nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-022-01400-x

The SpliZ generalizes 'percent spliced in' to reveal regulated splicing at single-cell resolution

Julia Eve Olivieri[1,2,4], Roozbeh Dehghannasiri[1,3,4] and Julia Salzman[1,3] ✉

Detecting single-cell-regulated splicing from droplet-based technologies is challenging. Here, we introduce the splicing Z score (SpliZ), an annotation-free statistical method to detect regulated splicing in single-cell RNA sequencing. We applied the SpliZ to human lung cells, discovering hundreds of genes with cell-type-specific splicing patterns including ones with potential implications for basic and translational biology.

Splicing is a core function of eukaryotic genomes that generates proteins with diverse and even opposite functions from a single gene[1], changes translation efficiency[2], controls localization[3], and generates noncoding RNAs[4]. So far, 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. However, the enormous complexity of splicing in eukaryotic genomes and the low sequencing depth per cell in scRNA-seq experiments, especially in droplet-based data, makes it extremely challenging to precisely quantify RNA isoform expression and its differential regulation in single cells.

With few exceptions[6], the field has typically attempted to either estimate isoform expression using model-based approaches[7,8] and then perform differential splicing analysis, or directly quantify exon inclusion[9,10] using percent spliced in (PSI). Annotation-based isoform quantification methods give notoriously unstable point estimates in the presence of nonuniform read sampling[7], incomplete annotations or low-depth and 3′-biased sequencing. The second approach, which is to use PSI, quantifies the fraction of transcripts skipping the exon[11]. However, tests based on PSI must proceed exon by exon, requiring thousands of tests, and cannot detect splicing events beyond simple exon skipping. These issues have led to the view that robust differential splicing analysis in droplet-based scRNA-seq is out of reach[5,11,12].

Here, we introduce the splicing Z score (SpliZ), a scalar value assigned to each gene–cell pair that quantifies how deviant a cell's splicing is compared to a population average. The SpliZ can be applied to compiled junctional read counts from any plate- or droplet-based single-cell data, aligned by STAR[13] or SICILIAN[14]. The SpliZ integrates all nonconstitutive spliced reads on a per-gene basis to detect deviant splicing patterns in single cells under low and biased sampling, while remaining uncorrelated with gene expression (Fig. 1a and Extended Data Fig. 1). It also requires one test per gene, greatly reducing the number of tests required by PSI and has the power to detect isoform expression changes beyond exon skipping.

The SpliZ quantifies splicing for each cell–gene pair by (1) assigning a rank to each read aligning to a splice junction based on the relative size of the intron compared to the set of observed introns for that splice site; (2) converting the rank to a residual measuring its statistical deviation compared to 'the typical' intron length rank and (3) statistically grounded scaling and summing of these values to a single SpliZ value per gene and cell (Fig. 1b). A cell–gene pair has a large negative (respectively positive) SpliZ, if, on average, introns have significantly smaller (respectively longer) intron length compared to the population of cells profiled in the experiment (Extended Data Fig. 2). Note that this definition of the SpliZ allows its values to change not only based on exon skipping, but also other events such as intronic alternative polyadenylation that could shift read coverage within the gene. The SpliZ is a mathematical generalization of PSI while increasing power in cases of more complicated exon skipping (Methods). Also, the theoretical properties of the SpliZ allow it to be efficiently integrated into significance testing: under the null hypothesis, the SpliZ has median 0 for every cell type, while under the alternative hypothesis the median SpliZ value per cell type is not necessarily zero (Fig. 1c and Supplementary Information).

As predicted by theory, the SpliZ has higher power than PSI in simulation when there are multiple isoforms for a gene, each of which includes different exons (Extended Data Fig. 3 and Fig. 1d). The SpliZ builds strength across these isoforms to identify real differences between cell types at lower read depths than PSI, while maintaining the same type II error rate (Fig. 1d). The Singular Value Decomposition (SVD) of the residual matrix is also used for biological interpretability. Splice sites corresponding to up to the three largest magnitude components of the first eigenvector are nominated as the statistically most variable splice sites or SpliZsites (Fig. 1e and Methods).

To increase the power of the SpliZ to detect alternative isoform expression in cases where the SpliZ has low power, we also developed the SpliZVD. The SpliZVD modifies the SpliZ by computing by projecting splicing residuals onto eigenvectors of its SVD (we consider only the first projection in this work) (Fig. 1f and Extended Data Fig. 4 and Methods). The SpliZVD maintains the power of PSI in simulation with two cassette exons (Fig. 1f), a situation where the SpliZ has no power to detect differences because both isoforms contain one 'long' and one 'short' intron. In these simulations, the splice sites involved in differential alternative splicing were correctly identified as SpliZsites (Extended Data Fig. 5).

