SOFTWARE

Open Access

Mining alternative splicing patterns in scRNA-seq data using scASfind

Yuyao Song[1,2], Guillermo Parada[1,3], Jimmy Tsz Hang Lee[1] and Martin Hemberg[1,4*]

*Correspondence:
mhemberg@bwh.harvard.edu

[1] Wellcome Sanger Institute, Hinxton CB10 1SA, UK
[2] European Molecular Biology Laboratory-European Bioinformatics Institute, Hinxton CB10 1SD, UK
[3] Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada
[4] The Gene Lay Institute of Immunology and Inflammation, Brigham and Women's Hospital, Massachusetts General Hospital, and Harvard Medical School, Boston, MA 02115, USA

Abstract

Single-cell RNA-seq (scRNA-seq) is widely used for transcriptome profiling, but most analyses focus on gene-level events, with less attention devoted to alternative splicing. Here, we present scASfind, a novel computational method to allow for quantitative analysis of cell type-specific splicing events using full-length scRNA-seq data. ScASfind utilizes an efficient data structure to store the percent spliced-in value for each splicing event. This makes it possible to exhaustively search for patterns among all differential splicing events, allowing us to identify marker events, mutually exclusive events, and events involving large blocks of exons that are specific to one or more cell types.

Keywords: Alternative splicing, Single cell RNA-seq, Cell type-specific events

Introduction

Alternative splicing (AS) is an essential, ubiquitous regulatory mechanism in eukaryotes. Through AS, a single gene can yield multiple mRNA isoforms, greatly expanding the protein diversity encoded by eukaryotic genes [1]. Fine-tuned regulation of alternative splicing has a critical role in the development and function of a diversity of tissues and cell types, including muscles, neurons, and immune cells [2–6]. Splicing errors can also lead to an array of human diseases, such as neurodegenerative diseases, autoimmunity, and cancer [7–9].

Decades of research using bulk methods have shown that many AS events are tissue-regulated [10], yet cell type-specific splicing remains incompletely understood. Using single cell RNA-seq (scRNA-seq), cell types can be comprehensively identified based on their expression profile, paving the way for studying splicing patterns. Although most scRNA-seq studies use droplet-based technologies such as 10X Chromium, which only profiles one end of the transcript, there are full-length scRNA-seq technologies, such as Smart-seq2 [11] and VASA-seq [12], that provide coverage of the entire transcript. Full-length technologies make it possible to conduct a local, event-level splicing quantification per cell type. In an event-level AS quantification, transcripts can be split into non-overlapping exonic regions, referred to as splicing nodes [13–16]. Nodes are further classified based on their behavior during splicing, e.g., core exons (CEs) or alternative donors (ADs). Then, the percent spliced-in (PSI) value for splicing nodes can be calculated based on reads spanning node junctions [13, 15, 17]. PSI is an informative indicator of exon usage frequency, providing an intuitive and easily interpretable metric to describe complex splicing events.

There are plenty of computational methods for event-level splicing quantification in bulk RNA-seq, such as MISO [17], dSpliceType [18], rMATS [16], MAJIQ [15], and SUPPA2 [19], but they are poorly suited due to the high sparsity and large size of scRNA-seq datasets. To overcome these issues, several methods aiming to detect and quantify AS in single-cell data have been developed. They include SingleSplice [20] which compares biological variation and technical noise in a population of single cells to find genes with isoform usage differences. Expedition [21] is a suite of tools that can detect differences among the usage of splicing modalities. Huang and Sanguinetti have developed BRIE and BRIE2 [22, 23], which use Bayesian models for PSI estimation to overcome sparsity. SICILIAN [24] assigns probabilities to called splice junctions to improve precision for their detection, and SpliZ [25] generalizes PSI to enhance splicing quantification at the single-cell level. A recent software tool is MARVEL [26], which integrates splicing and gene expression analyses. However, MARVEL analysis is limited to splicing events involving a single exon and it can only detect differential splicing between pairs of cell types. None of the methods presented to date can leverage event-level splicing quantification to comprehensively characterize cell type-specific splicing patterns, involving either single or multiple exons, without using a parametric model or imputing missing values.

To facilitate comprehensive de novo detection of cell type-specific AS events, we developed scASfind [27], a flexible and intuitive method for mining complex AS patterns in large single-cell datasets. scASfind is an open-source R package which is freely available at https://github.com/hemberg-lab/scASfind. scASfind uses a similar data compression strategy as our previous work scfind [28] to transform the cell pool-to-node differential PSI matrix into an index. This efficient data structure enables rapid access to cell type-specific splicing events, making it possible to use an exhaustive approach when carrying out pattern searches across the entire dataset. Importantly, scASfind does not involve any imputation or model fitting, instead cells are pooled to avoid the challenges presented by sparse coverage. Moreover, there is no restriction on the number of exons, or the inclusion/exclusion events involved in the pattern of interest. Building on these fast searches, scASfind allows interactive searching of cell type specificity of splicing patterns, such as differential splicing, mutually exclusive exons, and coordinated splicing events. We applied scASfind to mouse primary visual cortex [29], mouse embryonic development [12], and human fetal liver [30] to characterize cell type-specific splicing patterns.

Results

Data compression enables fast searching of splicing patterns

scASfind takes full-length scRNA-seq data, such as Smart-seq2 [31], RamDA-seq [32], Smart-seq3 [33], VASA-seq [12], and FLASH-seq [34], as input for splicing quantification. Tag-based methods such as 10X Genomics Chromium are unsuitable since they only capture the transcript's 3′ or 5′ end, and typically do not provide enough reads that span splice junctions. It is assumed that the data has been clustered and annotated so that each cell is assigned a cell type. Several cells of the same type are first combined into cell pools to provide sufficient reads for robust and accurate PSI quantification with Whippet [13] using the MicroExonator workflow [35].

The size of cell pools is an important hyperparameter in the analysis, which should be carefully determined on a per-case basis. Pooling single cells aims to reduce the technical sparsity of scRNA-seq data, which could be due to technology or tissue. In our examples, the dataset by VASA-seq (mouse embryo) was two orders of magnitude sparser than the two Smart-seq2 datasets (mouse cortex, human fetal liver); therefore, we employed a much larger pool size (Table 1). Choosing a pool size needs to balance two aspects: ensuring a proper coverage for PSI quantification, while preserving enough cell pools per cell type to faithfully represent the variation of cell abundance and allow effective statistics tests in scASfind. An examination of the impact of pool size on scASfind results in the mouse cortex data can be found in the "Methods" section.

After obtaining a splicing node x cell pool matrix of PSI values (Fig. 1a), scASfind first centers each column of the matrix to obtain the deviation of PSI values from the dataset mean (default: $|\Delta\text{PSI}| > 0.2$). This is to capture the biologically informative PSI variation across cells. The differential PSI matrix is further split into two to encode positive (spliced-in) and negative (spliced-out) PSI values. In both matrices, a non-zero value indicates that there is differential inclusion or exclusion of the splicing node in that particular cell pool. By ensuring that the matrices are sparse, we can achieve a high compression rate and fast pattern matching, even for large datasets.

We adopted the indexing strategy in scfind [28] to compress the two sparse PSI matrices into two scASfind indexes, and we then combined them into a single meta-index object. The index efficiently stores the splicing nodes that have a PSI value deviated from the mean (see "Methods" for details about data compression into an index).

Table 1 Datasets used in this study, parameters, on-disk size and the time of building scASfind index. Splicing nodes which had a PSI deviating from the dataset mean in at least one cell pool were encoded. The size of raw data includes the raw PSI matrix and relevant metadata objects, while the size of the scASfind index object involves the compressed PSI index and the same metadata objects (see "Methods"). Both sizes are on-disk file sizes. The time to build the scASfind index is calculated by running the index-building script in scASfind on a HPC cluster. *UMI*: unique molecular identifier; *MB*: megabyte; *std*: standard deviation

Dataset	Number of single cells	Number of cell types	UMIs per single cell (mean ± std)	Cell pool size used	Number of cell pools	Number of encoded splicing nodes	Size of raw data (MB)	Size of scASfind index object (MB)	Time to build the scASfind index (seconds)
Mouse cortex [29]	1654	49	1,594,655 ± 491,715	5	339	90,834	14,142.8	55.5	1185
Mouse embryo [12]	33,662	37	14,936 ± 12,319	200	191	75,837	27,484.3	21.9	1366
Human fetal liver [30]	4503	23	451,901 ± 393,948	10	451	199,361	29,374.5	101.4	1864

