Robust and annotation-free analysis of alternative splicing 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

Abstract

Although alternative splicing is 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. Here, we propose a new computational method, scQuint, that performs accurate quantification, dimensionality reduction, and differential splicing analysis using full-length, short-read, single-cell RNA-seq data. 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 alternative splicing, complementary to total gene expression, and identify a large volume of previously unannotated splice junctions. To further elucidate the regulation of alternative splicing, we build a predictive model based on splicing factor activity, which recovers several known interactions and generates new hypotheses, including potential regulatory roles of novel alternative splicing events in *Khdrbs3* and *Rbfox1*.

*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. Despite affecting more than 90% of human pre-mRNAs [1], isoform-level variation in gene expression has often been ignored because it's difficult to quantify 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. Many methods, including kallisto [11], Census [12], BRIE [13], SCATS [14], Quantas [15], and VALERIE (meant only for visualization) [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. DESJ [22] quantifies the proportion of an intron relative to all introns in a gene, but we find it is greatly affected by coverage biases, as are other methods based on global variation across the gene.

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 annotation-free quantification approach, based on *de novo* junction assembly, captures a wide range of alternative splicing events and accurately identifies previously unannotated isoforms. Moreover, the local rather than global nature of our quantification scheme makes it more robust to coverage biases, a heretofore largely unappreciated mode of technical variation which we have encountered across several data sets. We accompany our quantification approach with computationally efficient methods for further analysis and interpretation via dimensionality reduction and hypothesis testing. Specifically, we implement a custom variational autoencoder which empirically improves upon the dimensionality reductions produced by PCA, and we also provide a statistically principled test of differential splicing based on the Dirichlet-Multinomial distribution.

To demonstrate its effectiveness and scalability, we apply scQuint to characterize novel transcript variation in large, atlas-sized scRNA-seq data sets from the *Tabula Muris* consortium [23] and BRAIN Initiative Cell Census Network (BICCN) [24] which span 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. We also advance the understanding of alternative splicing regulation by building a predictive model incorporating the activity of splicing factors. Our splic-

Method	Isoform quantification	Annotation-free	Differential transcript usage
Quantas [15]	Local		✓
SingleSplice [20]	Global	✓	*
kallisto [11]	Global		*
Census [12]	Global		*
BRIE [13]	Local		✓
Expedition [18]	Local	✓	
ODEGR-NMF [17]	Global	✓	*
SCATS [14]	Local		✓
RNA-Bloom [21]	Global	✓	
ASCOT [19]	Local	✓	
DESJ [22]	Global	✓	✓
scQuint	Local	✓	✓

Table 1: **Summary of methods available to analyze transcript variation in short-read full-length scRNA-seq.** *Isoform quantification*: Does the method quantify global transcript variation or local alternative events? *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.

ing results, which we make available for visualization via interactive browsers, should be a useful resource for the genomics community.

Results

Methods overview

Robust, annotation-free quantification based on alternative introns. Most methods rely on the assumption that coverage depth 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 with increasing distance from the 3' ends of transcripts. For example, in mammary gland basal cells from the *Tabula Muris* data set [23], *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. Similar coverage bias artifacts are also apparent in the BICCN primary motor cortex data [24] (Figure S2).

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 expression levels. For the *Tabula Muris* mammary gland data set, a low-dimensional embedding of cells based on gene

expression reveals that cell type clusters exhibit internal stratification by plate (Figure 1a). 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, especially if it is based on coverage differences across the whole length of the transcript (denoted as *global* isoform quantification in Table 1). The UMAP embedding of isoform proportions (kallisto [11]), exon proportions (DEXSeq [25]), 100 bp bin coverage proportions (ODEGR-NMF [17]) or junction usage proportions across the whole gene (DESJ [22]) depicts a stronger plate clustering pattern which supersedes the anticipated cell type clusters (Figure 1b-e).

With these considerations in mind, we sought to quantify isoform variation in a local manner that would be more robust to coverage differences along the transcript. One such approach is alternative intron quantification, as performed by bulk RNA-seq methods MAJIQ [26], JUM [27] and LeafCutter [28]. Promisingly, quantification via LeafCutter (Figure 1f) yields an embedding that displays less clustering by plate than global quantifications such as kallisto or DESJ. Inspired by LeafCutter, our method scQuint also quantifies alternative introns but restricts to those sharing a common 3' acceptor site (Figure 2). As a result, alternative events are equidistant from the 3' end of transcripts and are not affected by the coverage biases described above. The embedding of cells based on our quantification approach (Figure 1g) shows less clustering by plate than LeafCutter and other methods. Some clustering by plate can still be seen in luminal epithelial cells, which could be caused by other kinds of technical artifacts such as sequence bias. As we describe later, scQuint also provides more accurate differential splicing p-values compared to LeafCutter.

Another advantage of alternative intron quantification is the ability to easily discover novel isoforms. 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. This is important since, as detailed later, we estimate up to 30% of novel cell-type-specific differential splicing events. Expedition [18] and ASCOT [19] also perform annotation-free quantification based on local intron usage, but they consider only simple binary events and ignore other important events such as alternative transcription start sites, which we find to be plentiful in our analysis. Moreover, neither Expedition nor ASCOT provide a statistical test for differential splicing across cell subpopulations.

Dimensionality reduction with Variational Autoencoder. To perform dimensionality reduction using splicing profiles, we developed a novel Variational Autoencoder (VAE) [29] with a Dirichlet-Multinomial noise model (Figure 2b, Materials and Methods). VAEs are flexible and scalable generative models which have been successfully applied to analyze gene expression [30] but have not yet been employed to investigate alternative splicing. We compared the latent space obtained with the VAE to the one obtained using Principal Component Analysis (PCA), a standard dimensionality reduction technique used in the LeafCutter software package. The VAE is able to better distinguish cell types than can PCA (Figure 3). While cortex cells are already quite well separated in the PCA latent space, the increased resolution offered by the VAE is better appreciated in other tissues such as mammary gland and diaphragm.

Differential splicing hypothesis testing with Generalized Linear Model. 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 (Figure 2c, Materials and Methods). We leverage one of the models proposed by LeafCutter for bulk RNA-seq, apply it to our Smart-seq2 intron quantification, and find that it yields well-calibrated p-values (Figure S3). LeafCutter's