We applied the SpliZ(VD) to 60,550 carefully annotated human lung cells across two individuals from the Human Lung Cell Atlas (HLCA)[15]. The HLCA, sequenced by the 10X (53,469 cells) and Smart-seq2 (7,081 cells) platforms, includes 57 annotated cell types with highly variable depth of sampling per cell type (Extended Data Fig. 6). Differential splicing analysis was performed on the 1,754 genes with a computable SpliZ in at least ten cells in one of the

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [2]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA. [3]Department of Biochemistry, Stanford University, Stanford, CA, USA. [4]These authors contributed equally: Julia Eve Olivieri, Roozbeh Dehghannasiri. ✉e-mail: julia.salzman@stanford.edu

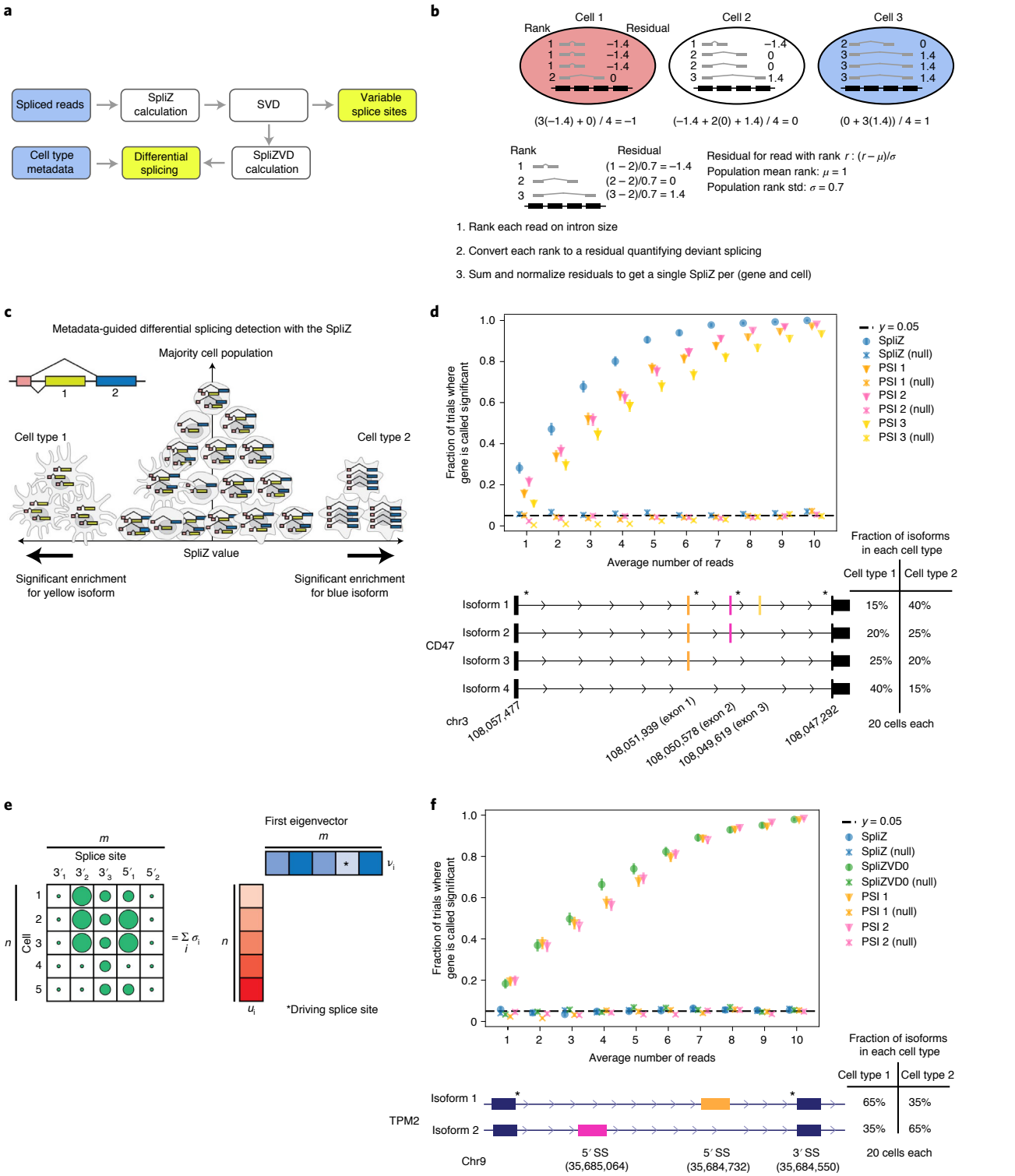
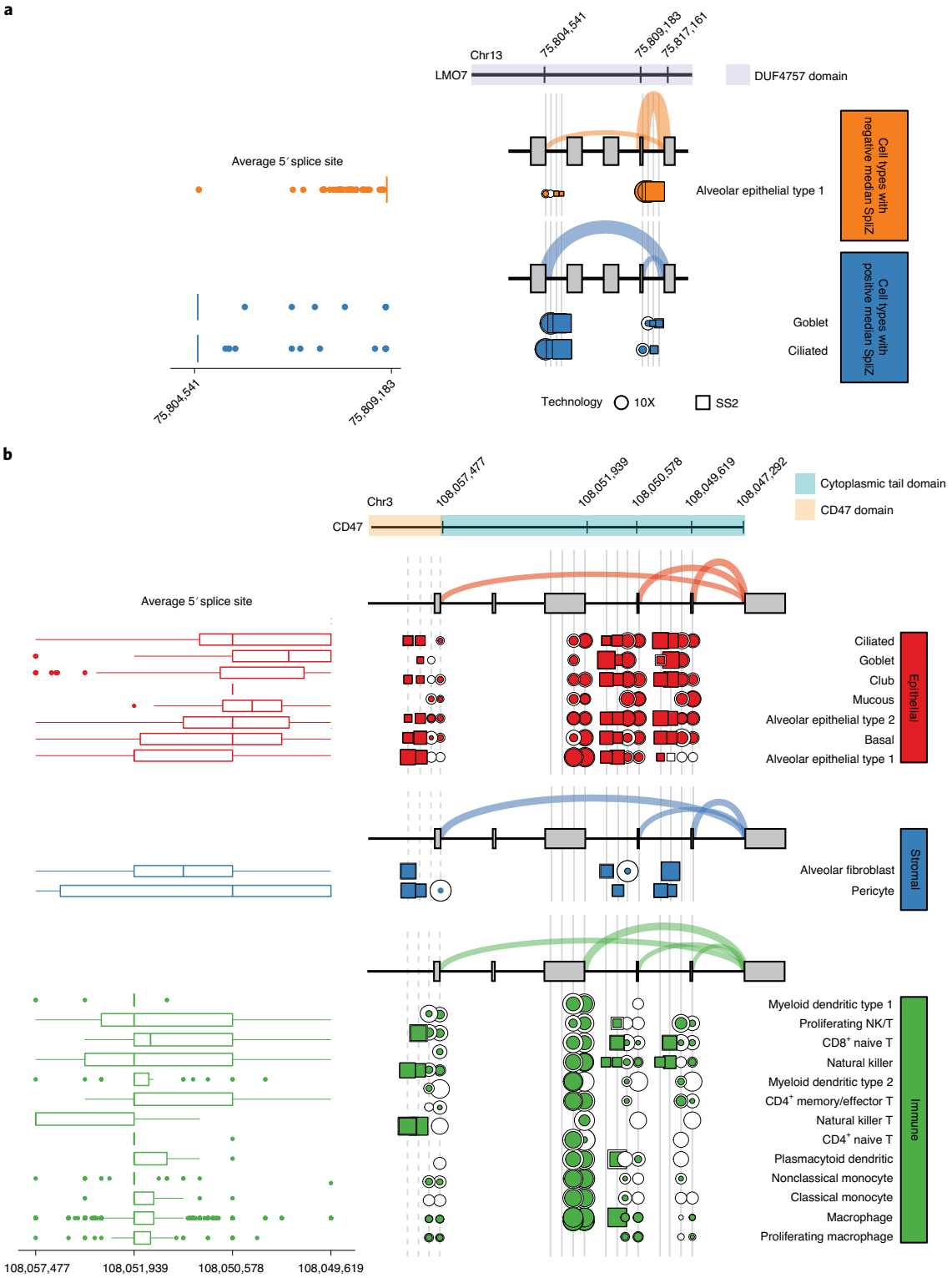