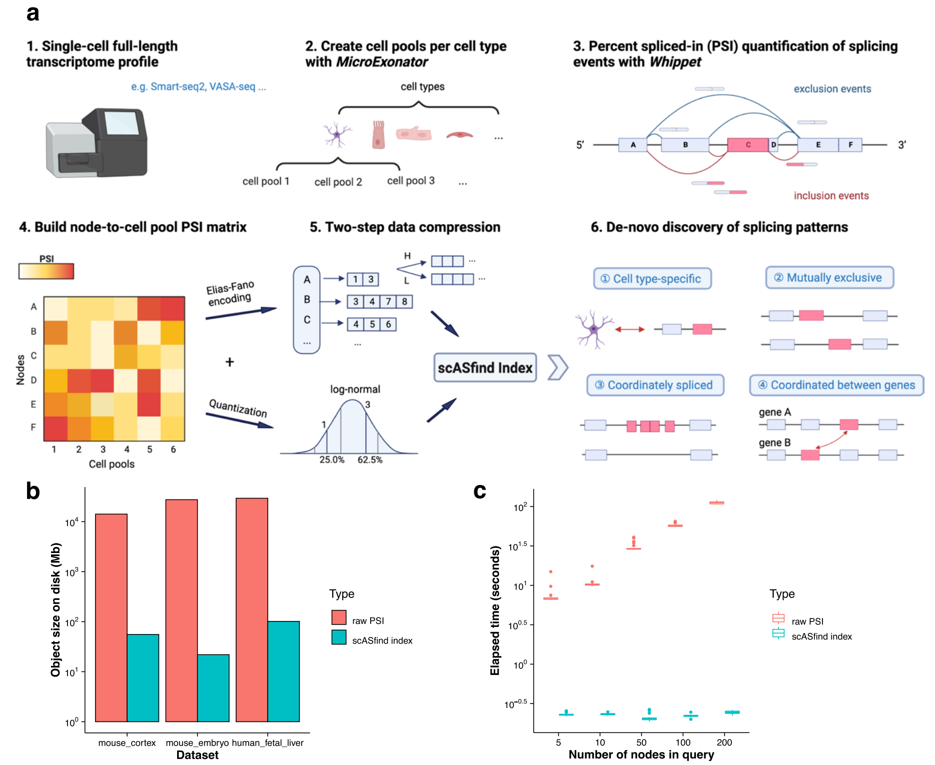

Fig. 1 Overview of scASfind. **a** Schematic of the scASfind workflow. Single-cell full-length transcriptome sequencing data such as Smart-seq2 or VASA-seq are suitable inputs for the scASfind workflow. Cells from the same cell type are pooled to increase the accuracy of splicing event detection (default 5 cells per pool) with MicroExonator [35]. The PSI value for each splicing node is calculated by Whippet [13] to obtain a splice node-by-cell pool PSI matrix, and we then build a scASfind index containing information about splicing events that are differentially spliced in or spliced out in each cell pool. Finally, we query the index to search for cell type-specific differential splicing events, mutually exclusive node pairs and consecutive nodes that are similarly spliced-in or coordinated splicing events. **b** The size of the file saved to disk containing either the raw PSI values and metadata objects or the scASfind index with metadata objects built with a two-bit quantization. **c** The elapsed time of searching all cells with increased inclusion, i.e., has a PSI no less than 0.2 higher than the dataset mean, in any of five randomly selected splicing nodes. The process is repeated 30 times. The bar in the boxplot shows the arithmetic mean, lower and upper hinges correspond to the first and third quartiles, whiskers extend from the hinge to the largest value no further than 1.5 * interquartile range from the hinge, and outliers beyond this range are plotted as individual data points. PSI, percent spliced-in

In addition to providing efficient storage, the scASfind index also allows for rapid access to the raw PSI values associated with each splicing node. In particular, it allows us to use AND or OR queries to find the set of cell pools that match a set of inclusion/exclusion criteria, e.g., nodes 1, 2, and 3 need to be included well above mean while nodes 4 and 5 are excluded well below mean. By combining multiple queries, we can carry out more complex searches, and since each operation is fast it becomes possible to adopt an exhaustive approach to search the entire dataset.

In the following, we set out to demonstrate how the scASfind index allows for thorough identification and characterization of cell type-specific splicing events. We used scASfind to analyze three technically and biologically distinct datasets: a mouse cortex data profiled using Smart-seq2 (hereafter referred to as mouse cortex) [29], a mouse embryonic development dataset profiled using VASA-seq (hereafter referred to as mouse embryo) [12], and a human fetal liver dataset profiled using Smart-seq2 (hereafter referred to as human fetal liver) [30].

For all three datasets, the scASfind representation required two to three orders of magnitude less disk space (Fig. 1b) when using two bits for the quantizer. We also benchmarked the search times: compared to an implementation using only standard data structures, scASfind was hundreds or thousands of times faster (Fig. 1c). For example, for the mouse embryo dataset, finding all pools that have increased inclusion of any of 200 randomly selected nodes with scASfind took 0.24 s on average, compared to 112 s for the naive approach. scASfind was also highly robust to increased search size. One additional overhead for scASfind is the time to build the index, but this is relatively minor as none of the datasets took more than 30 mins (Table 1).

Splicing events are more precise markers of cell types

Cell types are typically associated with a set of marker genes, i.e., genes that are highly expressed compared to other cell types. Since the most widely used single cell protocols do not provide enough information to distinguish isoforms, transcripts are usually evaluated at the gene level. However, identifying a reliable set of marker genes can be challenging, especially for neuronal tissues with complex cell type taxonomy [36]. Since AS is known to be more prevalent in the brain [2, 37], we hypothesized that splicing events are more reliable for distinguishing cell types than gene expression. We refer to splicing nodes that are highly included or excluded in only one cell type as marker nodes in analogy with marker genes.

To identify cell type marker nodes, we used each cell type to query the scASfind index for nodes that have high or low inclusion. Benefiting from the speed at which these quantities can be extracted using the scASfind index, we carried out an exhaustive search to identify the best marker nodes for each cell type. Nodes were evaluated using the precision, recall, and F1 scores for their ability to detect the cell type of interest (see “Methods” for details). We used a similar procedure for marker genes using scfind, and we compared the quality of the markers by the precision, recall, and F1 scores. In the mouse cortex and the mouse embryo datasets, we observed higher F1 scores in splicing markers, compared with expression markers across the board (Fig. 2a, d). Interestingly, the higher F1 of splicing markers was largely driven by higher precision (Fig. 2b, e), suggesting that they yielded few false positives. The F1 and precision of splicing and expression markers showed comparative scores in the human fetal liver dataset (Fig. 2g, h). The lack of benefit in using splicing markers for the human fetal liver suggested that the splicing landscape in this dataset was less complex compared to the mouse cortex and embryo, possibly due to the tissue. We conjecture that the poor recall (Fig. 2c, f, i) for splicing markers could be due to the sparsity of splicing quantification leading to a high number of false negatives as there is insufficient information to accurately quantify splicing nodes in many pools.

Moreover, the inclusion or exclusion of splicing nodes was independent of increased or decreased expression, suggesting that splicing markers were largely independent of the expression level (Fig. 3). For instance, in astrocytes, *Dtna_27* was excluded while the cell type had higher expression of the *Dtna* gene. On the other hand, astrocytes had similar expression of the *Hnrnpa2b1* gene with other cell types while it had higher inclusion

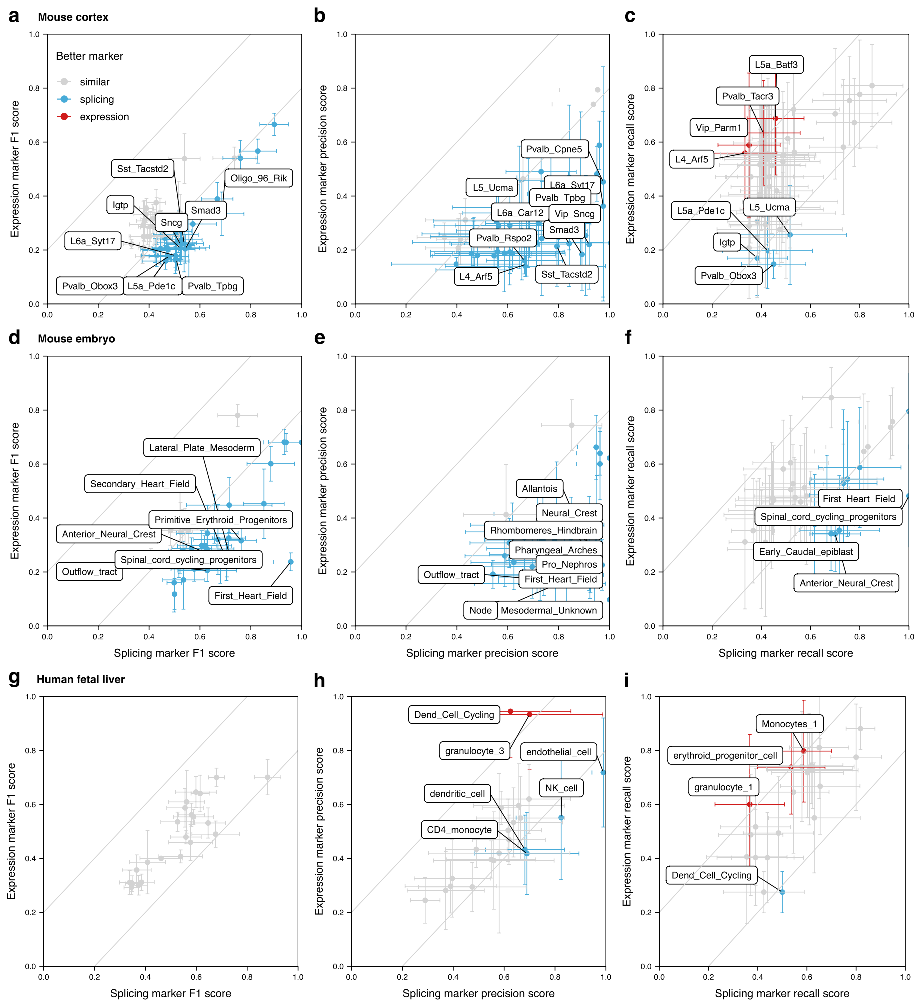

Fig. 2 Comparing gene expression and splicing as cell type markers. The top 20 expression and splicing markers, ranked by F1 scores, and their precision, recall, and F1 scores are calculated via scfind [28] or scASfind for **a**–**c** mouse cortex, **d**–**f** mouse embryo, and **g**–**i** human fetal liver. We compare the mean scores per cell type and consider a 0.2 difference between expression and splicing to indicate a better marker (gray lines in the figure). The dots represent the mean, while the whiskers indicate the minimum and maximum for the 20 markers. Cell types with either better splicing or expression markers are colored (blue for splicing, red for expression). For visual clarity, cell types have score differences of 0.5, 0.6, or 0.2 for precision; 0.2, 0.3, or 0.2 for recall; or 0.3, 0.4, or 0.2 for F1 in mouse cortex, mouse embryo, and human fetal liver data are labeled, respectively. These values are chosen based on the respective number of cell types with a better marker for each dataset

of *Hnrnpa2b1_32* (Fig. 3a, b). Another example is glutamatergic neuron subtype L4_Scnn1a (Fig. 3c, d), here we found both inclusion and exclusion splicing markers, while the expression levels of the corresponding genes could hardly distinguish this cell type from others. This was in line with observations by Wen et al. [26] that only a fraction of differentially spliced genes have expression changes in the same direction, indicating that differential splicing provides another layer of transcriptomics regulation that contributes to cell type heterogeneity.