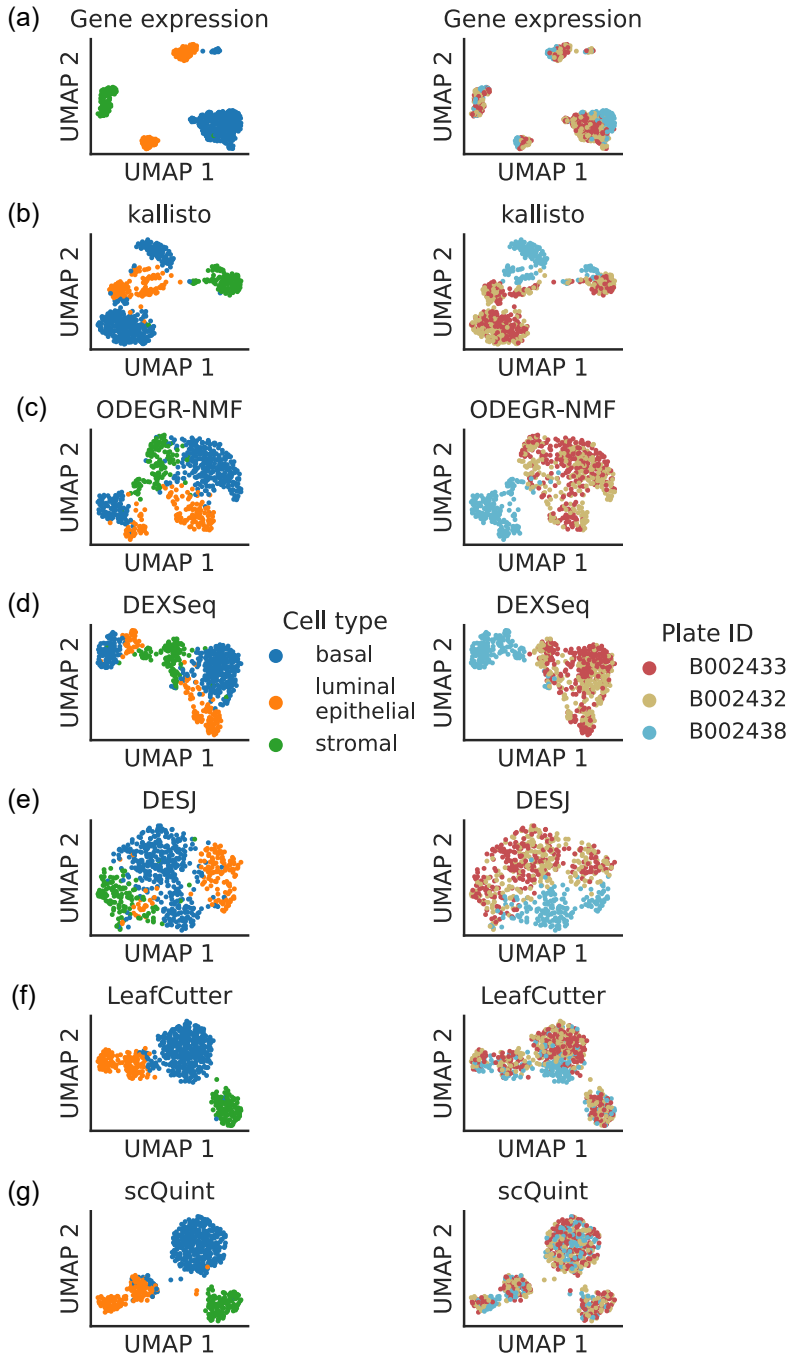

Figure 1: **Clustering patterns by cell type and plate in the mammary gland from a three month-old female mouse in Tabula Muris.** Cell embeddings based on different features were obtained by running PCA (gene expression) or VAE (the rest) followed by UMAP and subsequently colored by cell type (left column) and the plate in which they were processed (right column). **(a)** Gene expression, quantified using featureCounts (log-transformed normalized counts). **(b)** Isoform proportions. Isoform expression was estimated with kallisto and divided by the total expression of the corresponding gene to obtain isoform proportions. **(c)** Coverage proportions of 100 base-pair bins along the gene, as proposed by ODEGR-NMF. **(d)** Exon proportions, as proposed by DEXSeq. **(e)** Intron proportions across the whole gene, as proposed by DESJ. **(f)** Alternative intron proportions quantified by LeafCutter. **(g)** Alternative intron proportions (for introns sharing a 3' acceptor site) as quantified by scQuint.

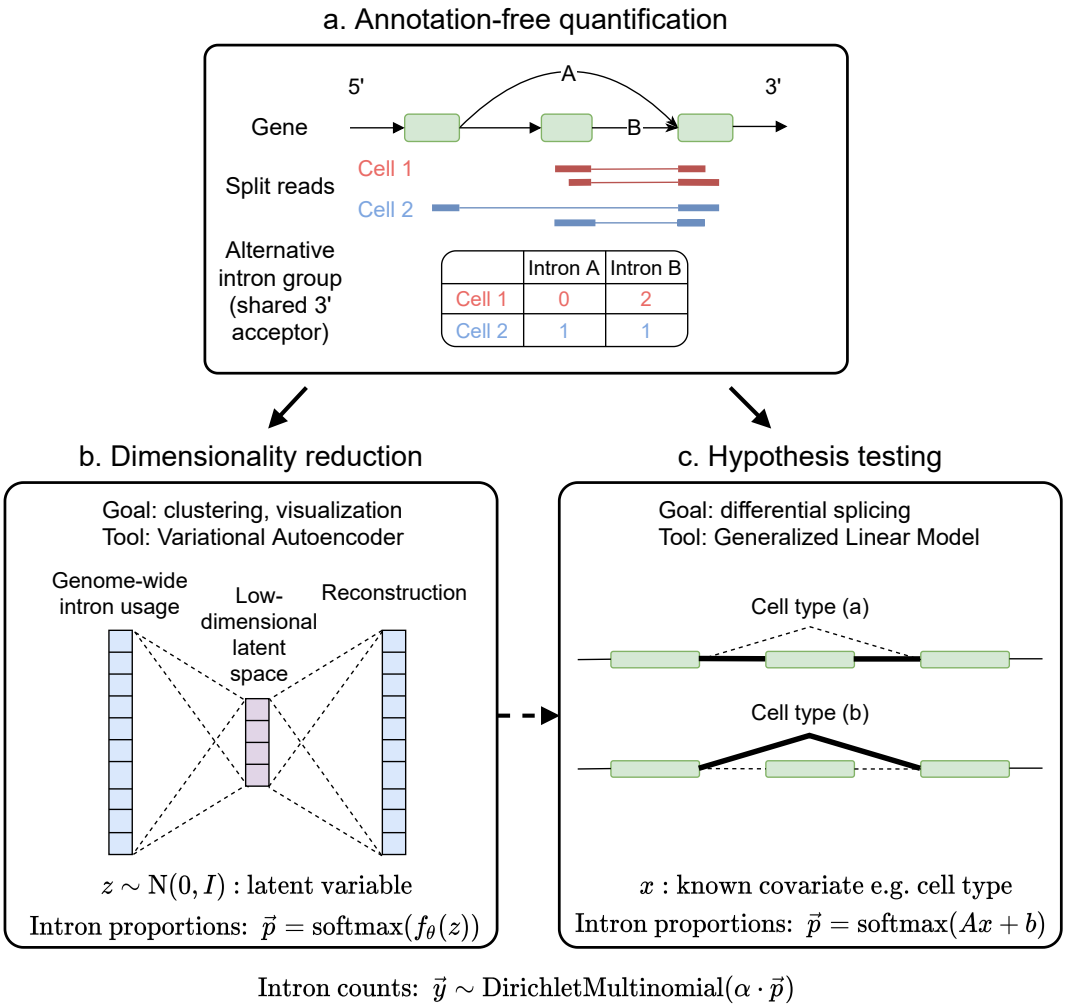

Figure 2: **Overview of scQuint. (a)** Intron usage is quantified from split reads in each cell, with introns sharing 3' splice sites forming alternative intron groups. **(b)** Genome-wide intron usage is mapped into a low dimensional latent space using a Dirichlet-Multinomial VAE. Visualization of the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

differential splicing p-values, however, are not as well calibrated (Figure S3), which could be due to differences in the model or its optimization. Since LeafCutter's p-values are anti-conservative, note that it can lead to more false positives than expected.

