single-cell mrna profiling reveals cell-type-specific expression of neurexin isoforms. *Neuron* 87(2):326–340.

[35] Wamsley B, et al. (2018) Rbfox1 mediates cell-type-specific splicing in cortical interneurons. *Neuron* 100(4):846–859.

[36] Kingma DP, Welling M (2013) Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[37] Saha A, et al. (2017) Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Research* 27(11):1843–1858.

[38] La Manno G, et al. (2018) Rna velocity of single cells. *Nature* 560(7719):494–498.

[39] Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ (2020) Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology* pp. 1–7.

[40] Hagemann-Jensen M, et al. (2020) Single-cell RNA counting at allele and isoform resolution using smart-seq3. *Nature Biotechnology* 38(6):708–714.

[41] Pisco AO, et al. (2019) A single cell transcriptomic atlas characterizes aging tissues in the mouse. *bioRxiv* p. 661728.

[42] Köster J, Rahmann S (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28(19):2520–2522.

[43] Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17(1):10–12.

[44] Dobin A, et al. (2013) Star: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15–21.

[45] Amemiya HM, Kundaje A, Boyle AP (2019) The ENCODE blacklist: identification of problematic regions of the genome. *Scientific Reports* 9(1):1–5.

[46] Liao Y, Smyth GK, Shi W (2014) featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30(7):923–930.

[47] Wolf FA, Angerer P, Theis FJ (2018) Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology* 19(1):1–5.

[48] Liu DC, Nocedal J (1989) On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 45(1-3):503–528.

[49] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1):289–300.

[50] McInnes L, Healy J, Melville J (2018) Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Supplementary Figures

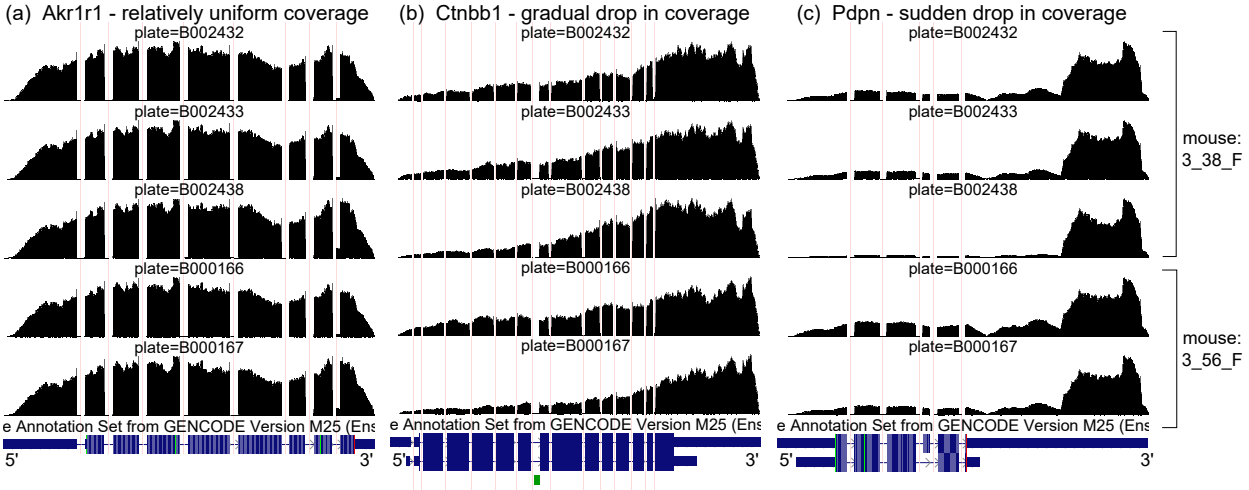

Figure S1: **Coverage artifacts in mammary gland basal cells from *Tabula Muris*.** Aggregate read coverage of basal cells is shown for three genes in two female mice: *3_38_F*, processed in three different plates, and *3_56_F*, processed in two different plates. Visualization on the UCSC Genome Browser. (a) *Akr1r1*, with relatively uniform coverage, what we expect. (b) *Ctnbb1*, with a gradual drop in coverage away from the 3' end. The rate of coverage decay varies across plates. (c) *Pdpn*, with a sudden drop in coverage halfway through the 3' UTR. The magnitude of the drop varies across plates.