While 67%, 57%, and 62% of the top 20 splicing markers were from different genes in mouse cortex, mouse embryo, and human fetal liver datasets, respectively, a single gene could contribute a large portion of marker nodes in some cases. We observed

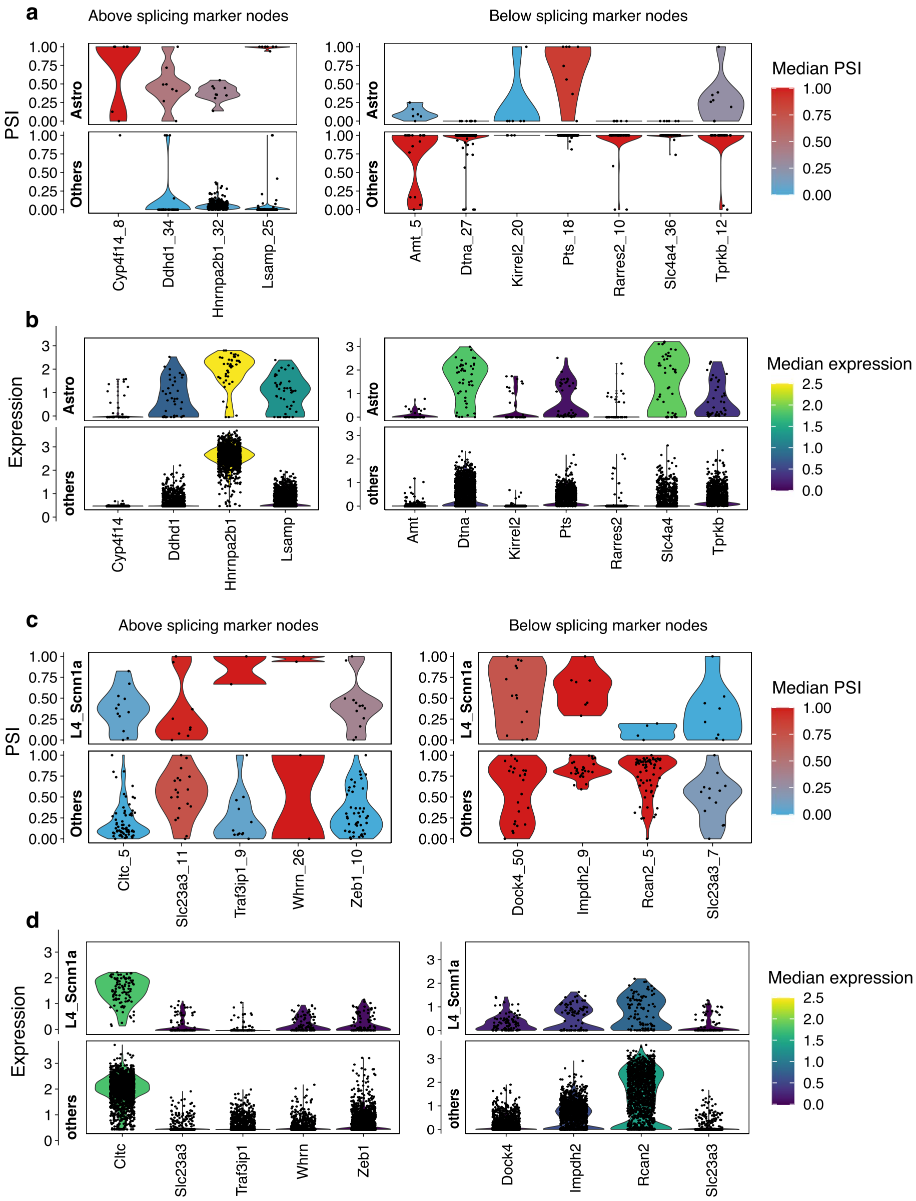

Fig. 3 PSI values of marker splicing nodes and expression levels of corresponding genes. **a** PSI values for astrocyte splicing markers from mouse cortex data, compared to the mean of all other cell types. **b** Expression levels for the same genes contribute to splicing markers in astrocytes. **c** PSI values for L4_Scnn1a neuron splicing markers from mouse cortex data. **d** Expression levels for the same genes in **c**. In all panels, each dot represents a cell pool, and the color scale shows the median PSI or gene expression level among the cell pools. PSI, percent spliced-in

that the *Ttn* gene, which encodes titin—the largest protein in the genome [38], contributed most of the above splicing markers in the first heart field (Fig. 2f) for the mouse embryo dataset. These results were consistent with previous analyses of Ttn splicing profiles that showed 50–219 exons to be developmentally regulated [5]. Taken together, this result suggested that splicing events frequently show higher cell type specificity than gene expression, and the splicing marker events reported by scASfind can be superior in terms of distinguishing cell types.

We compared our analysis of cell type-specific splicing markers with differentially spliced nodes calculated by MARVEL [26], a recent method focusing a combined analysis of expression and splicing in scRNA-seq data. Since MARVEL was designed for pairwise comparison, we also ran scASfind pairwise to detect differentially spliced CEs between all pairs of cell types in the mouse cortex data (see details in "Methods"). We saw a remarkable degree of consistency between the two tools. Among the 1176 pairs of cell types, MARVEL returned a total of 441 SE markers in 318 pairs of cell types. Among the MARVEL SE markers, 232 were ranked top 1 in scASfind and 330 were among the top 5 (Additional file 1: Fig. S1). For those marker nodes less significant in MARVEL, they also had a lower ranking and smaller F1 score in scASfind (Additional file 1: Fig. S1). The comparison suggested that MARVEL focused on returning the strongest signal while scASfind provided more information, reporting events in a custom range of F1 scores. Both tools were highly consistent in capturing the highest specificity splicing nodes that maximally distinguish the pair of cell types. Nevertheless, scASfind could provide results for all cell type pairs while many of these pairs did not have significant differentially spliced SEs in MARVEL analysis.

We also evaluated the PSI values of scASfind node markers in the mouse cortex data in VastDB [10], an atlas of alternative splicing profiles based on bulk RNA-seq data. For some non-neuronal cell types, there were purified single cell type tissues in VastDB, making them directly comparable with scASfind marker nodes. We found that for microglia, oligodendrocytes, and OPCs, the top 10 scASfind marker nodes clearly exhibited increased PSI in the respective VastDB cell type sample (Additional file 1: Fig. S2). Meanwhile, we observed that cortical, cerebellar, and whole brain data, which encompass mixed neuronal and glial cell types, did not exhibit high PSI for these marker genes. Furthermore, a probabilistic principal component analysis using the VastDB PSI values for all cell type marker nodes in scASfind across all VastDB tissues showed that these nodes have neural specificity (Additional file 1: Fig. S3). In summary, scASfind results were well supported by the PSI quantification from purified cell types in VastDB. The analysis further highlights the enhanced resolution provided by single-cell results compared to bulk RNA-seq from mixed cell types.

Detecting mutually exclusive exon pairs

Mutually exclusive splicing event is a special type of AS event where only one of two consecutive exons is included in the final mRNA product [39, 40]. In the largest study of mutually exclusive exons (MXEs) in bulk RNA-seq data carried out to date, Hatje et al. identified 855 exon pairs from 515 datasets [37, 39]. MXE splicing is known to be regulated by different molecular mechanisms that enable tissue-specific patterns [40–44]. We hypothesized that some of the observed tissue specific splicing profiles arise from the cumulative effects of cell type specific MXE preferences within each tissue. Hence, we leveraged scASfind to systematically discover MXEs and explore their cell type specificity.

