The SZS is an efficient statistical method to identify regulated splicing events in droplet-based RNA sequencing

Julia Eve Olivieri[1,2], Roozbeh Dehghannasiri[1], Julia Salzman[1,3,†]

Author affiliation
[1]Department of Biochemistry, Stanford University, Stanford, CA 94305
[2]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305
[3]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305

[†]Corresponding author: julia.salzman@stanford.edu

Abstract

To date, the field of single-cell genomics has viewed robust splicing analysis as completely out of reach in droplet-based platforms, preventing biological discovery of single-cell regulated splicing. Here, we introduce a novel, robust, and computationally efficient statistical method, the Splicing Z Score (SZS), to detect differential alternative splicing in single cell RNA-Seq technologies including 10x Chromium. We applied the SZS to primary human cells to discover new regulated, cell type-specific splicing patterns. Illustrating the power of the SZS method, splicing of a small set of genes has high predictive power for tissue compartment in the human lung, and the SZS identifies un-annotated, conserved splicing regulation in the human spermatogenesis. The SZS is a method that can rapidly identify regulated splicing events from single cell data and prioritize genes predicted to have functionally significant splicing programs.

Main Text:

Splicing is a core function of eukaryotic genes that generates proteins with diverse and even opposite functions[1], changes translation efficiency[2], controls localization[3], and generates non-coding RNAs[4]. Very few genes' splicing programs have been characterized at single-cell resolution, and the function of splicing remains a critical open problem in biology[5]. Constant advances in single-cell RNA sequencing (scRNA-Seq) technology now provide an unprecedented opportunity to understand how splicing is regulated at the single cell level.

The enormous complexity of splicing in eukaryotic genomes makes it challenging to precisely quantify RNA isoform expression and its differential regulation. With few exceptions[6], the field has typically either attempted to directly quantify exon inclusion[7,8], or to first estimate isoform expression using model-based approaches[9,10] and then perform differential splicing analysis. These methods suffer from technical issues including high variance of point estimates that become more extreme in single-cell experiments and hamper robust statistical inference. This, together with the 3' read bias in droplet-based platforms have led to the view that robust differential splicing in droplet-based scRNA-Seq is impossible[5,11,12].

Here, we introduce a new, theoretically grounded approach, the Splicing Z Score (SZS), that quantifies statistical deviation of splicing patterns at a single-cell level to analyze single cell data including 10x Chromium (10x). The SZS is a direct statistical method that detects differential splicing at single cell resolution. Each pair of gene and cell is assigned an SZS which is large and negative if introns in that gene are statistically shorter than the average across all cells, and large and positive if the introns are statistically longer than average (Fig. 1A-B, Methods).

The SZS sidesteps isoform estimation to draw inferences from sparse and strongly 3'-biased data, such as that from the 10x sequencing platform. Also, the score is nonparametric, avoiding requirements to estimate nuisance parameters such as isoform abundance[13] and resampling to estimate confidence intervals for these parameters that would be required for statistical inference[14].

The theoretical properties of the SZS allow it to be directly integrated into downstream analyses such as clustering or correlation with scalar variables such as gene expression or pseudotime assignment without the need for computationally intensive resampling approaches (Fig. 1C,E). Downstream analysis with the SZS which we call "the SZS pipeline" also detects cell-type-specific regulation based on annotated cell types using the median SZSs for a gene across the cells within a cell type and other statistical measures such as the effect size, significance value, and, in the case of data with biological replicates, an FDR (Fig. 1D). To identify splice sites that drive the differential splicing, the SZS pipeline provides splice-site-level scores for donors and acceptors that drive observed extreme values of the SZS.

Unlike previous tools for analyzing differential RNA splicing[7,8,10], the SZS gathers strength across all spliced reads from a gene to increase power to detect differential splicing. This decreases the variance of the SZS compared to point estimates of exon inclusion, a ubiquitous problem in low-sampling depth scenarios such as scRNA-Seq[12]. It also avoids parametric assumptions that can generate false positive statistical significance[15].

The SZS quantifies cell-type-specific splicing for each cell-gene pair by (1) converting each read aligning to a splice junction to a rank based on the relative size of the intron compared to the set of observed introns for that donor and acceptor (Methods); (2) converting the rank to a residual measuring its statistical deviation compared to "the typical" intron length rank (Methods); (3) statistically grounded scaling and aggregating of these values to a single SZS per gene and cell. A cell-gene pair has a large negative (resp. positive) SZS, if, on average, introns have statistically significantly smaller (resp. longer) intron length compared to the population of cells profiled in an experiment.

Given metadata, the SZS detects differential splicing as a function of cell class, such as annotated cell type and/or compartment. For each cell type or compartment, the median SZS in each cell type is reported with a statistical effect size and p-value corrected for multiple hypothesis testing[16] (Methods, Supp. Fig 1). The p-value measures the probability of observing this value of the median given the number of sampled cells. When two experimental replicates are available, the SZS pipeline also provides an FDR. The FDR is based on calculating, for each gene, the fraction of cell type pairs that are significant in the same direction in both experiments versus all called

pairs (Methods) and conservative compared to randomization methods[15] (Supp. Fig 2, Supplement).

Here we applied the SZS to the output of SICILIAN[17], a statistical post-filter on STAR[18] that removes false positive junction calls and enables discovery of un-annotated splicing contributing to cell type-specific regulation.

First, we analyzed 28,792 and 24,676 carefully annotated human lung cells from the Human Lung Cell Atlas (HLCA) study[19] across two individuals. The HLCA, sequenced on the 10x platform, includes 55 annotated cell types with highly variable depth of sampling per cell type. A median of 1,979 genes detected per cell and 67 genes with genes with a computable SZS per cell. 1,743 genes have a computable SZS in at least 10 cells (Methods).

101 (resp. 142) pairs of genes and cell types are called as having significant cell-type specific splicing including 51 (resp. 64) genes and 28 (resp. 43) cell types in the first (resp. second) individuals. The most cell type-specifically regulated splicing event was identified in MYL6, a gene known to be regulated in smooth muscle[20], as measured by the number of cell types with significant splicing deviation of the SZS with 18 cell types in both individuals, while all other genes are called in at most 7 cell types (Fig. 2A). Other highly cell type-specifically spliced genes are ANXA1, a phospholipid binding protein, SOD2, an antioxidant enzyme, LMNA, a gene which produces lamin A, lamin C, and progerin and whose dysregulation causes a variety of developmental disorders[21], and ATP5F1C, a core component of the mitochondrial ATP-synthase machinery (Supp. Fig 3-9 and Supp. Table 1-2).

Given the high rate of SZS-called cell type-specific splicing, used ANOVA to test whether splicing alone could define a cellular compartments, broad categories of cell types that define tissues: immune, endothelial, epithelial and stromal. For each gene and cell type with a computable SZS in >20 cells, we used ANOVA to test if the cellular compartment was a significant predictor of median SZS (Methods). MYL6, RPS24 and CD47 had significant compartment effects in both individuals (effect size > 0.3 and p-value < 0.05) (Fig. 2, Supp. Table 3-4).