Augmenting cell atlases with splicing information

We applied scQuint to two of the largest available Smart-Seq2 data sets. The first comprehensively surveys the mouse primary motor cortex (*BICCN Cortex*) [24] while the second contains over 100 cell types distributed across 20 mouse organs (*Tabula Muris*) [23] (Table S1). 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 [31]. Booeshagi et al. [32] analyzed *BICCN Cortex* at the transcript level , but focused on changes

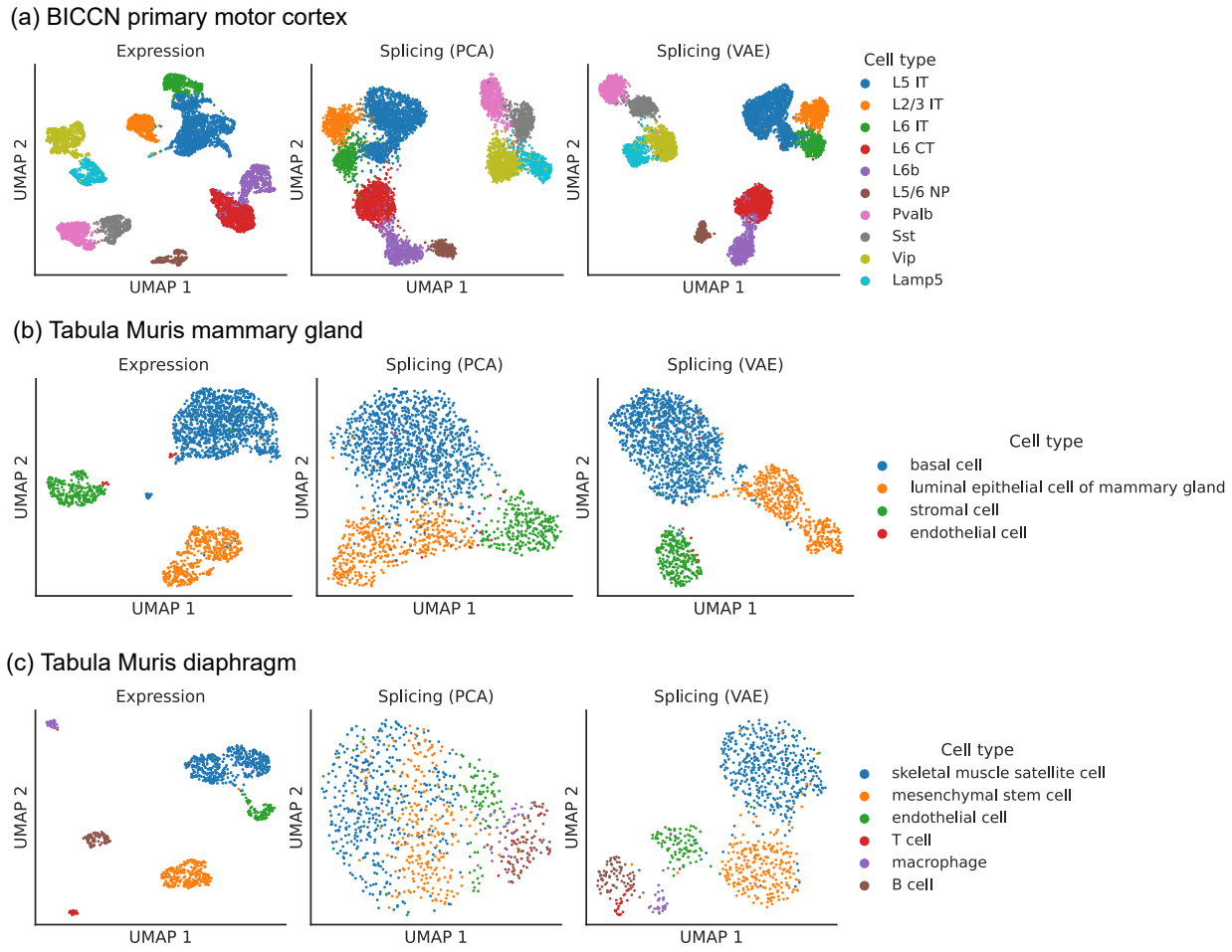

Figure 3: **Comparison of splicing latent spaces obtained with PCA and VAE.** Cells from **(a)** the cortex, **(b)** mammary gland and **(c)** diaphragm are projected into a latent space using PCA or VAE and visualized using UMAP. Cell type labels are obtained from the original data sources and are based on clustering in the expression latent space. The VAE is able to better distinguish cell types in the splicing latent space than PCA.

in absolute transcript expression rather than proportions. While the authors indirectly find some differences in transcript proportions by inspecting genes with no differential expression, this is not a systematic analysis of differential transcript usage. Meanwhile, only microglial cells in *Tabula Muris* [21] have been analyzed at the isoform level. (*Tabula Muris* also contains 10x Chromium data analyzed at the isoform level [33]).

As a community resource, we provide complementary ways to interactively explore splicing patterns present in these data sets (Figure 4), available at `https://github.com/songlab-cal/scquint-analysis/` with an accompanying tutorial video. The UCSC Genome Browser [34] 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 [35] (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 5a). Cells in the splicing latent space strongly cluster by cell type (annotated by Yao *et al.* [24] based on gene expression). A similar analysis was recently performed [36] 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). However, the authors only considered annotated skipped exons, a subset of the events we quantify, and used a different dimensionality reduction technique.

Figure 5b (top) highlights some differentially spliced genes between Glutamatergic and GABAergic neurons, including the glutamate metabotropic receptor *Grm5* as well as *Shisa9/Ckamp44*, which associates with AMPA ionotropic glutamate receptors [37]. The expression pattern of these genes, meanwhile, does not readily distinguish the neuron classes. In *Pgm2*, a gene of the glycolysis pathway thought to be regulated in the developing cortex by mTOR [38], we discover a novel exon preferentially included in Glutamatergic neurons (Figure 5c).

Our differential splicing test reveals thousands of cell-type-specific splicing events, highlighting marker introns that distinguish neuron subclasses, while the expression of their respective genes does not (Figure 5b, bottom). As another example of the many novel isoforms we discover, 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 [39] (Figure 5d). This novel TSS (`chr16:5763871-5763913`), which lies in a highly-conserved region, is (partially) used by only L6b neurons. We are also able to detect well-known cell-type-specific alternatively spliced genes such as *Nrxn1*, which encodes a key pre-synaptic molecule (Figure 5e) [40]. In this case, we observe an exon (known as splice site 2) exclusively skipped in Vip and Lamp5 neurons.