We performed an exhaustive search of all three datasets to observe cell type specificity for all exon pairs that could be MXEs (Fig. 4a). That is, for all consecutive exons, we identified cell types in which one of them is always included and the other excluded. To ensure high-quality results, several additional filters were employed. First, we required

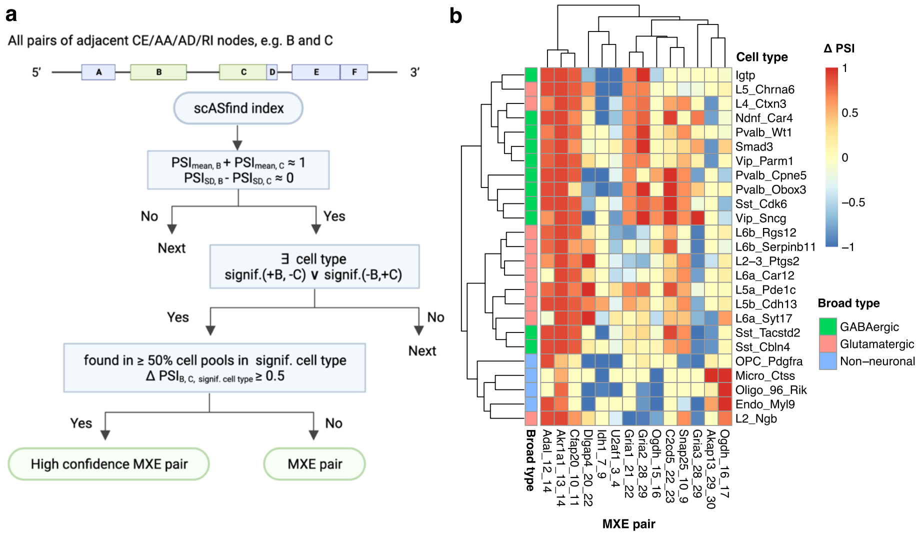

Fig. 4 Summary of 14 high-confidence MXE pairs in mouse cortex data. **a** Schematic overview of how MXE pairs are identified. **b** High confidence adjacent MXE pairs detected by scASfind. The heatmap color scale indicates the difference in raw PSI value (ΔPSI) between the downstream exon and the upstream exon in the pairs of MXEs. Cell types are grouped by their broad type; genes are organized by the number of cell types in which the pair is significant (decrease from left to right). MXE, mutually exclusive exon; PSI, percent spliced-in; CE, core exon; AA, alternative acceptor; AD, alternative donor; RI, retained intron; signif., significant

the pair to have mean PSI values summing to 1 ± 0.1, and that PSI standard deviation scores differ by less than 0.1 across all cell pools in the dataset. Second, we required at least one cell type to be significantly enriched for the pattern when one exon is included, and the other is excluded. Statistical significance was determined using hypergeometric tests. Third, we considered MXEs detected in over half of the cell pools and having a difference of cell type mean PSI value between the two exons ≥ 0.5 as highly confident pairs. The default criteria we have used were rather stringent to obtain a small amount of highly confident results.

We detected 63, 17, and 35 significant pairs of MXEs in mouse cortex, mouse embryo, and human fetal liver data, respectively. Among these 14, 2, and 2 were adjacent and highly confident. The high-confidence pairs in the mouse cortex dataset are summarized in Fig. 4b. Hierarchical clustering across cell types using the ΔPSI of these exons indicates that generally, cells of the same broad type have similar splicing patterns in these MXEs, with a few exceptions. This is concordant with the tissue-specific splicing observed in bulk RNA-seq. However, there are examples of cell types within each broad type that have distinctive MXE preferences, suggesting a more complex pattern.

A known example of MXEs can be found in the ionotropic glutamate receptor genes, AMPA 1/2/3 (*Gria1*, *Gria2*, *Gria3*), which have been studied extensively in mouse brain [45–47]. Reassuringly, the top candidates reported by scASfind include the three Gria genes. To the best of our knowledge, this is the first detailed study of cell type-specific MXE preferences for these genes (Fig. 5a). During development, some neurons switch from node 28 to using node 29, and this has important consequences for their responses to electric stimuli [48, 49]. We detected a significant preference for node 29 in glutamatergic neurons including L2 Ngb and L2/3 Ptgst, as well as L6a Car12, L6b Rgs 12 and L6b_Serpinb11, GABA-ergic neurons including Vip_Sncg, Pvalb_Obox3, Smad3,

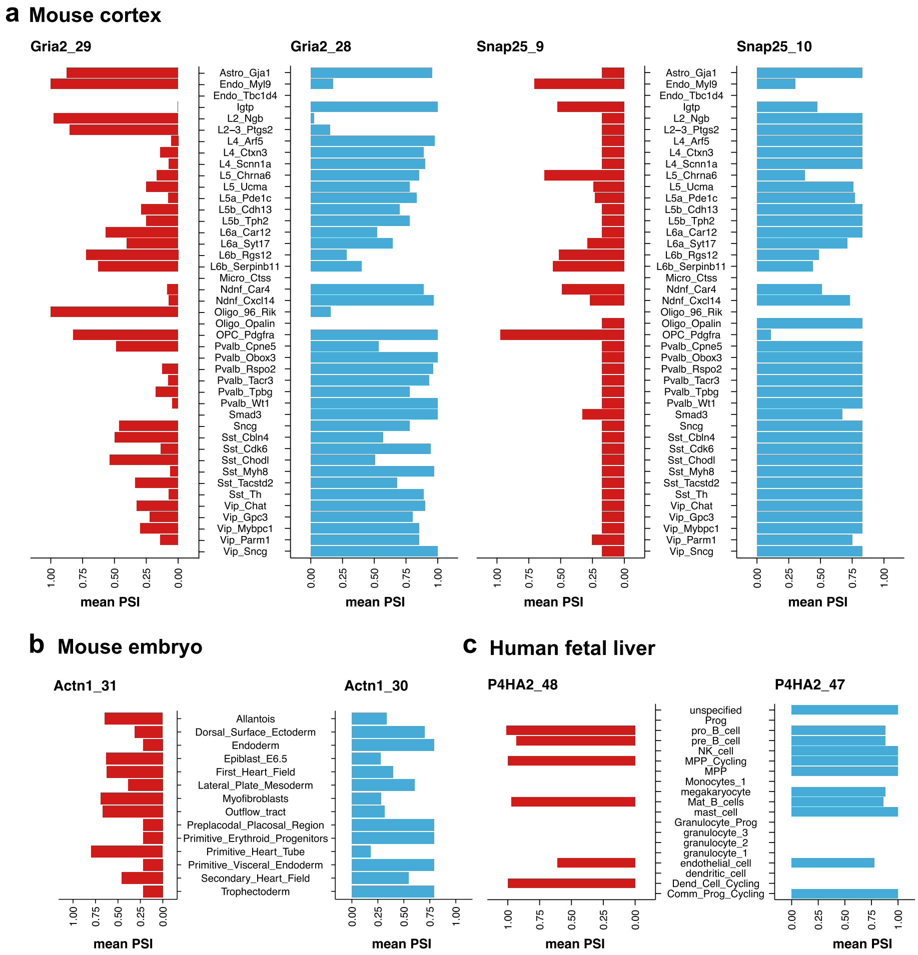

Fig. 5 PSI values of detected high confidence adjacent mutually exclusive exons. The mean PSI across cell pools in each cell type in **a***Gria2* (left) and *Snap25* (right) in the mouse cortex dataset, **b***Actn1* in the mouse embryo dataset, and **c***P4HA2* in the human fetal liver dataset. PSI, percent spliced-in

Igtp and Pvalb_Wt1, as well as in oligodendrocytes Oligo_96_Rik. By contrast, several glutamatergic and GABA-ergic neurons included node 28, suggesting that the cell type-specific pattern is complex. Another example was in the SNARE protein *Snap25*, whose MXE preference switches during mouse brain development [50, 51], and is related to regulating synaptic transmission and long-term synaptic plasticity [52, 53]. In our analysis, glutamatergic neurons L5_Chrna6, L6b_Rgs12, L6b_Serpinb11; GABA-ergic neuron Igtp, Ndnf_Car4, and oligodendrocyte progenitor cell OPC_Pdgfra showed a strong preference for node 9, while other cell types utilized node 10 (Fig. 5a). Taken together, our results recapitulate some of the complex cell type specific pattern of isoform switching for both the Gria genes and *Snap25*. In the mouse embryo data, we detected highly confident MXEs in *Actn1* and *Actn4*. *Actn1* has been found to have tissue-specific mutually exclusive splicing in adult mice. Compared to other tissues, muscle cells select an alternative exon which makes the protein's EF-hand motif insensitive to Ca2+, while brain cells include both exons [54, 55]. We were the first to describe the cell type preference of this MXE pair in the mouse embryo (Fig. 5b). For *Actn4*, we found primitive_heart_tube cells prefer *Actn4_14* while first_heart_field and secondary_heart_field cells chose *Actn4_13.* Finally, for the *P4HA2* gene in the human fetal liver dataset, Common_Prog_cycling and Dend_cell_cycling selected node 47 or 48, respectively (Fig. 5c).

Observing the cell type mean PSI of high-confidence MXEs, we found that the mutually exclusive pattern is not strictly followed in some cell types (Fig. 5). For example, Sst_Cbln4 cells showed comparable PSI for nodes 28 and 29 in *Gria2*, whereas astrocytes included both exons. This is in parallel with previous observations of MXEs only being mutually exclusive in specific tissues but not in all tissues [1, 40, 56]. For example, *TCL6* only shows exclusive patterns in specific tissues while on the basis of all known transcripts, the pattern is lost [1, 40, 56]. Similarly, our results concurred that MXEs can show a mutually exclusive pattern only in some cell types.

Identification of coordinately spliced exon blocks

By definition, the splicing nodes considered in scASfind only involve a single event. Though a single event can lead to drastic shifts of gene function [57], splicing events are often coordinated, resulting in multiple consecutive exons being simultaneously included or excluded. Hereafter we refer to such coordinated groups of splicing nodes as node blocks. Coordinated events are more likely to have a more substantial impact on protein function as a larger proportion of coding sequences are affected, but when using a splicing node representation, they are difficult to detect as one must find a stretch of consecutive nodes. Given the large number of nodes across the transcriptome and the high noise level of splicing quantification, this search can be very time-consuming.