MYL6, RPS24 and CD47 are ubiquitously expressed, essential genes with partially or poorly characterized splicing programs at cell-type or compartment level in any organ. MYL6 is an essential component of the actin cytoskeleton with partially characterized splicing[20,22–24]. RPS24 is a core component of the ribosome and one of the most highly expressed genes in the cell[25,26]. CD47 is a recently identified immune-regulatory, membrane protein that has recently been identified as a therapeutic target in a set of myeloid malignancies[27] but has undescribed cell type-specific splicing programs.

We then tested the predictive power of the set of three most ANOVA-significant genes for classification of compartments in each lung data set using multinomial logistic regression (Methods). We restricted analysis to cells that have computable SZS for all three genes to control for potential confounding with gene expression. Using equal splits of test and training, compartment was predicted with > 98% accuracy in each class in each individual, driven by high discriminatory power of the model between abundant immune and epithelial cells; and secondarily, endothelial cells; very few stromal cells had detectable levels of all three genes. ~92% classification accuracy was achieved with MYL6 and RPS24 alone.

To provide biological interpretability, the SZS can be decomposed to automatically identify donors and acceptors with most intron-length variability driving the significance of the SZS (Methods). Formally, the SZS is a sum of terms that measure per-cell statistical difference in intron lengths for each donor or acceptor. We call these terms “z-contrib”s. The highest z-contribs in absolute value are ones likely to most contribute to detected differential alternative splicing and we call “driving”.

Driving z-contribs in MYL6 are in donor 56160320, and acceptor 56161387, splice sites correspond to the most well characterized exon skipping events in MYL6 (Figure 2A) that alters the terminal helix in the final EF-hand domain, known to be preferentially included in smooth muscle and excluded in lymphocytes[20,23,24]. Thus, in MYL6, the SZS automatically re-identifies both these exon skipping events and the cell types in which they are regulated. Similarly, the highest z-contribs for RPS24 are for donors 78037304 and 7802854 and acceptors 78040204 and 7802270, respectively, all of which are known as annotated alternative exons (Supp. Table 5).

Overall, ~39% (171 out of 437) of driving z-contribs for significantly differentially spliced genes (FDR<.05) in the HLCA corpus are annotated alternative exons (Methods). The gene with second-highest magnitude of median SZS is PPP1R12A a protein phosphatase regulatory subunit: its driving z-contrib donor is not annotated as being alternatively spliced. The SZS and Z-contrib analysis identified it as a novel un-annotated exon skipping event alternatively regulated between airway smooth muscle cells and macrophages (Fig. 2E). This supports the use of the SZS to re-identify known alternative splicing events and as a purely statistical, annotation-free, approach to discover new alternative splicing events that are cell type-specifically regulated.

To further test the power of the SZS, we used it to detect splicing changes during human spermatogenesis, a highly regulated, temporary dynamic process characterized by three major stages of cell identity[28]. While splicing in sperm is known to be complex, regulated splicing programs during developmental are not well-understood[29]. We correlated pseudotime assignments with the SZS for 7,134 human sperm cells in a recent study of the spermatogenic lineage[28]. 200 genes out of the 1,092 genes (with a computable SZS) expressed in > 300 cells were significantly correlated with the pseudotime defined as abs(spearman) >0.1 and Bonferroni corrected p-value < 0.05) (Supp. Table 6).

The five genes with highest magnitude correlation between pseudotime and the SZS are TPPP2, a gene regulating tubulin polymerization implicated in male infertility [30], FAM71E1, identified in mouse as testes-predominantly expressed[31], SPATA42 a long non-coding RNA implicated in azoospermia[32], MTFR1, a gene regulating mitochondrial fission and MLF1, an oncogene known to be regulated in drosophila testes[33] (Fig. 3A). Z-contrib analysis revealed alternative splicing in MTFR1 driven by one donor which splices predominantly to an annotated acceptor in mature sperm, while splicing predominantly to an unannotated acceptor in immature sperm. This acceptor is at the boundary of the PDEase protein domain in PDE7A1, a cAMP-specific Phosphodiesterase immediately downstream of MTFR1, supporting a set of un-annotated transcriptional readthrough fusion transcripts (Fig. 3C). Z-contrib analysis also identified highly expressed un-annotated splice isoforms of TPPP2 including ones that modify the functionally annotated p25-alpha domain thought to be brain-specifically expressed[34] (Supp. Fig. 10-15).

Together these data support the use of the SZS to rapidly identify regulated splicing events during cellular differentiation and importance of including statistically supported un-annotated splicing events in the SZS computation[17].

Undoubtedly, the 3' bias of 10x genomics data derived from priming on the poly-A tail restricts discovery of regulated splicing in this study by biasing data and measurement towards only a subset of splice sites. Moreover, because the SZS collapses all splice events to a single scalar, the SZS lacks power to detect some differential splicing events. We predict that additional discoveries will be enabled by applying this method to droplet-based 5' sequencing and or full length transcript profiling. The methodology in the SZS pipeline can be applied to analyze Smart-Seq2 and any RNA-seq dataset, including bulk sequencing.

Because the SZS has a tractable statistical distribution and is single-cell resolved, it enables new biological analysis beyond the scope of this paper, including clustering approaches on the basis of the SZS alone or in combination with gene expression. In addition, the SZS can be correlated with any other numerical phenotype, such as pseudotime, cell cycle trajectory, gene expression or spatial position in the context of single cell spatial genomics. In summary, by deconvolving technical and biological noise in splicing, the SZS provides a new method for the field to identify and prioritize splicing events that are regulated and functional and to address decades-long debates regarding the degree to which alternative splicing is regulated at a single cell level.

Acknowledgments
We thank Peter Wang and Steve Quake for insight and instrumental comments during the development of the method, and Elisabeth Meyer and Robert Bierman for comments on the manuscript. We thank Brian Hermann and Keren Cheng for generously providing pseudotime data, Kyle Travaglini and Mark Krasnow for providing access to the Human Lung Cell Atlas data prior to publication, and Angela Wu for insights. J.O. is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1656518 and a Stanford Graduate Fellowship. R.D. is supported by the Cancer Systems Biology Scholars Program Grant R25 CA180993 and Clinical Data Science Fellowship Grant T15 LM7033-36. J.S. is supported by the National Institute of General Medical Sciences Grant R01 GM116847 and NSF Faculty Early Career Development Program Award MCB1552196.

Code Availability
The SZS codes along with the codes used for data analysis and to create the figures are available through a GitHub repository: https://github.com/juliaolivieri/SZS_pipeline/.

Data Availability
Human lung cell atlas (HLCA) data was downloaded from the EGA archive at accession number EGAS00001004344. We refer to Patient 2 in HLCA as Individual 1 and Patient 3 as Individual 2 in this manuscript. Human spermatogenesis data was downloaded from the NIH Gene Expression Omnibus (GEO) databases under accession number GSE109037.

List of Figures:

Figure 1. (A) Simple SZS toy example: Because the cell on the left contains "average" splicing for the population, its SZS = 0. Because the cell on the right has longer introns on average for the gene, its SZS is positive. (B) Complex SZS toy example:The cell on the left has shorter introns on average than the cell on the right, giving it a lower SZS. (C) The SZS can be used to correlate splicing changes in any gene with any scalar variable (eg.pseudotime). (D) Given cell type annotation of the dataset, SZS scores can be aggregated for each cell type, allowing identification of cell types with statistically deviant splicing. (E) The SZS pipeline and possible downstream analyses.