B cell development in the marrow. 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 are harder to distinguish from the other B cell subpopulations (Figure 6a), reflecting less refined splicing programs or limitations in transcript capture efficiency. The several differential splicing events we identified throughout development displayed splicing trajectories mostly independent from the trajectories of gene expression (Figure 6b). We highlight alternative TSSs (one of them novel) in two

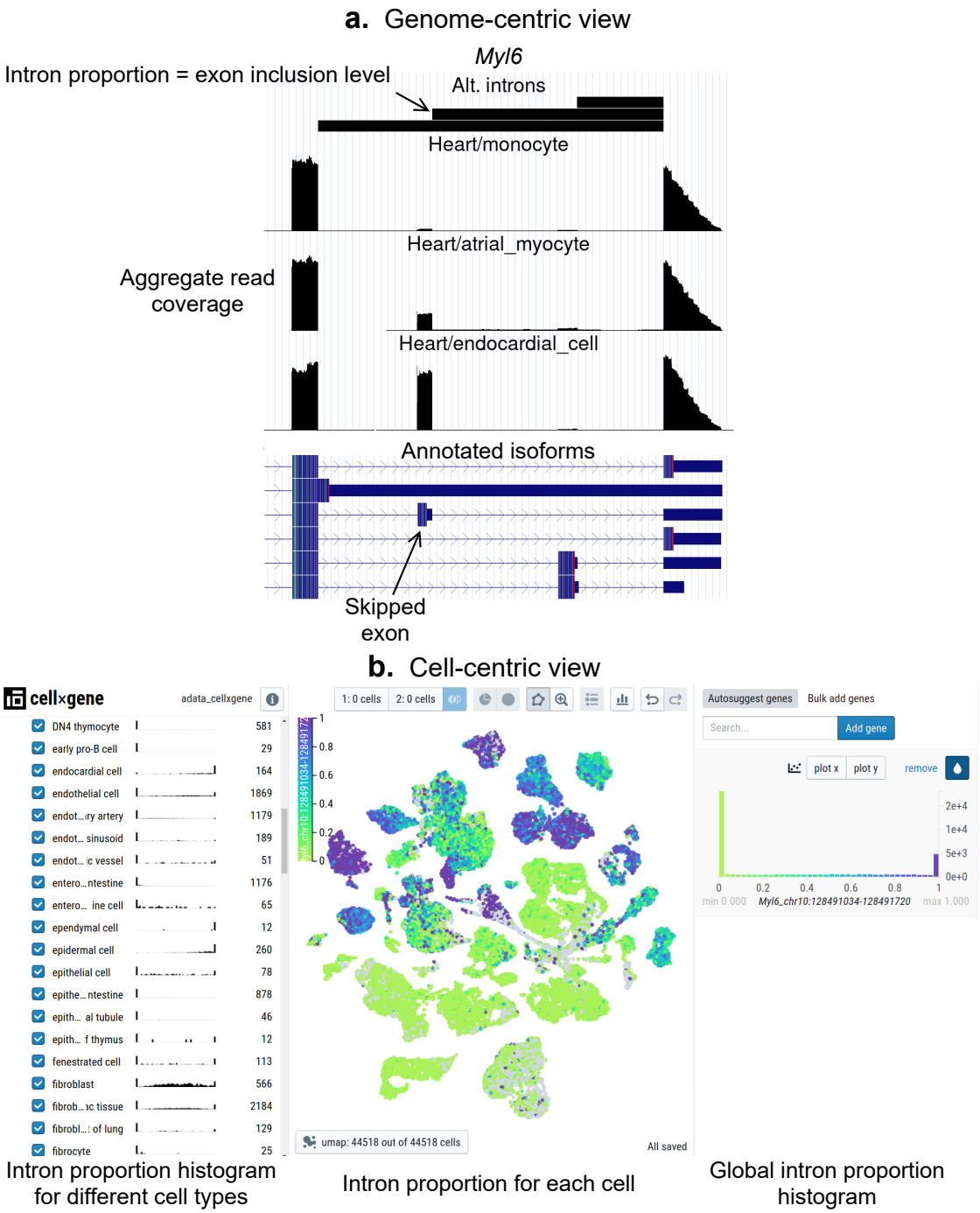

Figure 4: **Interactive visualizations of splicing patterns.** As an example, a skipped exon in *Myl6*. **(a)** The UCSC Genome browser visualization of this locus. Bottom: annotated isoforms of *Myl6*, including a skipped exon. Center: aggregate read coverage in three cell types with varying inclusion 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.