We used scASfind to detect node blocks. For each gene, we first identify consecutive nodes of type "core exon" with similar mean and standard deviation of their PSI values (the default is to require the absolute differences for both to be <0.1 for all exons in the block). Next, the search is expanded to identify additional neighboring nodes to find cell type specific blocks. Events where a block of nodes is coordinately spliced-in (the "above" events) and spliced-out (the "below" events) are detected separately. To ensure high quality results, we only keep blocks composed of at least three nodes from different actual exons. Cell type specific node blocks that are detected in over half of the cell pools are reported as high confidence blocks.

Overall, we detected 263 node blocks with lengths ranging from 3 to 21 in the mouse cortex dataset, with 8 high-confidence ones (3–5 splicing nodes long). For the mouse embryo data, we found 306 blocks containing 3–26 nodes and 14 high-confidence ones with lengths ranging from 3 to 6. For the human fetal liver data, there were 526 node blocks ranging from 3 to 37 nodes, 19 of which were high confidence with lengths from 3 to 9 (Fig. 6a–f). For example, in the mouse cortex data, we found that node 5–7 in *Haus7* is significantly spliced-out in L2_Ngb and L5_Pde1c while it is spliced-in in Sncg (Fig. 6d). Reassuringly, this is in line with two documented isoforms as shown in the GENCODE annotation [58, 59] (Additional file 1: Fig. S4). In the human fetal liver, we found a block of exons between chr2:95,895,399–95,901,206 in *ANKRD36C* that is spliced-in for dend_cell_cycling, also in concordance with known isoforms (Additional file 1: Fig. S5).

We proceed to analyze whether the detected highly confident node blocks correspond to known isoforms, and what are their functional implications on protein domains in the mouse embryo data. We detected known coordinated events in *Ttn* in

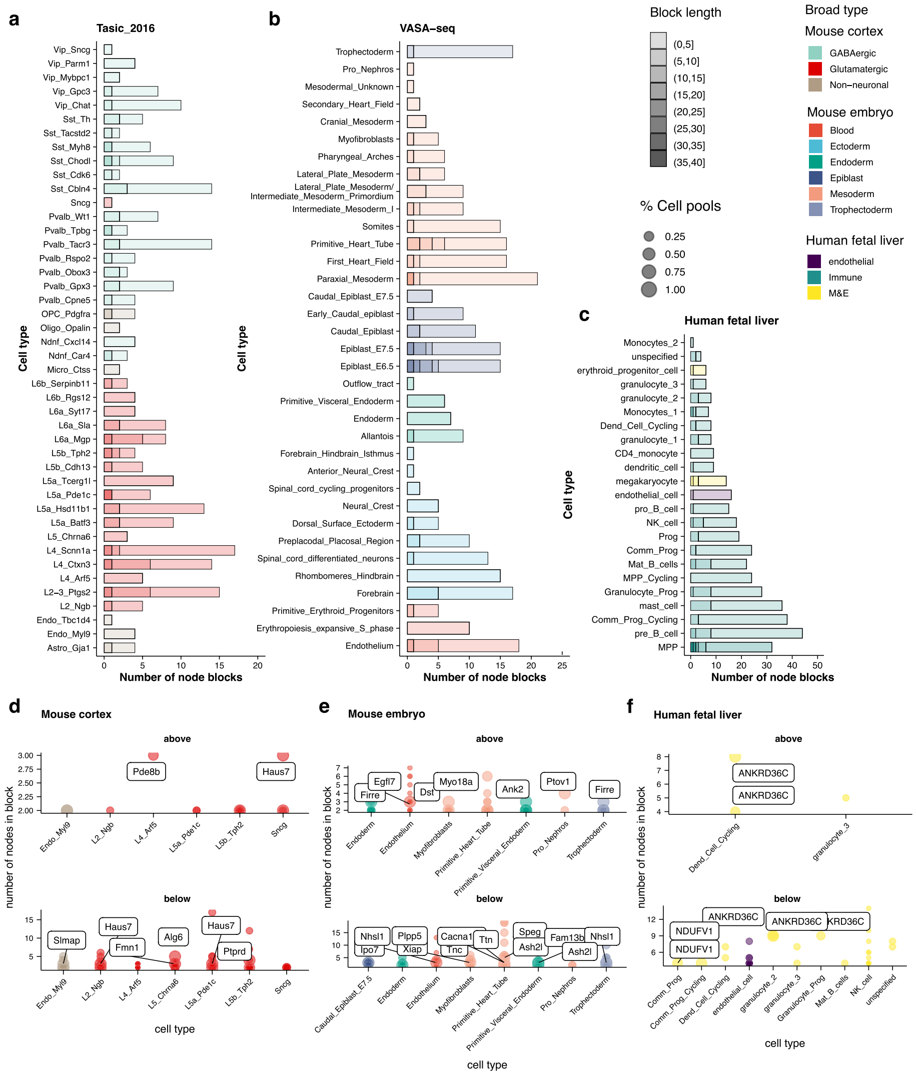

Fig. 6 Cell type-specific coordinated splicing. Overall, **a** 263 node blocks were detected in the mouse cortex dataset, **b** 306 in the mouse embryo, and **c** 526 in the human fetal liver. The gene name of highly confident node blocks, which include at least 3 nodes from different exons and were detected in ≥ 50% cell pools, are shown in **d**–**f** for each dataset. In **d**–**f**, "above" means that the node block is significantly spliced-in in the respective cell type and "below" means significantly splice-out of the node block

primitive_heart_tube and first_heart_field. This corresponds to a regulated isoform switching event during heart development, upon which the stiffness of the protein changes [60, 61]. Another key gene for heart development is *Dst*. Here we found a block including 5 exons and spanning 3 of the 7 major protein domains significantly spliced-in in myofibroblasts (Fig. 7A). This event is in line with the documented muscle-specific *Dst-b* isoform [62]. Mutation studies have shown that *Dst-b* is essential for strained muscle maintenance [63]. We have also found a node exclusion block in *Myo18A*, which is consistent with shorter annotated isoforms that are derived from an internal transcription start site [64, 65] (Fig. 7B). One of these shorter isoforms, known as *Myo18Aγ*, lacks the PDZ-containing N-terminus but includes an alternative N-terminal extension [66] and showed well-marked cell type specific profile associated to primitive_heart_tube. In

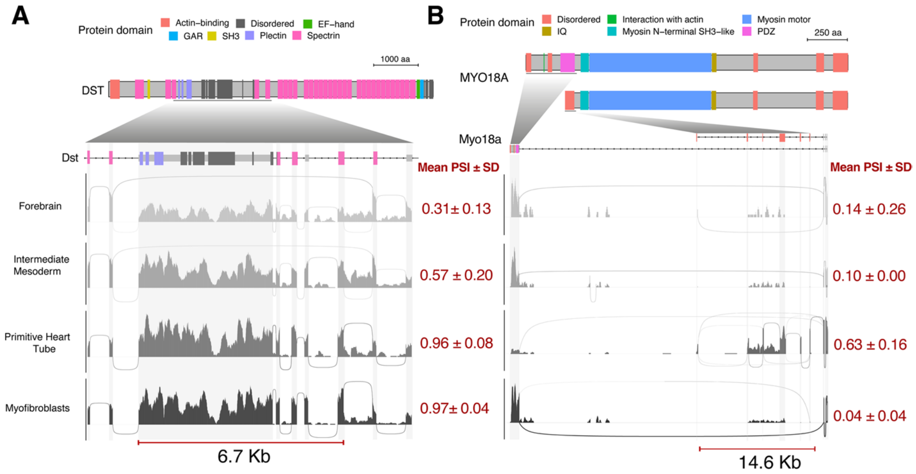

Fig. 7 Coordinated inclusion of splicing nodes across mouse embryonic cell types. Sashimi plots showing the read coverage and splice site usage across **A** *Dst* and **B** *Myo18a* transcripts. Top schematics show domains annotated for these proteins and coordinated splice nodes are highlighted with a red segment (bottom). Different gray shades are meant to distinguish cell types. Mean ± SD of PSI values across all coordinated splice nodes are indicated in red. PSI, percent spliced-in; SD, standard deviation

general, the detected highly confident node blocks correspond well to known regulated isoforms. The analysis demonstrated the versatility of scASfind to detect diverse isoform switching events.

Even though some blocks are detected in < 50% of cell pools, they often match documented isoforms in the annotation. For example, in the human fetal liver, 20% of NK_cell cell pools did not include chr6:42,632,552–42,655,723 (14 splicing nodes) in the *UBR2* gene, in line with an isoform corresponding to early termination (Additional file 1: Fig. S6). We also found that 25% of endothelial cells express a shorter isoform of *TBC1D19* (Additional file 1: Fig. S7). Moreover, we have detected several high confidence blocks that suggest undocumented isoforms. For example, four exons between chr7:44,864,913–44,865,509 in *Ptov1* are shown to be coordinately included in all cell pools of Pro_Nephros in mouse embryos while there are no recorded isoforms of this gene in the GENCODE [59] annotation (Fig. 6f, Additional file 1: Figs. S8, 9). This could have relevance to human biology since *PTOV1* has been associated with prostate cancer, and the splicing event is likely to have a disruptive impact on one of the two major domains involved in the interaction with multiple other genes [67].