Figure 2. (A) The ribosomal subunit RPS24 has three splice junctions that uniquely identify each RPS24 isoform. The fraction of each identifying isoforms for each cell type in both HLCA individuals are plotted, with error bars showing the empirical 25th and 75th quantiles of these fractions per cell. Cell types are colored by cellular compartment (B) MYL6 donor with driving z-contrib: the dot plot shows the fraction of reads from each donor splicing to each acceptor for each HLCA individual, colored by cellular compartment. The box plot shows the average rank of the acceptor paired to this donor by cell type for the first HLCA individual. Both smooth muscle ontologies in the data are starred and exhibit high levels of exon inclusion. (C) CD47 acceptor with driving z-contrib: Plots are as described in 2.B. (D) Heatmap of median SZS per cell type for top compartment-specific genes MYL6, CD47, and RPS24 in both HLCA individuals. (E) PPP1R12A donor with driving z-contrib: Plots are as described in 2.B.

Figure 3. (A) 2D heatmap showing the relationship between spermatogenesis pseudotime and SZS for the gene MTFR1 in human sperm data. (B) Dot plot and box plot for MTFR1 as described in 2.B, with groups defined by binning pseudotime values. (C) Dot plot and box plot for MTFR1 as described in 2.B, with groups defined by developmental state. One transcriptional readthrough between MTFR1 and PDE7A is annotated in refSeq; the SZS and SICILIAN identify a large repertoire of transcriptional readthrough splice variants.

Supplementary Tables:

Table 1. Outlying median differential splicing with FDR for HLCA P2 and P3. Table containing a row for each gene and cell type with at least 20 cells in the dataset. Columns: gene, cell type, effect size, p-value, FDR estimate based on comparing HLCA P2 and HLCA P3.

Table 2. ANOVA compartment differential splicing with FDR for HLCA P2 and P3. Table containing a row for each gene present in at least 2 compartments, 2 tissues, and 5 cell

types in the dataset. Columns: gene, cell type, effect size, p-value, FDR estimate based on comparing HLCA P2 and HLCA P3

Table 3. SZS pipeline values per cell for HLCA P2. Table with one row per gene per cell with columns indicating the cell, tissue, compartment, free annotation, z score, and z contrib

Table 4. SZS pipeline values per cell for HLCA P3. Table with one row per gene per cell with columns indicating the cell, tissue, compartment, free annotation, z score, and z contrib

Table 5. The highest z-contrib scores for the donors and acceptors for each pair of gene and cell type.

Table 6. Pseudotime correlation values for human spermatogenesis dataset[28]. Table with one row per gene with splicing expression in at least 100 cells from the time course. Columns indicate gene, pearson correlation, spearman correlation, number of cells the gene is expressed in, and the p-value for significance of the correlation based on spearman

Bibliography:

1. Shao, Y. *et al.* Alternative splicing-derived intersectin1-L and intersectin1-S exert opposite function in glioma progression. *Cell Death Dis.* **10**, 431 (2019).
2. Nakka, K., Kovac, R., Wong, M. M.-K. & Dilworth, F. J. Intron Retained, Transcript Detained: Intron Retention as a Hallmark of the Quiescent Satellite Cell State. *Dev. Cell* **53**, 623–625 (2020).
3. Oleynikov, Y. & Singer, R. H. RNA localization: different zipcodes, same postman? *Trends Cell Biol.* **8**, 381–383 (1998).
4. Yang, Y. & Carstens, R. P. Alternative splicing regulates distinct subcellular localization of Epithelial splicing regulatory protein 1 (Esrp1) isoforms. *Sci. Rep.* **7**, 3848 (2017).
5. Arzalluz-Luque, Á. & Conesa, A. Single-cell RNAseq for the study of isoforms—how is that possible? *Genome Biology* vol. 19 (2018).
6. Vaquero-Garcia, J. *et al.* A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5**, e11752 (2016).
7. Trincado, J. L. *et al.* SUPPA2 provides fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. doi:10.1101/086876.
8. Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E5593–601 (2014).
9. Salzman, J., Jiang, H. & Wong, W. H. Statistical Modeling of RNA-Seq Data. *Statistical Science* vol. 26 62–83 (2011).
10. Li, J. J., Jiang, C.-R., Brown, J. B., Huang, H. & Bickel, P. J. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 19867–19872 (2011).

11. Westoby, J., Artemov, P., Hemberg, M. & Ferguson-Smith, A. Obstacles to detecting isoforms using full-length scRNA-seq data. *Genome Biol.* **21**, 74 (2020).
12. Buen Abad Najar, C. F., Yosef, N. & Lareau, L. F. Coverage-dependent bias creates the appearance of binary splicing in single cells. *Elife* **9**, (2020).
13. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
14. Katsevich, E. & Roeder, K. Conditional resampling improves calibration in single cell CRISPR screen analysis. doi:10.1101/2020.08.13.250092.
15. Barber, R. F. & Candès, E. J. Controlling the false discovery rate via knockoffs. *The Annals of Statistics* vol. 43 2055–2085 (2015).
16. Maritz, J. S. & Jarrett, R. G. A Note on Estimating the Variance of the Sample Median. *Journal of the American Statistical Association* vol. 73 194–196 (1978).
17. Dehghannasiri, R., Olivieri, J. E. & Salzman, J. Specific splice junction detection in single cells with SICILIAN. doi:10.1101/2020.04.14.041905.
18. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
19. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single cell RNA sequencing. doi:10.1101/742320.
20. Fisher, S. A. Vascular smooth muscle phenotypic diversity and function. *Physiol. Genomics* **42A**, 169–187 (2010).
21. Schreiber, K. H. & Kennedy, B. K. When lamins go bad: nuclear structure and disease. *Cell* **152**, 1365–1375 (2013).
22. Brozovich, F. V. *et al.* Mechanisms of Vascular Smooth Muscle Contraction and the Basis for Pharmacologic Treatment of Smooth Muscle Disorders. *Pharmacol. Rev.* **68**, 476–532 (2016).
23. Morano, I. Chapter 11 Essential Myosin Light Chains Regulate Myosin Function and Muscle Contraction. *Muscle Contraction and Cell Motility* 273–302 (2016) doi:10.1201/9781315364674-12.
24. Lenz, S., Lohse, P., Seidel, U. & Arnold, H. H. The alkali light chains of human smooth and nonmuscle myosins are encoded by a single gene. Tissue-specific expression by alternative splicing pathways. *J. Biol. Chem.* **264**, 9009–9015 (1989).
25. Bortoluzzi, S., d'Alessi, F., Romualdi, C. & Danieli, G. A. Differential expression of genes coding for ribosomal proteins in different human tissues. *Bioinformatics* **17**, 1152–1157 (2001).
26. Zhou, X., Liao, W.-J., Liao, J.-M., Liao, P. & Lu, H. Ribosomal proteins: functions beyond the ribosome. *J. Mol. Cell Biol.* **7**, 92–104 (2015).
27. Chao, M. P. *et al.* Therapeutic Targeting of the Macrophage Immune Checkpoint CD47 in Myeloid Malignancies. *Front. Oncol.* **9**, 1380 (2019).
28. Hermann, B. P. *et al.* The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids. *Cell Rep.* **25**, 1650–1667.e8 (2018).
29. Hannigan, M. M., Zagore, L. L. & Licatalosi, D. D. Ptbp2 Controls an Alternative Splicing Network Required for Cell Communication during Spermatogenesis. *Cell Rep.* **19**, 2598–2612 (2017).
30. Zhu, F. *et al.* Deficiency of TPPP2, a factor linked to oligoasthenozoospermia, causes subfertility in male mice. *J. Cell. Mol. Med.* **23**, 2583–2594 (2019).