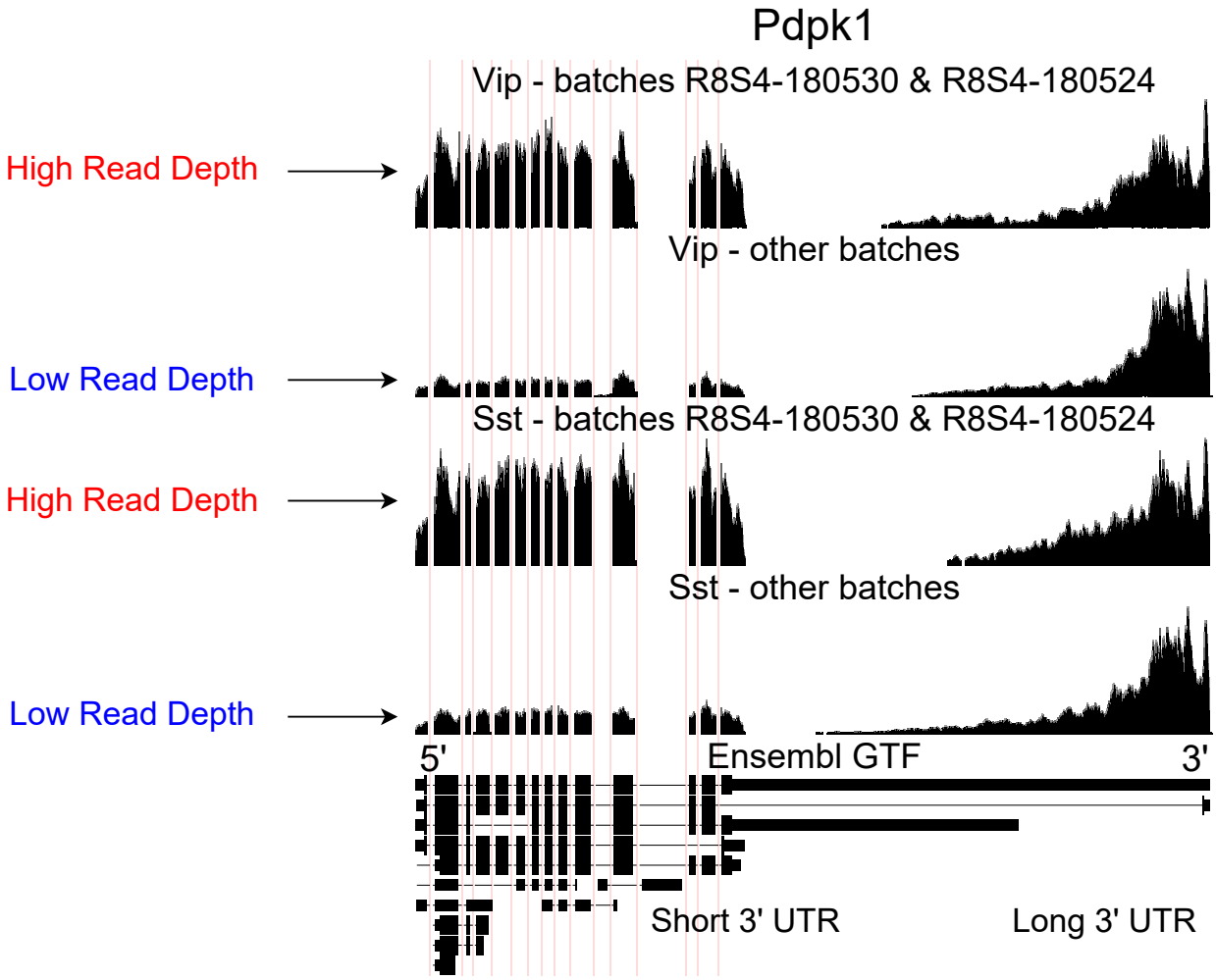

Figure S2: **Technical artifacts in *BICCN Cortex*.** Aggregate read coverage in *Pdpk1* in two cell types, Vip and Sst, further separated according to the "batch" metadata label into two groups. The first group contains cells from batches *R8S4-180530* and *R8S4-180524*, while the second group contains the remaining batches. Cells from different batches belong to different mice and were processed on different dates. In all groups, coverage decreases rapidly in the 3' UTR of the isoform with the longest 3' UTR, eventually reaching zero. Additionally, the relative coverage in this region compared to the rest of the gene (seeming to originate from a different isoform with a short 3' UTR) varies drastically across batches, and consistently in different cell types. In principle, this could be due to biological differences between mice from different batches, but an explanation based on technical factors such as amplification bias may be more plausible.

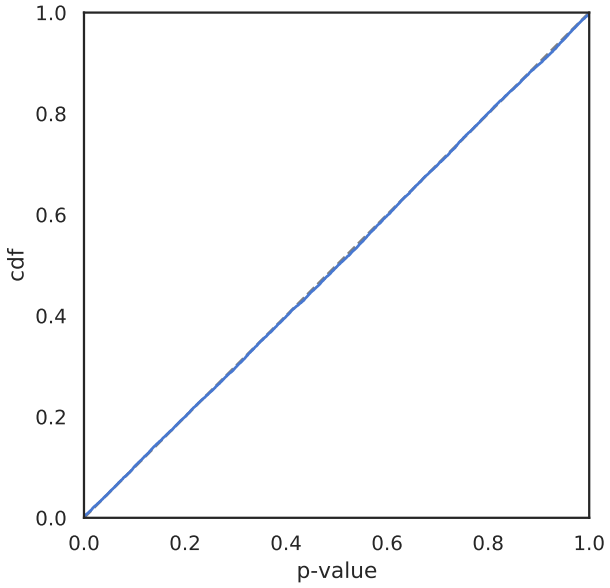