Discussion

Splicing is a highly regulated process with a key role in cellular identity and function [5, 68]. Here we present scASfind, a toolkit for mining cell type-specific splicing patterns from large, single-cell, full-length transcriptomics datasets. It is challenging to analyze splicing in scRNA-seq data due to the vast number of splicing nodes and the high degree of sparsity. To overcome these challenges, we utilized cell pooling and data compression to build an index which can support efficient queries of the cell type pattern of splicing events. We demonstrate that an index for thousands of cells can be created in 20–30 min, resulting in compression by 2–3 orders of magnitude, while at the same time speeding up queries by hundreds of folds. Building on this data structure, we provide functions for discovering cell type-specific splicing events such as finding marker nodes, mutually exclusive exons, or coordinately spliced node blocks. Using mouse cortex, mouse embryonic development, and human fetal liver datasets, we demonstrated scASfind's utility for carrying out tasks that would have been prohibitive without the tool.

Quantification of individual splicing events, compared with transcript-level analysis, is more tractable with short-read data and does not rely on complete annotation models [13, 15]. The PSI quantification used by scASfind comes from the event-level splicing quantification tool Whippet. Whippet is an established method which is efficient and showed high recall in a benchmark study [69]. However, the algorithm only detects and analyzes annotated AS events in its contiguous splice graphs index [13] and it is less effective on detecting events de novo [69]. Therefore, scASfind also only detected splicing patterns of nodes existing in the Whippet index.

Current single-cell platforms typically focus on sequencing only the $3'$ or $5'$ ends of transcripts, leaving alternative splicing largely unexplored at the single-cell level. Our analysis of the scASfind splicing node markers strongly suggests that single-node splicing patterns can provide higher cell type precision than gene expression. This is particularly relevant to tracing rare cell types. By leveraging rare cell types identified through specialized algorithms [70, 71], we can explore specific splicing patterns using scASfind, offering potential for higher precision and experimental validation. For instance, studies have demonstrated subtype-specific splicing in neurons [72–74], reinforcing the utility of splicing patterns in distinguishing neuron subtypes. Understanding AS events with strong cell population specificity is crucial for effectively studying cellular heterogeneity. Additionally, the systematic identification of MXEs and tissue-specific coordinated splicing events provides insights into cell type-specific AS regulation, enriching our understanding of the regulatory landscape.

The development of high throughput full-length protocols such as VASA-seq [12] will likely open opportunities for splicing analysis in a diversity of biological systems. Moreover, several studies have utilized long reads technologies for single-cell studies [75–77], allowing an entire transcript to be captured by a single read. We believe that these advances will allow single-cell studies to quantify alternative splicing events, but for this to become feasible novel computational methods are required. Given its efficient memory usage, low run times, and convenient search functionality, we believe that scASfind will be a valuable tool for researchers to decipher cell type-specific splicing using scRNA-seq data.

Conclusions

We provide scASfind, a freely available software for mining cell type-specific alternative splicing events in full-length scRNA-seq data. It utilizes an efficient data structure to detect marker splicing nodes and enables exhaustive searches of MXEs and node blocks.

Applying scASfind to three datasets from mouse and human demonstrated the high precision of marker splicing nodes compared to the more widely used marker genes. We also found known and novel MXEs and node blocks that show cell type specific splicing patterns. Splicing analysis with scASfind facilitates discovery of cell type-specific splicing events that may have functional implications.

Methods

AS quantification across cell types

To quantify AS events across cell types, we configured and ran MicroExonator's single-cell module, as described in [78]. As part of this workflow, MicroExonator qualifies AS events using Whippet [13] across cell pools derived from annotated cell clusters. Using this protocol, we processed mouse scRNA-seq data from brain visual cortex [29] and whole embryos [12], as well as scRNA-seq data derived from human immunophenotypic blood cells from fetal liver and bone marrow [30]. We used genome assembly mm10 and GENCODE transcript annotation v16 to process mouse scRNA-seq data. For human scRNA-seq analyses, we used genome assembly hg38 and GENCODE transcript annotation v19.

Filter for confidently quantified events

Before encoding the PSI data, we first filter for confidently quantified events. The sparsity of scRNA-seq data often results in an insufficient number of reads spanning splicing junctions that can be used to calculate node PSI values. By default, we require at least 10 reads available for PSI quantification. This gives roughly a confidence interval of PSI < 0.5 from Whippet.

Create a scASfind index

scASfind builds four types of data structures from the input data, and together they form a queryable index.

1. The splicing node x cell pool differential PSI matrices

   For each node, we first calculate the mean PSI across all cell pools, then calculate the difference from the mean for all PSI values. A 0.2 deviation was used as the default threshold to select sufficiently deviated events. Secondly, we separate nodes with differential inclusion (the "above" events) and those with differential exclusion (the "below" events). Both metrics are then multiplied by 100 so that the value is in the range of 0, 100 for the compression.

   The splicing node x cell pool differential PSI (ΔPSI) matrices are independently compressed using the strategy in scfind [28]. The compression is a two-step process: (1) storing the positions of non-zero values are compressed by Elias-Fano encoding, and (2) the actual differential PSI value is represented as quantiles of a log-normal distribution (Additional file 1: Fig. S10). The mean and variance of the log-normal distribution, along with quantiles of the original ΔPSI values, are stored. The first step is lossless while the second step is lossy. The approximation of actual ΔPSI in the index makes it possible to retrieve the approximate PSI value when giving the user the ability to tune the size of the storage based on the number of bits used for the quantization (default 2 bits).

2. The mean and standard deviation PSI values per node
   We store the mean and standard deviation for each dataset and node for retrieval of raw PSI values and for speeding up searches of MXEs and node blocks based on expected patterns in mean and standard deviation.
3. The mask for NA values from PSI quantification
   Since we only encode differential PSI values, cell pools with PSI values close to the dataset-wise mean are excluded. However, cell pools where the PSI values are unquantified (NA) or below the required number of reads for confident quantification are also excluded. Distinguishing these two circumstances is required to enable retrieval of raw PSI values from the index. For this purpose, we use a binary mask matrix. In this matrix, 1 indicates the cell pools with PSI value equal to the dataset-wise mean and 0 indicates unquantified. Typically, unquantified events are more frequent than equal mean events, resulting in a sparse matrix.
4. The annotation of nodes
   We use ENSEMBL [79] via the R package biomaRt (V2.46.3) to obtain annotations of all nodes included in the index to enable quick interpretation of results.
   In addition, metadata for each cell, providing information about its annotated cell type or state is required. The buildAltSpliceIndex function in scASfind takes the cell pool-by-splicing node differential PSI matrices and a table with the cell type annotation to build an index object. The "above" and "below" index objects are stored as two datasets, and the three other metrics are stored in the metadata slot in the scASfind object.

Benchmark of file size, index build time, and node search time

We benchmark the efficiency of the scASfind index, compared with a basic approach utilizing only R and Seurat functions. For file size, we take the sum of on-disk space taken by the raw PSI matrix (as.tsv files) and the three metadata objects (as.rds files) as "raw PSI," and the complete scASfind index (as.rds object), including the same three metadata objects stored in the metadata slot as "scASfind index." All file sizes are obtained with the "file.info" function in R. For index build time, we run the scASfind build index script on a high-performance computing cluster (Rocky Linux 8.5) for the three datasets with 4 cores and maximum 2 GB memory, 10 processes, and 200 threads. The time to build the index naturally depends on the computational resources available. For differential events search time, we randomly select 5, 10, 50, 100, and 200 splicing nodes, and search for cell pools with an above-mean PSI of any of the nodes using either the naive approach or the scASfind index. The elapsed times were measured in 30 repetitions per query length.

Cell type marker node search

We use a precision-recall framework to search for nodes that are specific to different cell types. For each node, we count the number of cell pools with the relative inclusion/exclusion of this node in a cell type of interest compared to all other cell types. We use precision, recall, and F1 scores to evaluate how well the node distinguishes the cell type of interest from all other cell types. A true positive (TP) is when a node is included or excluded in a cell pool from the cell type of interest for spliced-in and spliced-out events, respectively. False positives (FP) are when the same node is included/excluded in cell pools of other cell types, and false negatives (FN) are cell pools from the same cell type in which the node is not detected as included or excluded. The precision score is calculated by:

$$precision = \frac{TP}{TP + FP} \tag{1}$$

The recall score is calculated by:

$$recall = \frac{TP}{TP + FN} \tag{2}$$

The F1 score is the harmonic mean of precision and recall score:

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{3}$$

By default, we use F1 scores as a balanced metric to rank all nodes for each cell type to indicate the best marker nodes for either inclusion or exclusion events.

Comparing PSI and gene expression in splicing marker nodes

We calculate the top 20 gene expression markers (ranked by F1 score) using scfind in the mouse cortex data. For the genes containing splicing marker nodes, we used the R package Seurat (V4.1.0) to obtain the scaled expression values. Then, we used ggplot2 (V3.3.3), viridislite (V0.4.0), and cowplot (V1.1.1) to create the violin plot of PSI and scaled expression values.

MARVEL analysis

We ran MARVEL [26] (v2.0.5) following the tutorial for plate-based sequencing methods (https://wenweixiong.github.io/MARVEL_Plate.html). The function CompareValues were used to perform pairwise differential splicing analysis between cell types using the “ad” algorithm. Only exon-skipping type events were used to compare with CEs in scASfind.

VastDB comparison

We downloaded the main PSI table of the mouse data in VastDB [10] (mm10) (https://vastdb.crg.eu/wiki/Downloads#AS_events_3). Splicing nodes that were cell type markers in scASfind analysis of the mouse cortex data were subtracted from this table and subjected to tissue specificity analysis. PPCA was performed using pcaMethods (v1.64.0).