31. Kwon, J. T. *et al.* Expression of uncharacterized male germ cell-specific genes and discovery of novel sperm-tail proteins in mice. *PLoS One* **12**, e0182038 (2017).
32. Bo, H. *et al.* Long noncoding RNAs expression profile and long noncoding RNA-mediated competing endogenous RNA network in nonobstructive azoospermia patients. *Epigenomics* **12**, 673–684 (2020).
33. Singh, S. R., Liu, Y., Zhao, J., Zeng, X. & Hou, S. X. The novel tumour suppressor Madm regulates stem cell competition in the Drosophila testis. *Nat. Commun.* **7**, 10473 (2016).
34. Takahashi, M. *et al.* A novel brain-specific 25 kDa protein (p25) is phosphorylated by a Ser/Thr-Pro kinase (TPK II) from tau protein kinase fractions. *FEBS Lett.* **289**, 37–43 (1991).

Methods

A statistical score for identifying regulated splicing in single cells

The Splicing Z Score (SZS) is a general approach for statistically quantifying alternative splicing in scRNA-Seq. The SZS score gathers the information about the gene's splicing across all sequenced cells for each pair of cell and gene.

The SZS score computation for a gene consists of two parts: one relative to the donors in the gene (the donor SZS) and one relative to the acceptors (the acceptor SZS). We first explain how to calculate the donor SZS for one gene (suppressing the notation specifying the gene for simplicity) and the acceptor SZS can be computed similarly. Let i specify the donor, j specify the acceptor, k specify the cell, and ℓ specify the read. Therefore each junctional read for the gene in the dataset is specified by a unique combination of $ijk\ell$. Note that we only consider junctions for which the donor has multiple acceptors in the dataset.

The donor SZS score calculation proceeds by treating each donor separately. Assume that donor i has multiple acceptors across the whole dataset (otherwise we would have filtered it out). We rank these acceptors from closest to the farthest to donor i (i.e., from lowest to the highest genomic coordinate). For example, if there were four acceptors found for i across the whole dataset, we would rank them 1, 2, 3, and 4 in the order of their genomic coordinates. For genes that are on the minus strand, the acceptors are ranked differently as described in the Supplement. We use $r_{ijk\ell}$ to denote the acceptor rank for the read specified by $ijk\ell$. If $r_{ijk\ell} = 1$, it means the acceptor has the smallest genomic coordinate among all acceptors in the dataset.

Now, let N_i be the number of junctional reads observed across all cells for donor i. We can compute $\overline{r_i}$, the average rank of the acceptors for donor i, as

$$\overline{r_i} = \frac{\sum_{j,k,\ell} r_{ijk\ell}}{N_i}.$$

We can also find the variance of the ranks as:

$$\hat{\sigma}^2_{\overline{r}_i} = \frac{\sum_{j,k,\ell} \left(r_{ijk\ell} - \overline{r_i}\right)^2}{N_i}.$$

Now, we renormalize the rank $r_{ijk\ell}$ using the sample variance and mean as:

$$S_{ijk\ell} = \frac{r_{ijk\ell} - \overline{r_i}}{\hat{\sigma}_{\overline{r_i}}}.$$

We can see that $\mathbb{E}\left[S_{ijk\ell}\right] = 0$ and $\mathrm{var}\left(S_{ijk\ell}\right) = 1$ as we subtract the mean and divide by the standard deviation. The closer the acceptor of a read is to the donor, the smaller its corresponding $S_{ijk\ell}$ value would be . In practice, we take the 0.01 and 0.99 quantiles of $S_{ijk\ell}$, and replace further outliers with the 0.1 and 0.9 quantiles to avoid effects from extreme outliers.

We now aggregate these zero-mean and unit-variance variables for all donors in the gene to compute the donor SZS (z^d_k) in the kth cell as:

$$z^d_k = \frac{\sum_{i,j,\ell} S_{ijk\ell}}{\sqrt{M_k}},$$

where M_k is the number of junctional reads mapping to the gene in the kth cell. It is straightforward to see that $\mathbb{E}\left[z_k\right] = 0$. Knowing that the variance of the sum of independent random variables is the sum of their variances:

$$\mathrm{var}\left(z^d_k\right) \approx \frac{\sum_{ij\ell} \mathrm{var}\left(S_{ijk\ell}\right)}{M_k} = 1$$

The approximation is due to the fact that the $S_{ijk\ell}$'s are not necessarily independent but we expect them to be close enough to independent.

Similarly, we can compute the acceptor SZS z_k^a. We average the two scores z_k^a and z_k^d and further normalize for the read depth to compute the SZS z_k for the gene in cell k (Supplement):

$$z_k = \frac{1}{M_k}(z_k^d - z_k^a)/\sqrt{2}.$$

For computing SZS, we only consider cells with at least 6 junctional reads mapping to the gene, and the junctional reads for which the donor has more than one acceptor observed in the dataset.

Differential splicing analysis between ontologies using outlying median SZS method

We define an ontology as a unique combination of tissue, compartment, and cell type. Let n_p be the number of cells in ontology p that express at least on splice junction of the gene. For each given gene, We restrict ourselves to ontologies with $n_p > 20$. We calculate the median SZS per ontology for the gene:

$$m_p = \text{median}_{k \in p}\left(z_k\right).$$

Because the SZS score for each gene is a zero-mean variable, if the absolute value of median SZS m_p for an ontology is much larger than zero, it indicates that the ontology has a different splicing pattern than the average for that gene. We compute the p value for each median SZS by comparing it against a conservative and highly precise approximation of its null distribution.

Now we have a distribution of median SZS s (one for each ontology): how can we make inferences about which are "extreme" and represent differential splicing? Based on the null hypothesis (that the given gene does not exhibit differential splicing by ontology), we still do not expect all of the medians to be exactly the same. Rather, we expect them to vary based on the known distribution of medians from a set of random variables.

To formalize this difference, we calculate the median and standard deviation

$$\hat{\mu} = \text{median}\left(z_k\right)$$

$$\hat{\sigma} = 1$$

for all cells and all genes. We expect $\hat{\mu}$ to be close to 0. Because of the SZS scaling we expect z_k to be less than 1. Because the size of the variance we use determines the range of differential alternative splicing events we can detect as significant, with lower variance corresponding to more significance detected, we are being more conservative here by using a high estimate of variance. Then for a given ontology, we can calculate the standard deviation of the median SZS under the null hypothesis that there is no differential splicing (meaning that each ontology has the same distribution of SZS s).

$$\hat{\sigma}_p^{\text{med}} \approx \frac{1}{\sqrt{4n_p f^2(\hat{\mu})}} \approx \sqrt{\frac{\pi\hat{\sigma}^2}{2n_p}} = \sqrt{\frac{\pi}{2n_p}}.$$

The first approximation is based on the approximation given in Martiz *et al.*[16]. Here $f(\hat{\mu})$ is the density function of our sample median. We assume that our SZS 's are approximately normal with mean 0 and variance 1. Then because $\hat{\mu} \approx 0$, we can see that in our case $f^2(\hat{\mu}) \approx \frac{1}{2\pi\hat{\sigma}^2}$.