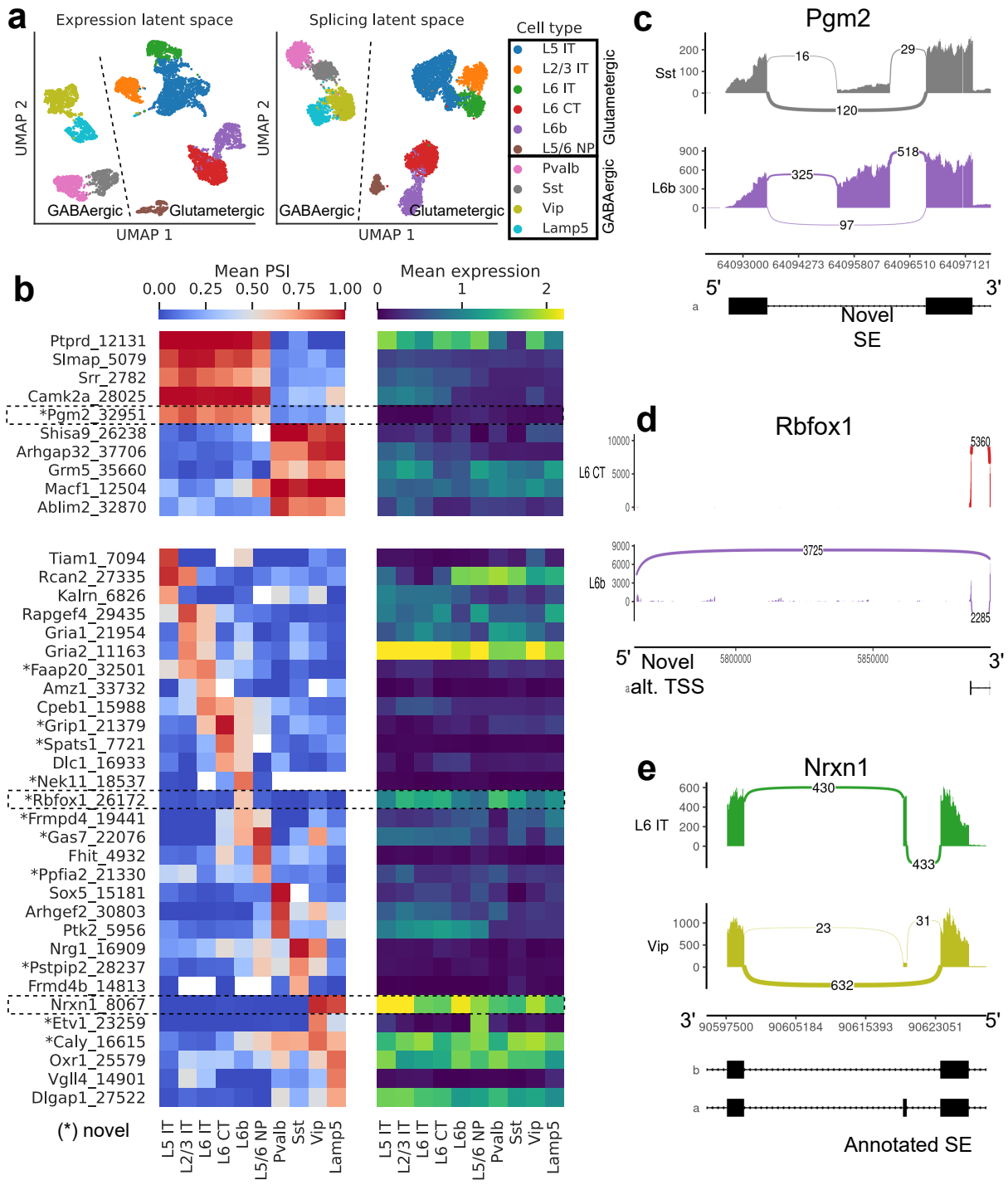

Figure 5: **Splicing patterns in *BICCN Cortex*. (a)** Expression and splicing latent spaces, visualized using UMAP. The expression (splicing) latent space is defined by running PCA (VAE) on the gene expression (alternative intron proportion, PSI) matrix. Cell types are well-separated in either latent space. **(b)** PSI of selected introns (left) and expression (log-transformed normalized counts) of their respective genes (right) averaged across cell types. Top: introns distinguishing Glutamatergic and GABAergic neuron classes. Bottom: introns distinguishing neuron subclasses. Intron coordinates are available in Table S2. **(c-e)** Sashimi plots [41] of specific alternative splicing events, displaying overall read coverage with arcs indicating usage of different introns (certain introns are shrunk for better visualization). **(c)** Novel skipped exon in *Pgm2*. **(d)** Novel alternative transcription start site (TSS) in *Rbfox1*. **(e)** Annotated skipped exon (SE) in *Nrxn1*.

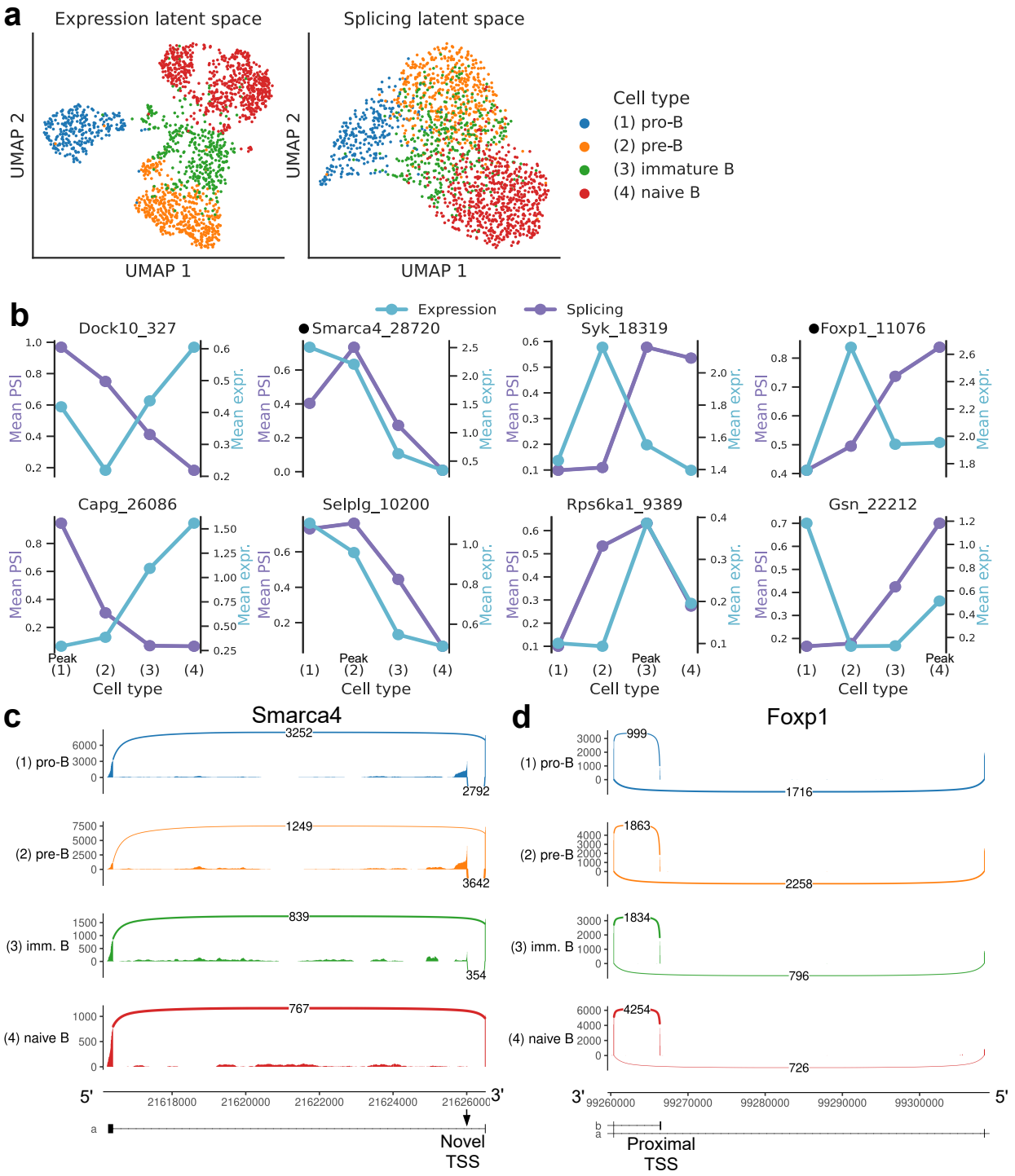

Figure 6: **Splicing in developing marrow B cells from *Tabula Muris*.** B cell developmental stages include pro-B, pre-B, immature B, and naive B. **(a)** Expression versus splicing latent space, as defined previously. In the splicing latent space, some cells types (pro-B) are better distinguished than others (immature B). **(b)** PSI of some introns that are differentially spliced throughout development, together with expression of the respective genes (log-transformed normalized counts). Expression and splicing can have very different trajectories. Intron coordinates are available in Table S3. **(c)** Sashimi plot of novel alternative transcription start site (TSS) in *Smarca4*. The novel TSS has maximum usage in pre-B cells, and then decays, while the expression peaks at pro-B cells. **(d)** Sashimi plot of an annotated alternative TSS in *Foxp1*. The proximal TSS in increasingly used as development progresses, while the expression peaks at pre-B cells.

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	880	8835	0.10	4
Cortex	6220	10	L5 IT	1571	1447	6402	0.23	2
Heart	4144	6	Endothelial cell of coronary artery	1126	465	7108	0.07	5
Large Intestine	3729	5	Enterocyte of epithelium	1112	586	10786	0.05	2
Marrow	4783	10	Hematopoietic stem cell	1363	692	9909	0.07	2

Table 2: **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.

transcription factors essential for B cell development; *Smarca4*, encoding BRG1 [42] (Figure 6c); and *Foxp1* [43] (Figure 6d). Though *Foxp1* expression peaks in pre-B cells and does not follow a monotonic trend over developmental stages, the alternative TSS is progressively included throughout B cell development. Combining gene-level expression with TSS usage, which can influence translation rate, provides a more nuanced characterization of the expression patterns of these important transcription factors. Some other differentially spliced genes with well-known roles in B cell development are *Syk* [44], *Dock10* [45], *Selplg/Psgl-1* [46], and *Rps6ka1* [47].