Detect cell type-specific mutually exclusive exons

Detection of cell type-specific MXEs is based on a hypergeometric test with the “hyperQueryCellTypes” function in scASfind. The hypergeometric distribution models the probability of k success in n draws without replacement, from a finite population with N subjects and K of them contains the pattern. In our case, k is the number of cell pools in a cell type which have the splicing pattern, n is the number of cell pools in that cell type, K is the total number of cell pools in which the splicing pattern is detected, and N is the total number of cell pools. A pattern with a hypergeometric test P value ≤ 0.05 in a cell type is considered significant.

We use a mutually exclusive combination of splicing nodes (include one and exclude the other) as the pattern in the hypergeometric test to detect MXEs. The query is performed exhaustively for all pairs of exons in the dataset. To reduce the search space, we first filter all the possible node pairs by (1) having a mean PSI sum of 1 ± 0.1 and (2) having a < 0.1 difference in the PSI standard deviation. Then, we query for significant cell types for the potential pairs of MXEs. Pairs with at least one cell type significant in one of the two possible patterns are kept as candidates.

Further filters are applied for candidate MXE pairs. First, we require the MXE pattern to be found in $\geq 50\%$ of the pools in the significant cell type. Then, we require the difference of absolute PSI value in the pair to be ≥ 0.5 to be considered high confidence.

Detecting cell type-specific coordinately spliced-in exons

We scan all genes from the $5'$ of all annotated “core exon” nodes in the scASfind index to find coordinately spliced-in exons. We extend an exon block by requiring the next exon to have at most ± 0.1 difference with both the mean and standard deviation of the PSI of the previous block of exons. If this criterion is not fulfilled, we initiate a new block, and the previously constructed block is tested for cell type specificity using the hypergeometric test as described previously. Blocks significant in at least one cell type are kept. Finally, we use a node-to-exon mapping table from Whippet [13] to combine nodes belonging to the same actual exon. If the resulting block contains at least 3 exons, we propose it as a potential coordinated sliced-in exon block. We also require the block to be found in $> 50\%$ of the pools in the significant cell type for it to be highly confident.

The impact of pool size on PSI quantification and scASfind results

We ran Whippet, followed by scASfind, using a cell pool size of 5, 10, 15, and 20 on the mouse cortex data. This was specified in the config file with cells_pseudobulks in the MicroExonator workflow.

First, we examined the percentage of nodes quantified among all annotated nodes (Additional file 1: Fig. S11a). We found that despite a slight increase in the total percentage of nodes quantified as the pool size increased, the percentage of confidently quantified nodes (with ≥ 10 reads) remained stable and saw a slight decrease at pool size 20. This is further supported by the fact that the coverage distribution of confidently quantified nodes remained similar across the pool sizes (Additional file 1: Fig. S11b). Additionally, comparing the percentage of nodes quantified among all nodes in the three datasets suggested that the selected pool sizes gave comparable numbers of confidently quantified nodes: 20.47 ± 4.30, 30.63 ± 13.78, and 15.72 ± 5.63 (mean ± stddev).

In summary, using a pool size 5 for the mouse cortex data ensures a balance between achieving sufficient coverage for PSI quantification and retaining the abundance variation between cell types.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03323-6.

Additional file 1: Supplementary figures S1–S12 for the publication "Mining alternative splicing patterns in scRNA-seq data using scASfind"

Additional file 2: Peer review history

Acknowledgements
We would like to thank Jae-Won Cho and Ruben Chazarra-Gil for feedback on the manuscript and Nikolaos Patikas for help in testing the software.

Review history
The review history is available as Additional file 2.

Peer review information
Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
GP, JTHL, and MH conceived the project. YS wrote the code and analyzed the data with assistance from GP and JTHL. YS and MH wrote the manuscript with input from GP and JTHL. MH supervised the research.

Authors' X handles
X handles: @YuyaoSong123 (Yuyao Song); @ge_parada (Guillermo Parada).

Funding
YS was funded by the EMBL international PhD program, GP and JTHL were funded by the Wellcome Trust, and MH was funded by the Wellcome Trust and the Evergrande Center.

Availability of data and materials
We provide scASfind (27) freely available via zenodo (https://doi.org/10.5281/zenodo.8241682) or GitHub (https://github.com/hemberg-lab/scASfind) under an MIT license. MicroExonator (78) is freely available via https://github.com/hemberg-lab/MicroExonator. All datasets used in this study are publicly available. Mouse cortex data is accessible from Gene Expression Omnibus (GSE71585) (29). Mouse embryo data is accessible from Gene Expression Omnibus (GSE176588) (12). Human fetal liver data is accessible from ArrayExpress (E-MTAB-9067) (30).

Declarations

Ethics approval and consent to participate
Ethics approval is not applicable to this study.

Competing interests
None declared.

Received: 17 August 2023 Accepted: 26 June 2024
Published online: 29 July 2024

References

1. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature. 2010;463(7280):457–63.
2. Dai A, Temporal S, Schulz DJ. Cell-specific patterns of alternative splicing of voltage-gated ion channels in single identified neurons. Neuroscience. 2010;168(1):118–29.
3. Ergun A, Doran G, Costello JC, Paik HH, Collins JJ, Mathis D, et al. Differential splicing across immune system lineages. Proc Natl Acad Sci U S A. 2013;110(35):14324–9.
4. Nikonova E, Kao S-Y, Spletter ML. Contributions of alternative splicing to muscle type development and function. Semin Cell Dev Biol. 2020;1(104):65–80.
5. Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol. 2017;18(7):437–51.
6. Zhang X, Chen MH, Wu X, Kodani A, Fan J, Doan R, et al. Cell-type-specific alternative splicing governs cell fate in the developing cerebral cortex. Cell. 2016;166(5):1147-1162.e15.
7. Scotti MM, Swanson MS. RNA mis-splicing in disease. Nat Rev Genet. 2015;17(1):19–32.
8. Ren P, Lu L, Cai S, Chen J, Lin W, Han F. Alternative splicing: a new cause and potential therapeutic target in autoimmune disease. Front Immunol. 2021;17(12): 713540.
9. Zhang Y, Qian J, Gu C, Yang Y. Alternative splicing and cancer: a systematic review. Signal Transduct Target Ther. 2021;6(1):78.
10. Tapial J, Ha KCH, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A, et al. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. 2017;27(10):1759–68.

11. Picelli S. Full-length single-cell RNA sequencing with Smart-seq2. Methods Mol Biol. 2019;1979:25–44.
12. Salmen F, De Jonghe J, Kaminski TS, Alemany A, Parada GE, Verity-Legg J, et al. High-throughput total RNA sequencing in single cells using VASA-seq. Nat Biotechnol. 2022;27:1–14.
13. Sterne-Weiler T, Weatheritt RJ, Best AJ, Ha KCH, Blencowe BJ. Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop. Mol Cell. 2018;72(1):187-200.e6.
14. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511–5.
15. Vaquero-Garcia J, Barrera A, Gazzara MR, González-Vallinas J, Lahens NF, Hogenesch JB, et al. A new view of transcriptome complexity and regulation through the lens of local splicing variations. Elife. 2016;1(5): e11752.
16. Shen S, Park JW, Lu Z-X, Lin L, Henry MD, Wu YN, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A. 2014;111(51):E5593–601.
17. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. 2010;7(12):1009–15.
18. Deng N, Zhu D. dSpliceType: a multivariate model for detecting various types of differential splicing events using RNA-Seq. In: Basu M, Pan Y, Wang J, editors. Bioinformatics Research and Applications. ISBRA 2014; 2014 Jun 28-30; Zhangjiajie, China. Heidelberg (DE): Springer Cham; 2014. p. 322–33.
19. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 2018;19(1):40.
20. Welch JD, Hu Y, Prins JF. Robust detection of alternative splicing in a population of single cells. Nucleic Acids Res. 2016;44(8): e73.
21. Song Y, Botvinnik OB, Lovci MT, Kakaradov B, Liu P, Xu JL, et al. Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. Mol Cell. 2017;67(1):148-161.e5.
22. Huang Y, Sanguinetti G. BRIE: transcriptome-wide splicing quantification in single cells. Genome Biol. 2017;18(1):123.
23. Huang Y, Sanguinetti G. BRIE2: computational identification of splicing phenotypes from single-cell transcriptomic experiments. Genome Biol. 2021;22(1):251.
24. Dehghannasiri R, Olivieri JE, Damljanovic A, Salzman J. Specific splice junction detection in single cells with SICILIAN. Genome Biol. 2021;22(1):219.
25. Olivieri JE, Dehghannasiri R, Salzman J. The SpliZ generalizes 'percent spliced in' to reveal regulated splicing at single-cell resolution. Nat Methods. 2022;19(3):307–10.
26. Wen WX, Mead AJ, Thongjuea S. MARVEL: an integrated alternative splicing analysis platform for single-cell RNA sequencing data. Nucleic Acids Res. 2023;51(5): e29.
27. Song Y, Parada GE, Lee JTH, Hemberg M. Mining alternative splicing patterns in scRNA-seq data using scASfind. scASfind. 2023. Available from: https://github.com/hemberg-lab/scASfind/10.5281/zenodo.8241681.
28. Lee JTH, Patikas N, Kiselev VY, Hemberg M. Fast searches of large collections of single-cell data using scfind. Nat Methods. 2021;18(3):262–71.
29. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci. 2016;19(2):335–46.
30. Ranzoni AM, Tangherloni A, Berest I, Riva SG, Myers B, Strzelecka PM, et al. Integrative single-cell RNA-Seq and ATAC-Seq analysis of human developmental hematopoiesis. Cell Stem Cell. 2021;28(3):472-487.e7.
31. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10(11):1096–8.
32. Hayashi T, Ozaki H, Sasagawa Y, Umeda M, Danno H, Nikaido I. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat Commun. 2018;9(1):1–16.
33. Hagemann-Jensen M, Ziegenhain C, Chen P, Ramsköld D, Hendriks G-J, Larsson AJM, et al. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. Nat Biotechnol. 2020;38(6):708–14.
34. Hahaut V, Pavlinic D, Carbone W, Schuierer S, Balmer P, Quinodoz M, et al. Fast and highly sensitive full-length single-cell RNA sequencing using FLASH-seq. Nat Biotechnol. 2022;40(10):1447–51.
35. Parada GE, Munita R, Georgakopoulos-Soares I, Fernandes HJR, Kedlian VR, Metzakopian E, et al. MicroExonator enables systematic discovery and quantification of microexons across mouse embryonic development. Genome Biol. 2021;22(1):43.
36. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14(5):483–6.
37. Hakim NHA, Majlis BY, Suzuki H, Tsukahara T. Neuron-specific splicing. Biosci Trends. 2017;11(1):16–22.
38. Tharp CA, Haywood ME, Sbaizero O, Taylor MRG, Mestroni L. The giant protein titin's role in cardiomyopathy: genetic, transcriptional, and post-translational modifications of TTN and their contribution to cardiac disease. Front Physiol. 2019;28(10):1436.
39. Hatje K, Rahman R-U, Vidal RO, Simm D, Hammesfahr B, Bansal V, et al. The landscape of human mutually exclusive splicing. Mol Syst Biol. 2017;13(12):959.
40. Pohl M, Bortfeldt RH, Grützmann K, Schuster S. Alternative splicing of mutually exclusive exons—a review. Biosystems. 2013;114(1):31–8.
41. Kalsotra A, Cooper TA. Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet. 2011;12(10):715–29.
42. Xu Q, Modrek B, Lee C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res. 2002;30(17):3754–66.
43. Rodriguez JM, Pozo F, di Domenico T, Vazquez J, Tress ML. An analysis of tissue-specific alternative splicing at the protein level. PLoS Comput Biol. 2020;16(10): e1008287.
44. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470–6.
45. Sommer B, Keinänen K, Verdoorn TA, Wisden W, Burnashev N, Herb A, et al. Flip and flop: a cell-specific functional switch in glutamate-operated channels of the CNS. Science. 1990;249(4976):1580–5.