Now given an ontology p, we know the expected mean of the median, $\hat{\mu}$, and the expected standard deviation $\hat{\sigma}_p^{\text{med}}$ under the null hypothesis. Assuming the medians m_p of the SZSs follow a normal distribution, we

can use $\hat{\sigma}_p^{\text{med}}$ to calculate a p-value p_p for each m_p as the probability that we would get a value at least as extreme as m_p given the null hypothesis. A small p_p implies that the median SZS m_p for ontology p is significantly different from the overall distribution of the median SZSs for the gene across all ontologies. We use the Benjamini-Hochberg procedure to correct the p_p values for multiple testing. To make inferences of differential splicing, we consider both the corrected p_p and the magnitude of m_p, which acts as an effect size for the splicing change.

To identify differentially spliced genes in each cell type, we chose cut-offs of a p-value < 0.05 and an effect size of deviant median SZS score greater than 0.3. This effect size cutoff was chosen based on the effect size of TPM1 in lung airway smooth muscle, where it is known to be differentially spliced[31].

Detecting differential splicing between compartments using the ANOVA analysis of the SZS

Not only do we want to identify ontology-specific splicing, we are also interested in identifying compartment-specific splicing. Because there are only four major compartments (immune, stromal, endothelial, and epithelial) in the HLCA data, we do not have enough measurements to use the median SZS approach outlined above. Instead, we test how well compartments can predict the median SZS m_p by first fitting the following linear model:

$$\mathbf{Y} = \beta_0 \mathbb{I}_{\text{endothelial}} + \beta_1 \mathbb{I}_{\text{epithelial}} + \beta_2 \mathbb{I}_{\text{immune}} + \beta_3 \mathbb{I}_{\text{stromal}},$$

where $\mathbf{Y}$ is the vector of median SZSs and $\mathbb{I}_X$ is the indicator function being 1 when the compartment of the cell is X. This model has one observation per ontology. After fitting this model, we used ANOVA to determine the significance of each variable (compartment) in our model. For each gene, we consider both the p-value obtained by the ANOVA analysis (after correction by the Benjamini-Hochberg procedure) and the magnitude of the coefficient for each compartment in the linear model, which serves as an effect size, to identify differential splicing at the compartment level. We use the same effect size and p-value cutoffs to define significance here as we did for the outlying median SZS method (0.3 and 0.05, respectively).

FDR calculation

When dealing with more than one dataset each being either an experimental or biological replicate, we would be able to compute an FDR for each pair of gene and cell type in each one of the data sets. The FDR is a function of the thresholds we use for both the effect size and p-value for identifying differential splicing in the replicates. We calculate the concordance c between the replicates as the fraction of genes and cell types called as differentially spliced with the same effect direction in the datasets. We then estimate the FDR given these thresholds as $2(1–c)$ (Supplement).

Multinomial regression analysis

We used the R package nnet to compute classification performance using models that included all interaction terms between the SZS for RPS24, CD47, and MYL6 the most significant genes at predicting compartment assignment by the ANOVA analysis.

Training and test errors were quantified for each dataset using two-fold cross-validation. We only included cells with detected expression of each of these three genes, which results in very few stromal cells being included in the analysis.

Finding the splice sites driving the alternative splicing of a gene using the Z-contrib score

We define the z-contrib score to find the splice site that drives the alternative splicing pattern of a differentially-spliced gene across different ontologies. The z-contrib score for each donor and acceptor in a gene is obtained by decomposing the donor SZS z_k^d and the acceptor SZS z_k^a, respectively. The z-contrib score for each donor can be computed by rearranging the summands in the sum for z_k^d and using the associative property of addition:

$$\begin{aligned} z_k^d = \frac{\sum_{i,j,\ell} S_{ijk\ell}}{\sqrt{M_k}} &= \sum_i \Big(\frac{\sum_{jl} S_{ijkl}}{\sqrt{M_k}} \Big) \\ &= \underbrace{\frac{\sum_{jl} S_{1jkl}}{\sqrt{M_k}}}_{\text{donor 1}} + \underbrace{\frac{\sum_{jl} S_{2jkl}}{\sqrt{M_k}}}_{\text{donor 2}} + \ldots + \underbrace{\frac{\sum_{jl} S_{\mathrm{D}jkl}}{\sqrt{M_k}}}_{\text{donor D}} \end{aligned}$$

This suggests that the acceptor SZS z_k^d can be viewed as the sum of donor-specific scores, being called the z-contrib scores, where each z-contrib score determines the fraction of z_k^d that is being driven by a specific donor. By doing so, we compute the z-contrib score for the i-th donor as:

$$z_{\text{cont}}^d = \frac{\sum_{jl} S_{ijkl}}{\sqrt{M_k}}$$

Similarly, we compute the z_{cont}^a for the j-th acceptor as:

$$z_{\text{cont}}^a = \frac{\sum_{il} S_{ijkl}}{\sqrt{M_k}}$$

When genes with significant differential splicing patterns are found using the effect size and p-value (in case of a single data set) or the FDR (when dealing with multiple data set), the z-contrib scores can be used to find the donors and acceptors that have the largest contributions to the deviant splicing pattern of the gene. The z-contrib score, equips us with a systematic approach to relate the deviant splicing pattern of a gene to its constituents exons (both donors and acceptors), and paves the way for further functional and evolutionary investigations on the exons that mainly drive the alternative splicing pattern of the gene.

For each splice site, in addition to the z-contrib score, we report on the annotation status of the driving splice sites, including whether they are annotated as alternatively spliced and whether their usage impacts protein domains. To determine whether an exon is known to be involved in alternative splicing based on the reference annotation, we extracted the splice sites from the GRCh38 annotation GTF file and then considered those donors (acceptors) that are observed to have at least two distinct acceptors (donors) across the extracted splice sites as known alternatively spliced donors (acceptors). With this definition for annotated alternative splice sites, we found that 40% of the splice sites with driving z-contribs in differentially spliced genes (FDR < 0.05) are not involved in annotated alternative splicing.

Calling splice junctions using SICILIAN

We used SICILIAN[16] for calling splice junctions. SICILIAN is a statistical wrapper that can be applied to the alignment output file from a spliced aligner and can distinguish false positive junction calls from true positives via assigning a statistical score to each splice junction reported by the aligner. The scRNA-Seq data sets were mapped to the reference human genome (GRCh38) using STAR version 2.7.5a with default parameters.