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 2 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.

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 7a). 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 7b). We find an increased proportion of differentially spliced non-coding RNA in the cortex, the majority of which are previously unannotated 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 7c). 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 lower degree of separation observed in the embedding of Figure 6a.

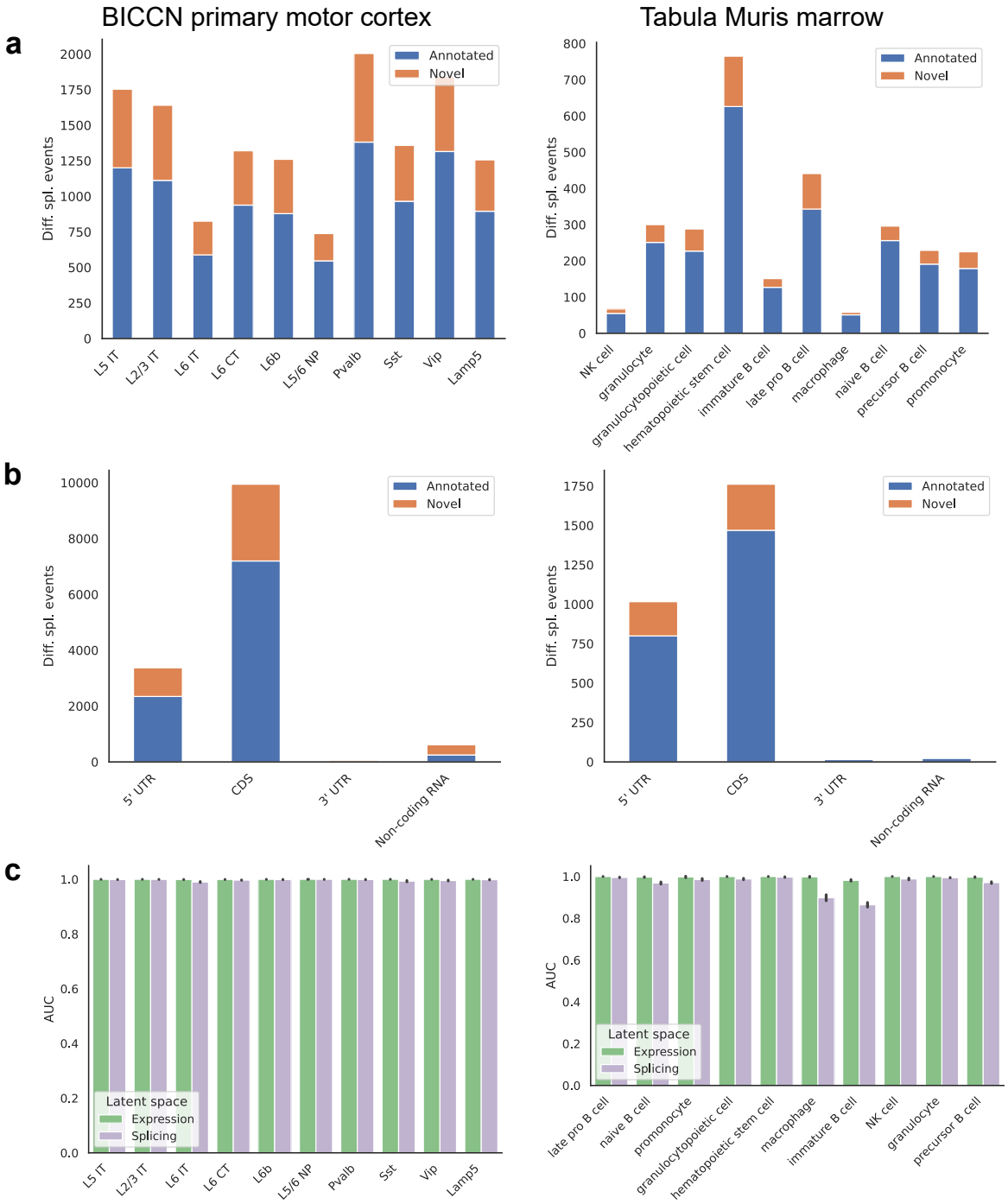

Figure 7: **Patterns across tissues.** **(a)** Number of differential splicing events detected in each cell type. Cortex cell types have more differential splicing events and larger proportions of novel events (those involving an intron absent from the reference). **(b)** Number of differential splicing 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.

Finding splicing factors associated with specific alternative splicing events

Several splicing factors have been identified as regulators of specific alternative splicing events, but most regulatory interactions remain unknown (see [48] for a review focused on the brain). To complement expensive and laborious knockout experiments, we sought to generate regulatory hypotheses by analyzing the correlation between splicing outcomes and splicing factor variation across cell types. Focusing on a subset of highly expressed genes in BICCN primary motor cortex neurons, we fit a sparse linear model regressing PSI of skipped exons on both splicing factor expression and splicing patterns (Figure 8a). Our model recovers several known regulatory interactions such as Khdrbs3's repression of splice site 4 (SS4) in neurexins, modulating their binding with postsynaptic partners [48]. Additionally, the proportion of a novel alternative TSS (though annotated in the human reference) in *Khdrbs3* (Figure 8b) is negatively associated with SS4 in *Nrxn1* and *Nrxn3*. The skipping of exon 5 (E5) of *Grin1*, which controls long-term synaptic potentiation and learning [49], is known to be regulated by Mbnl2 and Rbfox1 [48]. The model associates *Grin1* E5 PSI with the expression of *Rbfox1* but not *Mbnl2*; however, it does suggest an association with the PSI of two skipped exons in *Mbnl2* (Figure 8c) and further implicates the inclusion level of the novel alternative TSS in *Rbfox1* reported above (`Rbfox1_chr16:5763912-6173605`, Figure 5d). These results help clarify the disparate impacts of expression and alternative splicing in splicing factors, and encourage the use of regression models to suggest candidate regulators of cell-type-specific alternative splicing. Such computationally generated hypotheses are particularly valuable for splicing events in splicing factors because of the heightened difficulty to experimentally perturb specific exons rather than whole genes.

Discussion

Here we've introduced scQuint, a toolkit for the analysis of splicing in short-read scRNA-seq data which address three different tasks–quantification, dimension reduction/visualization, and differential splicing. Our localized approach to intron quantification via alternative intron usage was inspired by our observation of substantial coverage biases in Smart-seq2 data, adding to the list of inherent challenges when analyzing isoform variation in short-read single-cell data. Until these biases are better understood, including potential interactions between technical and biological factors, we advocate quantifying transcript variation in a local manner. Although this limits the scope of transcript variation one can investigate, we showed that this subset is robust to coverage biases and easily detects novel isoforms. To visualize quantified splicing data, we implemented a novel VAE which produces better separated and more refined cell type clusters than PCA. Finally, our testing procedure for differential splicing, adapted from existing bulk methods, demonstrated superior statistical calibration in single-cell data sets.