46. Koike M, Tsukada S, Tsuzuki K, Kijima H, Ozawa S. Regulation of kinetic properties of GluR2 AMPA receptor channels by alternative splicing. J Neurosci. 2000;20(6):2166–74.
47. Wright A, Vissel B. The essential role of AMPA receptor GluR2 subunit RNA editing in the normal and diseased brain. Front Mol Neurosci. 2012;11(5):34.
48. Hadzic M, Jack A, Wahle P. Ionotropic glutamate receptors: which ones, when, and where in the mammalian neocortex. J Comp Neurol. 2017;525(4):976–1033.
49. Monyer H, Seeburg PH, Wisden W. Glutamate-operated channels: developmentally early and mature forms arise by alternative splicing. Neuron. 1991;6(5):799–810.
50. Bark IC, Hahn KM, Ryabinin AE, Wilson MC. Differential expression of SNAP-25 protein isoforms during divergent vesicle fusion events of neural development. Proc Natl Acad Sci U S A. 1995;92(5):1510–4.
51. Prescott GR, Chamberlain LH. Regional and developmental brain expression patterns of SNAP25 splice variants. BMC Neurosci. 2011;28(12):35.
52. Irfan M, Gopaul KR, Miry O, Hökfelt T, Stanton PK, Bark C. SNAP-25 isoforms differentially regulate synaptic transmission and long-term synaptic plasticity at central synapses. Sci Rep. 2019;9(1):1–14.
53. Boschert U, O'Shaughnessy C, Dickinson R, Tessari M, Bendotti C, Catsicas S, et al. Developmental and plasticity-related differential expression of two SNAP-25 isoforms in the rat brain. J Comp Neurol. 1996;367(2):177–93.
54. Kremerskothen J, Teber I, Wendholt D, Liedtke T, Böckers TM, Barnekow A. Brain-specific splicing of α-actinin 1 (ACTN1) mRNA. Biochem Biophys Res Commun. 2002;295(3):678–81.
55. Waites GT, Graham IR, Jackson P, Millake DB, Patel B, Blanchard AD, et al. Mutually exclusive splicing of calcium-binding domain exons in chick alpha-actinin. J Biol Chem. 1992;267(9):6263–71.
56. Sammeth M, Foissac S, Guigó R. A general definition and nomenclature for alternative splicing events. PLoS Comput Biol. 2008;4(8): e1000147.
57. Zhang M, Zhu B, Davie J. Alternative splicing of MEF2C pre-mRNA controls its activity in normal myogenesis and promotes tumorigenicity in rhabdomyosarcoma cells. J Biol Chem. 2015;290(1):310–24.
58. Karolchik D, Hinrichs AS, Kent WJ. The UCSC genome browser. Curr Protoc Bioinformatics. 2009 Dec;Chapter 1:Unit1.4.
59. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760–74.
60. Lahmers S, Wu Y, Call DR, Labeit S, Granzier H. Developmental control of titin isoform expression and passive stiffness in fetal and neonatal myocardium. Circ Res. 2004;94(4):505–13.
61. Opitz CA, Leake MC, Makarenko I, Benes V, Linke WA. Developmentally regulated switching of titin size alters myofibrillar stiffness in the perinatal heart. Circ Res. 2004;94(7):967–75.
62. Leung CL, Zheng M, Prater SM, Liem RK. The BPAG1 locus: alternative splicing produces multiple isoforms with distinct cytoskeletal linker domains, including predominant isoforms in neurons and muscles. J Cell Biol. 2001;154(4):691–7.
63. Yoshioka N, Kurose M, Yano M, Tran DM, Okuda S, Mori-Ochiai Y, et al. Isoform-specific mutation in dystonin-b gene causes late-onset protein aggregate myopathy and cardiomyopathy. Elife. 2022;9(11): e78419.
64. Guzik-Lendrum S, Heissler SM, Billington N, Takagi Y, Yang Y, Knight PJ, et al. Mammalian myosin-18A, a highly divergent myosin. J Biol Chem. 2013;288(13):9532–48.
65. Ouyang Z, Zhao S, Yao S, Wang J, Cui Y, Wei K, et al. Multifaceted function of myosin-18, an unconventional class of the myosin superfamily. Front Cell Dev Biol. 2021;9(9): 632445.
66. Horsthemke M, Nutter LMJ, Bachg AC, Skryabin BV, Honnert U, Zobel T, et al. A novel isoform of myosin 18A (Myo18Aγ) is an essential sarcomeric protein in mouse heart. J Biol Chem. 2019;294(18):7202–18.
67. Cánovas V, Lleonart M, Morote J, Paciucci R. The role of prostate tumor overexpressed 1 in cancer progression. Oncotarget. 2017;8(7):12451–71.
68. Mazin PV, Khaitovich P, Cardoso-Moreira M, Kaessmann H. Alternative splicing during mammalian organ development. Nat Genet. 2021;53(6):925–34.
69. Fenn A, Tsoy O, Faro T, Rößler FLM, Dietrich A, Kersting J, et al. Alternative splicing analysis benchmark with DICAST. NAR Genom Bioinform. 2023;5(2):lqad044.
70. Wegmann R, Neri M, Schuierer S, Bilican B, Hartkopf H, Nigsch F, et al. CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. Genome Biol. 2019;20(1):142.
71. Wang S, Li H, Zhang K, Wu H, Pang S, Wu W, et al. scSID: A lightweight algorithm for identifying rare cell types by capturing differential expression from single-cell sequencing data. Comput Struct Biotechnol J. 2024;23:589–600.
72. Sato Y, Iijima Y, Darwish M, Sato T, Iijima T. Distinct expression of SLM2 underlies splicing-dependent trans-synaptic signaling of neurexin across GABAergic neuron subtypes. Neurochem Res. 2022;47(9):2591–601.
73. Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, et al. ASCOT identifies key regulators of neuronal subtype-specific splicing. Nat Commun. 2020;11(1):137.
74. Murphy D, Cieply B, Carstens R, Ramamurthy V, Stoilov P. The Musashi 1 controls the splicing of photoreceptor-specific exons in the vertebrate retina. PLoS Genet. 2016;12(8): e1006256.
75. Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. Genome Biol. 2021;22(1):310.
76. Philpott M, Watson J, Thakurta A, Brown T, Oppermann U, Cribbs AP. Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq. Nat Biotechnol. 2021;39(12):1517–20.
77. Lebrigand K, Bergenstråhle J, Thrane K, Mollbrink A, Meletis K, Barbry P, et al. The spatial landscape of gene expression isoforms in tissue sections. Nucleic Acids Res. 2023;51(8):e47
78. Parada GE, Hemberg M. Identification and quantification of microexons using bulk and single-cell RNA-Seq data. Methods Mol Biol. 2022;2537:129–47.
79. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res. 2022;50(D1):D988-95 2023.

Publisher's Note