Figure 1

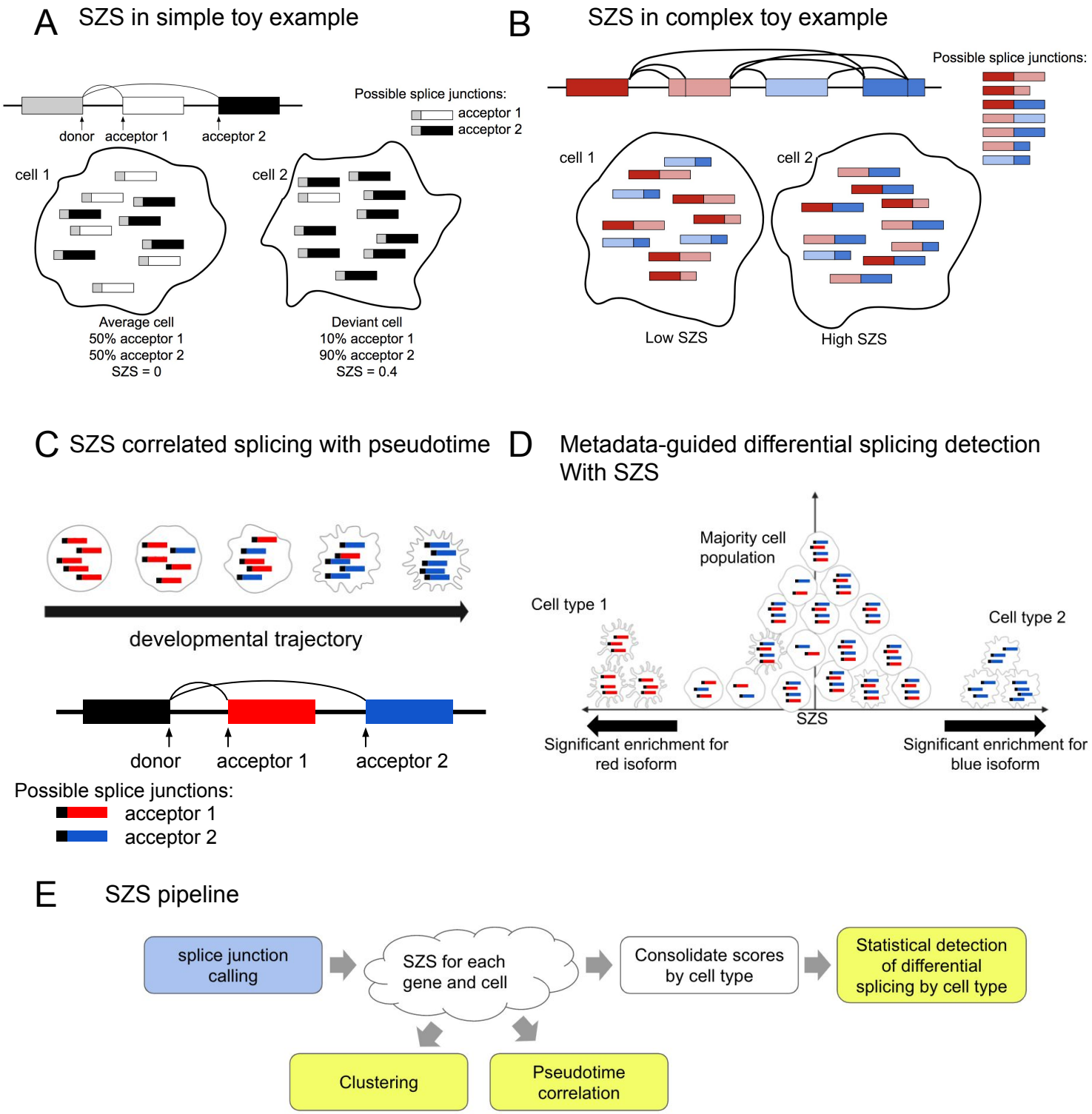


Figure 1. (A) Simple SZS toy example: Because the cell on the left contains “average” splicing for the population, its SZS = 0. Because the cell on the right has longer introns on average for the gene, its SZS is positive. (B) Complex SZS toy example:The cell on the left has shorter introns on average than the cell on the right, giving it a lower SZS. (C) The SZS can be used to correlate splicing changes in any gene with any scalar variable (eg.pseudotime). (D) Given cell type annotation of the dataset, SZS scores can be aggregated for each cell type, allowing identification of cell types with statistically deviant splicing. (E) The SZS pipeline and possible downstream analyses.