We successfully applied scQuint to data sets consisting of diverse mouse tissues, finding strong cell-type-specific splicing signals. This result, coupled with the lack of overlap between the most strongly differentially expressed and spliced genes (Table 2), should encourage tissue specialists to further study the contribution of isoform variation to cell identity and its role in cell and tissue homeostasis. To facilitate further exploration of these data, we've uploaded our results to cell and genome browsers (linked at `https://github.com/songlab-cal/scquint-analysis/`).

Despite the apparent association between splicing and cell identity, our analyses are yet to produce 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

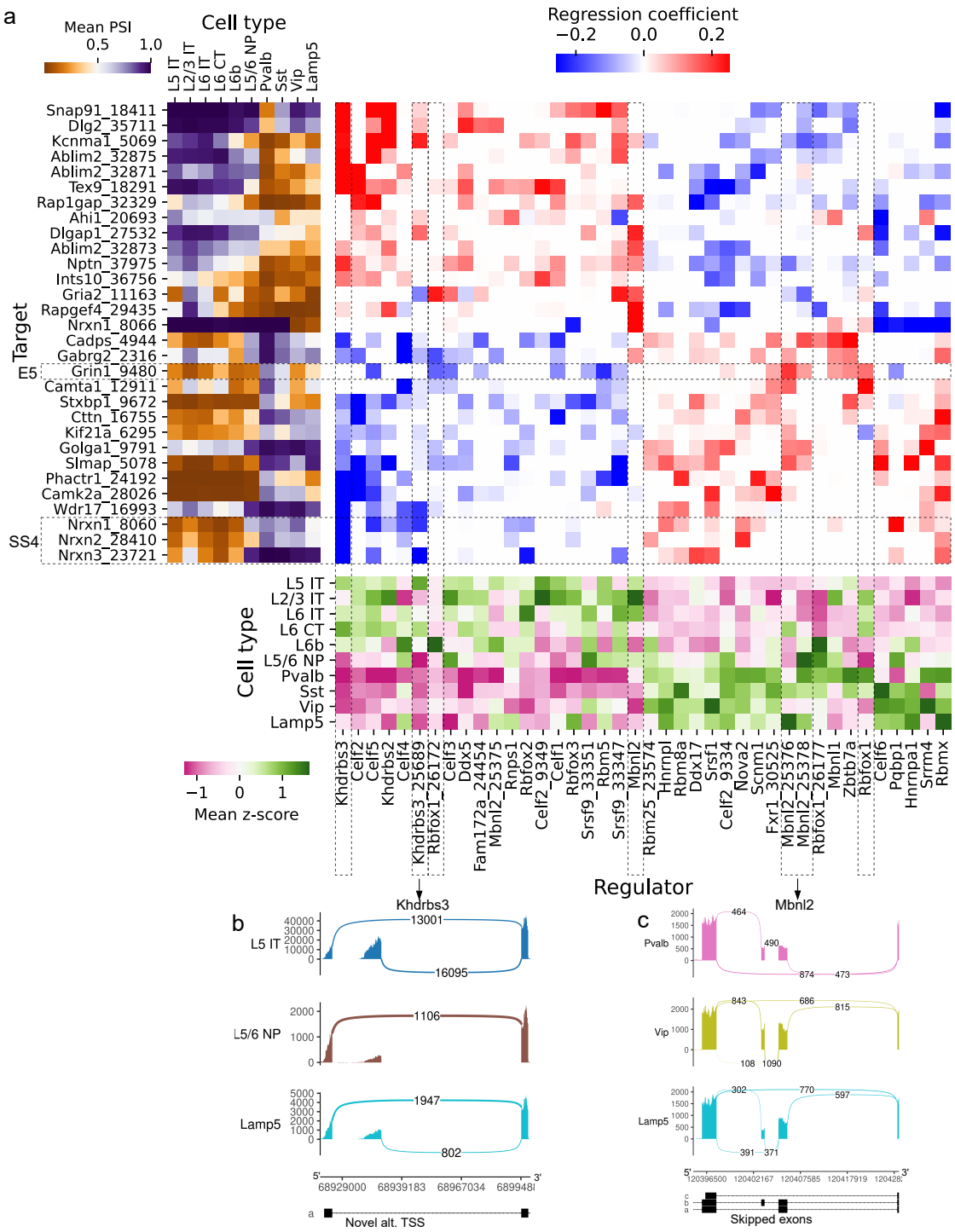

Figure 8: **Associations between splicing factors and alternative splicing. (a)** Regression analysis of exon skipping based on expression and splicing of splicing factors, using the BICCN mouse primary motor cortex dataset. Left panel: mean PSI of skipped exons across cell types. Bottom panel: mean z-scores of selected splicing factor features across cell types, including whole-gene expression (gene name) and PSI of alternative introns (gene name and numerical identifier). Intron coordinates are available in Table S4. Center panel: regression coefficients (log-odds) of each splicing factor feature used to predict skipped exon PSI in our sparse Dirichlet-Multinomial linear model. Full results are available in Figure S5. **(b)** Novel alternative TSS in *Khdrbs3*. **(c)** Annotated skipped exons in *Mbnl2*.

expression and splicing proportions provide two different projections of the same underlying cell state.

To augment our understanding of cell-type-specific splicing, we performed a regression analysis based on the natural variation of splicing factors in different cell types. We recovered several regulatory interactions which had been previously identified through knockout experiments and provide potential granularity on the nature of these interactions. A key component of our analysis is the decision to supplement the expression of splicing factors with their alternative splicing patterns as features in the model. Our model suggests several alternative splicing events in splicing factors themselves (including novel ones) can contribute to their regulatory activity, presenting interesting follow-up opportunities.

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 [50] estimates can be distorted by alternative splicing, and Bergen *et al.* [51] 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 [52] 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 [53]). *BICCN Cortex* data were downloaded from http://data.nemoarchive.org/biccn/lab/zeng/transcriptome/scell/SMARTer/raw/MOp/ and filtered as in [32].

Quantification. The bioinformatic pipeline was implemented using Snakemake [54]. Raw reads were trimmed from Smart-Seq2 adapters using Cutadapt [55] 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 [56] 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 [57] (downloaded from https://github.com/Boyle-Lab/Blacklist/raw/master/lists/mm10-blacklist.v2.bed.gz).

Gene expression was quantified using featureCounts [58], and total-count normalized such that each cell had 10,000 reads (as in the Scanpy [59] 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 run PCA, 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.

The VAE was implemented using PyTorch [60] and scvi-tools [61]. The following is the generative model, repeated for each cell:

1. Sample the latent cell state $z \sim \text{Normal}(0, \text{I})$
2. For each intron group g:
   (a) Obtain the underlying intron proportions:
   $\vec{p}_g = \text{softmax}(f_g(z))$
   (b) Sample the intron counts conditioning on the total observed n_g:
   $\vec{y}_g | n_g \sim \text{DirichletMultinomial}\left(n_g, \alpha_g \cdot \vec{p}_g\right)$

Here f_g, known as the decoder, can be any differentiable function, including linear mappings and neural networks. α_g is a scalar controlling the amount of dispersion. We optimize a variational posterior on cell latent variables $q(z|y)$ (Gaussian with diagonal covariance, given by an encoder neural network) as well as point estimates of global parameters f_g, α_g. The encoder takes as input the smoothed PSI values, as in PCA, but the likelihood is based on the raw intron counts. The objective to maximize is the evidence lower bound (ELBO), consisting of a reconstruction term and a regularization term:

$$\text{ELBO}(y) = \mathbf{E}_{z \sim q(z|y)}[\log p(y|z)] - \mathbf{KL}(q(z|y)||p(z))$$

Optimization is performed using Adam [62], a stochastic gradient descent method. To avoid overfitting in cases of relatively few cells with respect to the number of features, we considered a linear decoder [63], as well as a $\text{Normal}(0, \sigma)$ prior on the entries of the decoder matrix. Hyperparameters were tuned using reconstruction error on held-out data and are described in Table S5.