Figure S3: **Distribution of p-values under the null.** 6220 *BICCN Cortex* cells were randomly split into two equally sized groups and our differential splicing test performed. In this scenario, the null hypothesis of no difference in splicing proportions holds, and we expect the distribution of p-values to be uniform. The cumulative distribution function of p-values shows their distribution is indeed uniform, suggesting that the model is well-calibrated under the null.

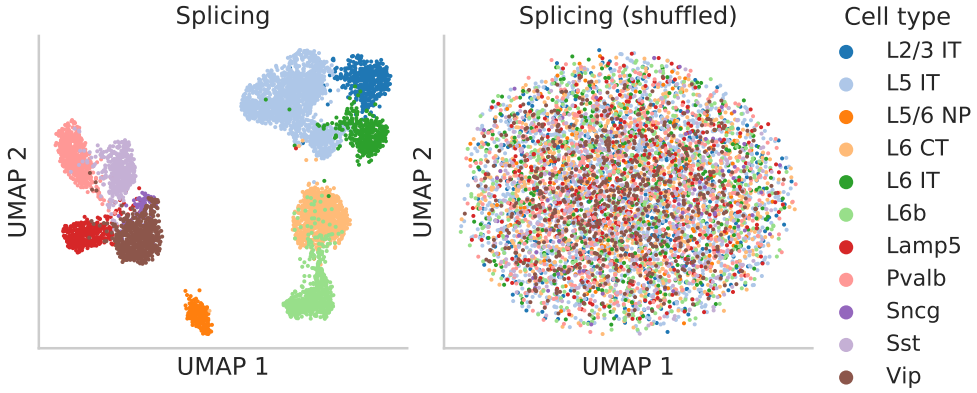

Figure S4: **Splicing latent space when alternative intron counts are shuffled.** To ensure that absolute gene expression does not affect the splicing latent space, we perturbed the *BICCN Cortex* data set by resampling alternative intron counts with a fixed proportion in all cells (the proportion in different alternative intron groups varied and was sampled from a uniform Dirichlet distribution). In this scenario, different cell types still vary in their gene expression but not in their splicing patterns. As hoped, the splicing latent space does not distinguish between cell types, meaning it is only capturing differences in splicing proportions and not changes in absolute gene expression.