Figure 2

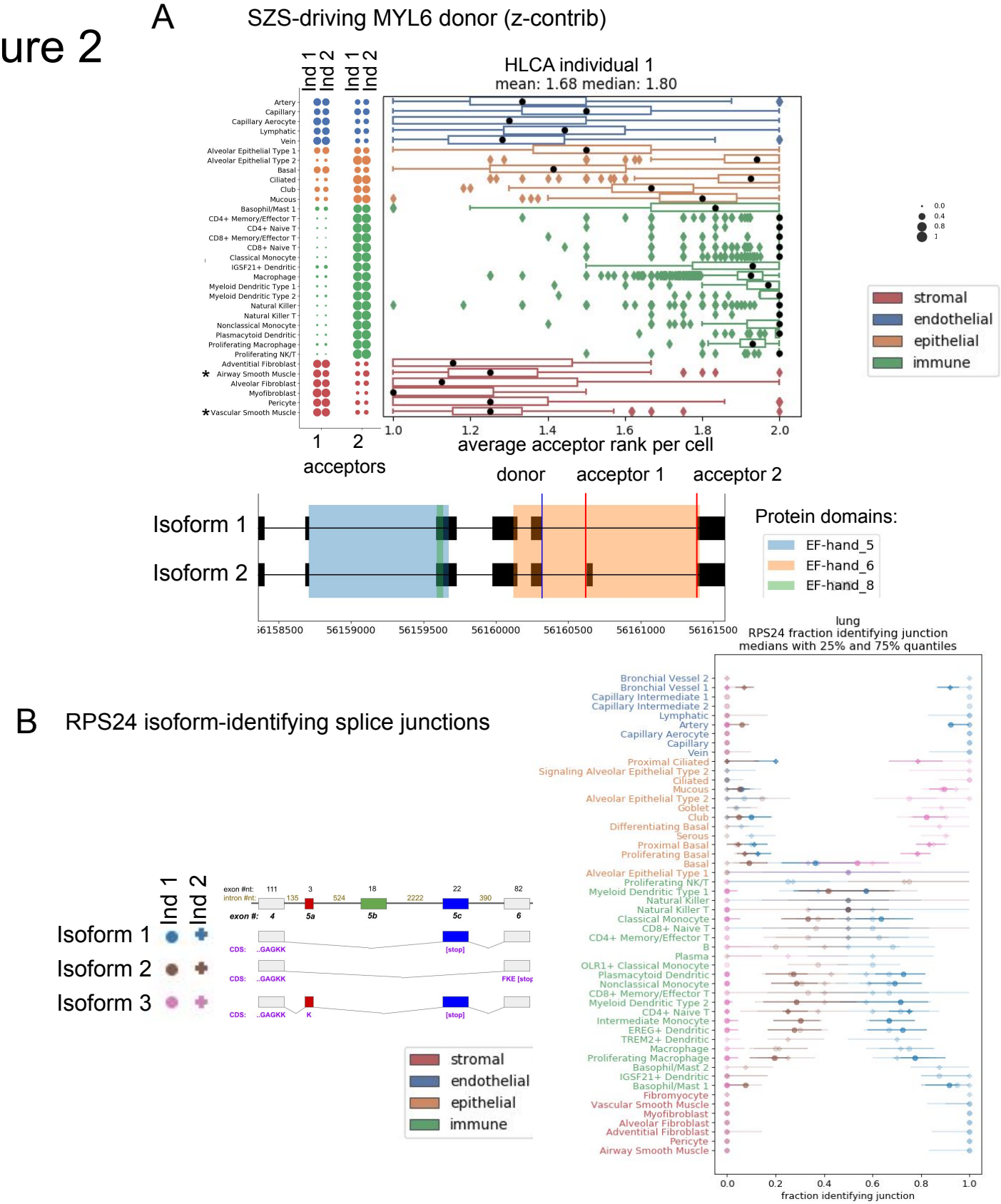


Figure 2. (A) The ribosomal subunit RPS24 has three splice junctions that uniquely identify each RPS24 isoform. The fraction of each identifying isoforms for each cell type in both HLCA individuals are plotted, with error bars showing the empirical 25th and 75th quantiles of these fractions per cell. Cell types are colored by cellular compartment (B) MYL6 donor with driving z-contrib: the dot plot shows the fraction of reads from each donor splicing to each acceptor for each HLCA individual, colored by cellular compartment. The box plot shows the average rank of the acceptor paired to this donor by cell type for the first HLCA individual. Both smooth muscle ontologies in the data are starred and exhibit high levels of exon inclusion.

Figure 2

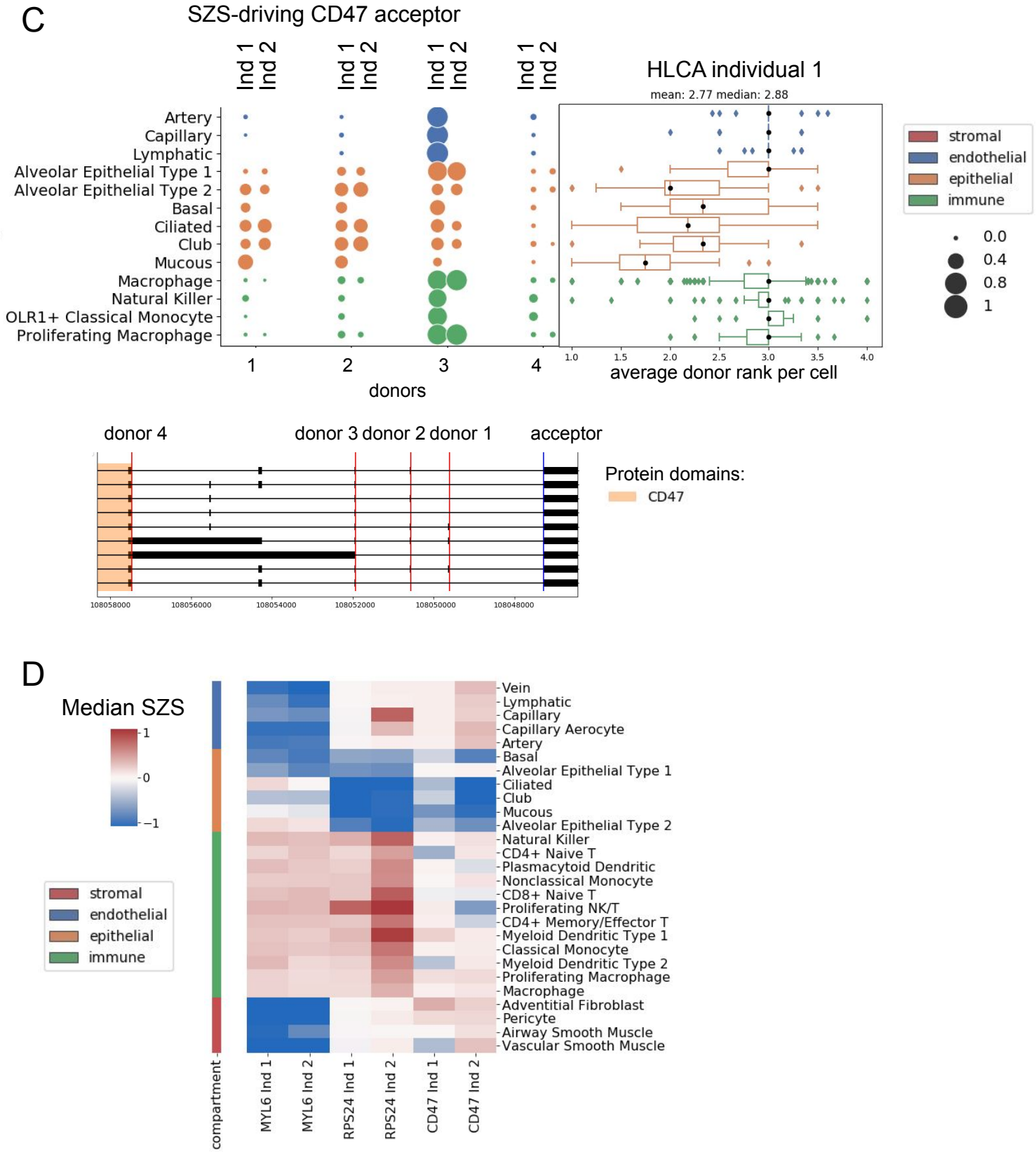


Figure 2. (C) CD47 acceptor with driving z-contrib: Plots are as described in 2.B.(D) Heatmap of median SZS per cell type for top compartment-specific genes MYL6, CD47, and RPS24 in both HLCA individuals.