To verify that we prevent leakage of gene expression information into our splicing profiles, we apply our VAE to embed 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 [28] 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 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 [64].

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 [65].

Latent space 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 the VAE on the alternative intron count matrix (or equivalent features, e.g., Kallisto transcript counts, DEXSeq exon counts). Both latent spaces were visualized using UMAP [66]. In the comparison of Figure 1, we used our own implementation of the quantifications proposed by ODEGR-NMF, DEXSeq, and DESJ for ease of application to large single-cell datasets.

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.

Cell-type-specific differential splicing. 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.

Splicing factor regression analysis. We obtained 75 mouse splicing factors using the Gene Ontology term "alternative mRNA splicing, via spliceosome" (http://amigo.geneontology.org/

amigo/term/GO:0000380). A skipped exon annotation, processed by Brie [13], was downloaded from https://sourceforge.net/projects/brie-rna/files/annotation/mouse/gencode.vM12/SE.most.gff3/download. Instead of using single cells as replicates, we partitioned the BICCN primary motor cortex dataset into roughly 200 clusters of 30 cells each that were pooled to create pseudobulks, aiming to reduce variance in the expression and splicing of splicing factors used as covariates in the model. We filtered target exon skipping events to those defined in at least 95% of the replicates, and those having a PSI standard deviation of at least 0.2. We used log-transformed normalized expression and PSI of alternative splicing events as input features. We chose to keep the PSI of only one intron per intron group to avoid the presence of highly correlated features and improve clarity, even if some information from non-binary events is lost. Input features were filtered to those having standard deviation of at least 0.05, and then standardized. A lasso Dirichlet-Multinomial GLM was fit to the data (in this instance, the model reduces to a Beta-Binomial because skipped exons are binary events), with the sparsity penalty selected via cross-validation. As a first approach, we fit a regular lasso linear regression model on PSI instead of raw counts, resulting in roughly similar patterns in the coefficients. Figure 8c shows the coefficients of the lasso Dirichlet-Multinomial model for the top 30 targets with the highest variance explained by the regular lasso model, all above 68%.

Code and data availability. scQuint implementation in Python 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 [59], Seurat [67], and other tools.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We would like to thank Angela Oliveira Pisco, Spyros Darmanis, and Kif Liakath-Ali for helpful discussions. We also thank the Chan Zuckerberg Biohub for hosting our cell×gene sessions and Aaron McGeever for assistance. 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 preprint*. https://doi.org/10.1101/2020.01.10.902361.

[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] Buen Abad Najar CF, 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] Liu S, et al. (2021) Single-cell differential splicing analysis reveals high heterogeneity of liver tumor-infiltrating t cells. *Scientific Reports* 11(1):1–12.

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

[24] Yao Z, et al. (2020) An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. *bioRxiv preprint*. https://doi.org/10.1101/2020.02.29.970558.

[25] Anders S, Reyes A, Huber W (2012) Detecting differential usage of exons from RNA-seq data. *Genome Research* 22(10):2008–2017.

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

[27] 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.

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

[29] Kingma DP, Welling M (2014) Auto-encoding variational Bayes in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. arXiv preprint arXiv:1312.6114.

[30] Lopez R, Gayoso A, Yosef N (2020) Enhancing scientific discoveries in molecular biology with deep generative models. *Molecular Systems Biology* 16(9):e9198.

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

[32] Booeshaghi A, et al. (2020) Isoform specificity in the mouse primary motor cortex. *bioRxiv preprint*. https://doi.org/10.1101/2020.03.05.977991.

[33] Patrick R, et al. (2020) Sierra: Discovery of differential transcript usage from polya-captured single-cell RNA-seq data. *Genome Biology* 21(1):1–27.

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

[35] Megill C, et al. (2021) cell×gene: a performant, scalable exploration platform for high dimensional sparse matrices. *bioRxiv preprint*. https://doi.org/10.1101/2021.04.05.438318.

[36] 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):e2013056118.

[37] Von Engelhardt J, et al. (2010) Ckamp44: a brain-specific protein attenuating short-term synaptic plasticity in the dentate gyrus. *Science* 327(5972):1518–1522.

[38] Schüle M, et al. (2021) mTOR driven gene transcription is required for cholesterol production in neurons of the developing cerebral cortex. *International Journal of Molecular Sciences* 22(11):6034.

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

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

[41] 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.

[42] Bossen C, et al. (2015) The chromatin remodeler brg1 activates enhancer repertoires to establish b cell identity and modulate cell growth. *Nature Immunology* 16(7):775–784.

[43] Hu H, et al. (2006) Foxp1 is an essential transcriptional regulator of b cell development. *Nature Immunology* 7(8):819–826.

[44] Cornall RJ, Cheng AM, Pawson T, Goodnow CC (2000) Role of syk in b-cell development and antigen-receptor signaling. *Proceedings of the National Academy of Sciences* 97(4):1713–1718.

[45] García-Serna AM, et al. (2016) Dock10 regulates cd23 expression and sustains b-cell lymphopoiesis in secondary lymphoid tissue. *Immunobiology* 221(12):1343–1350.

[46] Urzainqui A, et al. (2020) Relevance of psgl-1 expression in b cell development and activation. *Frontiers in Immunology* 11:2900.

[47] Stein M, et al. (2017) A defined metabolic state in pre b cells governs b-cell development and is counterbalanced by Swiprosin-2/EFhd1. *Cell Death & Differentiation* 24(7):1239–1252.

[48] Vuong CK, Black DL, Zheng S (2016) The neurogenetics of alternative splicing. *Nature Reviews Neuroscience* 17(5):265–281.

[49] Sengar AS, et al. (2019) Control of long-term synaptic potentiation and learning by alternative splicing of the nmda receptor subunit glun1. *Cell Reports* 29(13):4285–4294.

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

[51] Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ (2020) Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology* 38:1408–1414.

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

[53] Tabula Muris Consortium, , et al. (2020) A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* 583(7817):590–595.

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

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

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

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

[58] 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.

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

[60] Paszke A, et al. (2019) PyTorch: An imperative style, high-performance deep learning library in *Advances in Neural Information Processing Systems*. pp. 8026–8037.

[61] Gayoso A, et al. (2021) scvi-tools: a library for deep probabilistic analysis of single-cell omics data. *bioRxiv preprint*. https://doi.org/10.1101/2021.04.28.441833.

[62] Kingma DP, Ba J (2015) Adam: A method for stochastic optimization in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1412.6980.

[63] Svensson V, Gayoso A, Yosef N, Pachter L (2020) Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* 36(11):3418–3421.

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

[65] 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.

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

[67] Stuart T, et al. (2019) Comprehensive integration of single-cell data. *Cell* 177:1888–1902.