Figure 2

E

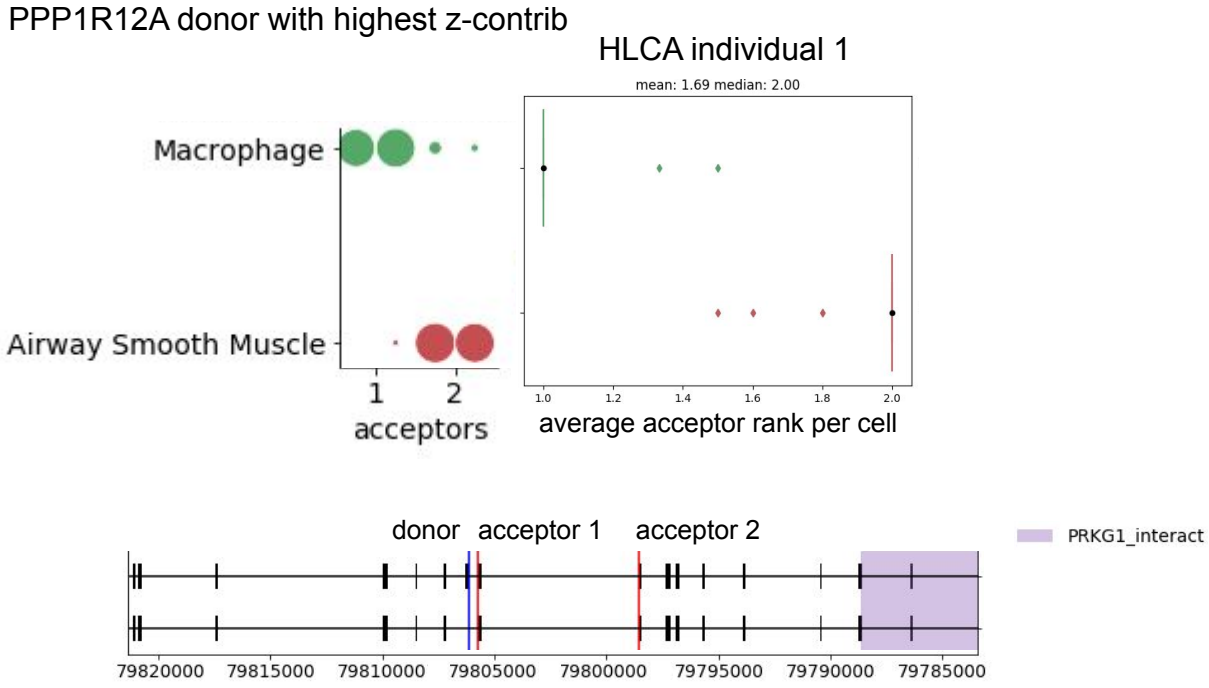
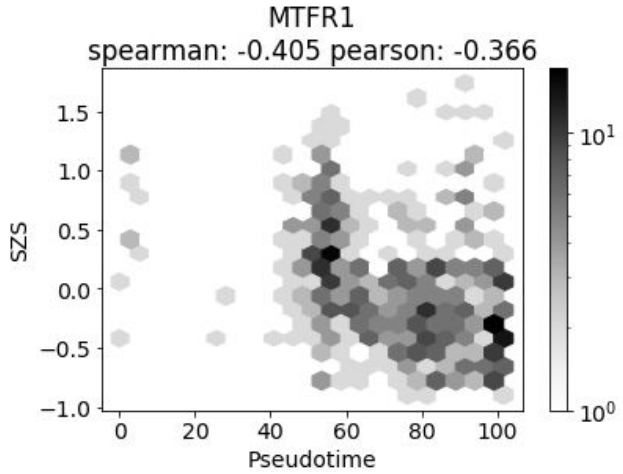


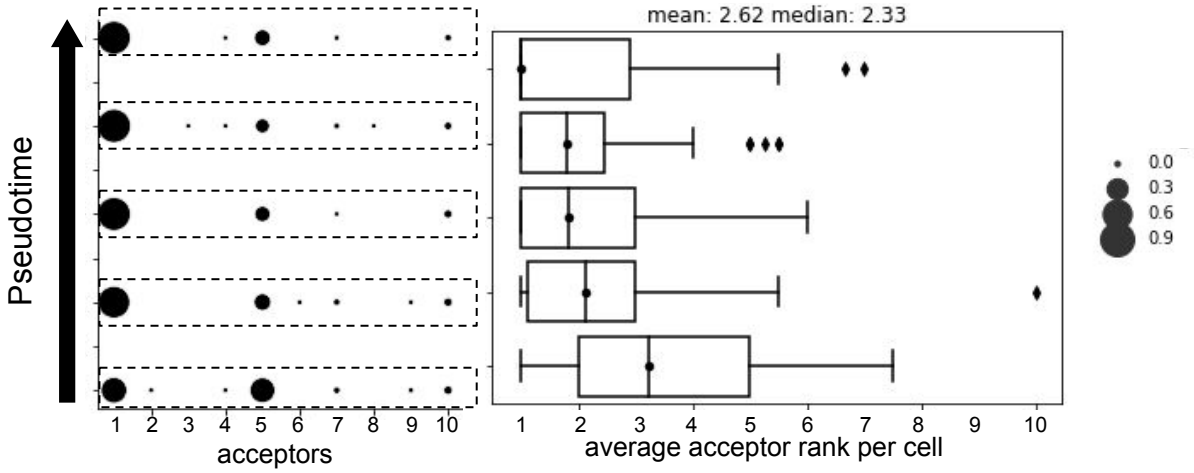
Figure 2. (E) PPP1R12A donor with driving z-contrib: Plots are as described in 2.B.

Figure 3

A Correlation between spermatogenesis pseudotime and SZS for MTFR1



B MTFR1 donor chr8 65708011 over spermatogenesis pseudotime



C Driving z-contrib donor in MTFR1

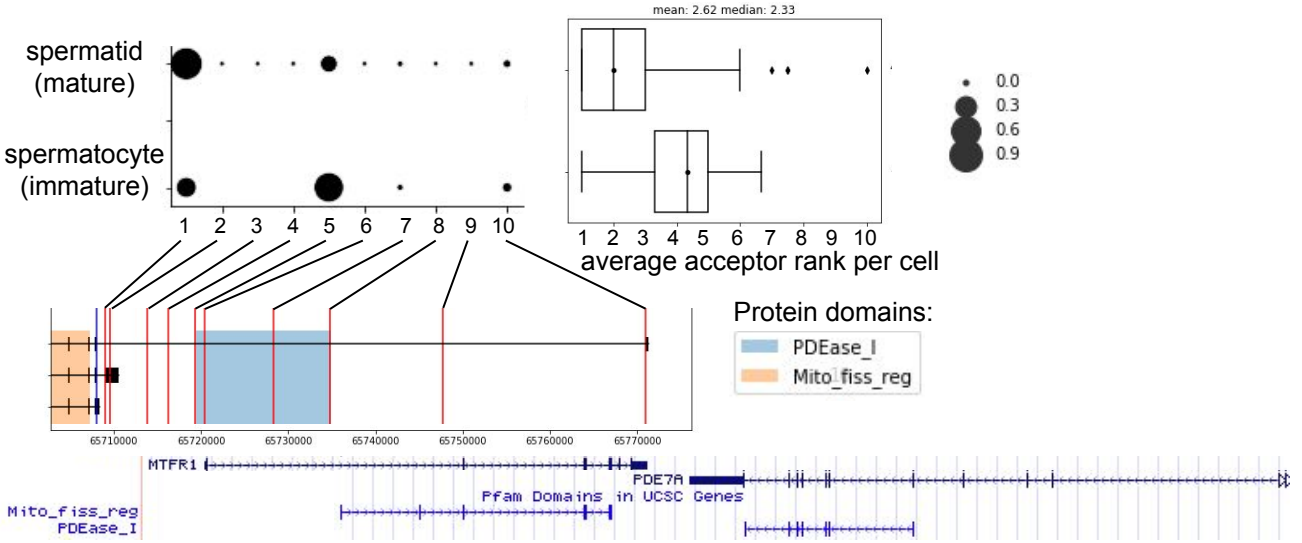


Figure 3. (A) 2D heatmap showing the relationship between spermatogenesis pseudotime and SZS for the gene MTFR1 in human sperm data. (B) Dot plot and box plot for MTFR1 as described in 2.B, with groups defined by binning pseudotime values. (C) Dot plot and box plot for MTFR1 as described in 2.B, with groups defined by developmental state. One transcriptional readthrough between MTFR1 and PDE7A is annotated in refSeq; the SZS and SICILIAN identify a large repertoire of transcriptional readthrough splice variants.