Fig. 1 | The SpliZ outperforms PSI in simulation. a, The SpliZ takes spliced reads and metadata and returns genes described as differentially spliced by cell type. **b**, Illustration of the SpliZ calculation as described in the Methods. **c**, SpliZ scores can be aggregated for each cell type, identifying cell types with statistically deviant splicing. The cell types with significant enrichment for short and long introns have large negative and positive median SpliZ values, respectively. **d**, Simulations of the SpliZ and PSI: reads of two cell types (20 cells each), each cell type with a unique fractional isoform abundance shown (1,000 trials). At each read depth *n*, Poisson(*n*) reads are sampled for each cell in proportion to isoform fractions. Cell populations with identical isoform expressions are used to estimate calls under the null hypothesis (Methods). SpliZsites identified with the SVD are marked with asterisks and coincide with simulated alternative splice sites. Error bars are 95% binomial confidence intervals. The SpliZ increases power over PSI to detect cell types expressing different proportions of *CD47* isoforms. **e**, A representation of a splicing matrix, along with the first eigenvector from its SVD decomposition, with the SpliZsites marked with an asterisk. **f**, Simulation regime identical to **d** except for a different ground truth *TPM2*-like exon structure and isoform proportion. Error bars are 95% binomial confidence intervals. The SpliZVD calls differential alternative splicing between the two cell types with different proportions of *TPM2*-like isoforms with power greater than or equal to PSI, while the SpliZ does not.



individuals (11,640 for Smart-seq2 data) (Methods and Supplementary Table 1).

The SpliZ and SpliZVD identified hundreds of differential alternative splicing events between 10X cell types in each individual. A total of 207 in individual 1 (respectively 219 in individual 2) genes were described as having significant cell-type-specific splicing by the SpliZ, and 135 (142) genes were called by the SpliZVD, with 88 (91) genes called by both (P value < 0.05, Supplementary Table 2). Also, 175 genes were called by either the SpliZ or SpliZVD in both individuals' 10X data ($P = 1.11 \times 10^{-16}$, Extended Data Fig. 7). Restricting analysis to the genes that are significant in both 10X and Smart-seq2 data in the same individual, there is a positive correlation between median SpliZ scores for the same gene and cell type between 10X and Smart-seq2 data (Pearson correlation of 0.315 and 0.650, Extended Data Fig. 8), suggesting that SpliZ results are reproducible between technologies.

Fig. 2 | The SpliZ detects cell-type-specific splicing in HLCA dataset. a, *LMO7* has significantly differential exon skipping affecting a domain of unknown function within epithelial cell types from splice site 75817161. Read fractions from two 10X and Smart-seq2 datasets are shown by dots and squares, respectively. Dot size represents the fraction of reads from each 5′ splice site to the 3′ site at 75817161. For each dot, the outer ring (in white) shows the upper confidence interval and the inner ring (color-coded) shows the lower confidence interval. Box plots show the distribution of the average alternative splice site per cell across all technologies and individuals. Each box shows 25–75% quantiles of average splice site per cell. The mid line of each box is the median and whiskers extend to 1.5 times the interquartile range. All points outside of this range are plotted individually. The cell types are grouped in two sets according to the sign of the median SpliZ for *LMO7*. The sashimi arcs for each cell-type group show the average splice site use across all cell types within that group and also across technologies and individuals. Thicker arcs correspond to higher fractional use measured in both 10X and Smart-seq2. **b**, Differential splicing of *CD47* is tissue-compartment-specific and affects the cytoplasmic tail protein domain. The alternative splicing involves one 3′ splice site (at 108047292) and four 5′ splice sites, including one unannotated splice junction (specified by vertical dashed lines). Cell types have highly tissue-compartment-specific inclusion of 5′ splice sites.

Genes with the most deviant SpliZ values include *ATP5F1C*, a gene encoding a core component of the mitochondrial ATP-synthase machinery, and *MYL6*, a gene encoding an essential component of the actin cytoskeleton with partially characterized splicing (Extended Data Fig. 9a,b and Supplementary Table 3). Other examples of highly cell-type-specifically spliced genes include *LMO7*, a gene encoding an emerin-binding protein with alternative splicing affecting a protein domain of unknown function (Fig. 2a), and *CD47* in which we identified differential alternative splicing involving three isoforms affecting the cytoplasmic tail[16] (Fig. 2b and Supplementary Table 3). *CD47* encodes an immune-regulatory membrane protein that has recently been identified as a therapeutic target in a set of myeloid malignancies[17] but has undescribed cell-type-specific splicing programs. One of the genes with the highest magnitude median SpliZ is *PPP1R12A*, a gene that encodes a protein phosphatase regulatory subunit, where the SpliZsite is not annotated as being alternatively spliced (Extended Data Fig. 9c). *PPP1R12A* is also called by the SpliZVD in Smart-seq2. This supports the use of the SpliZ to not only re-identify known alternative splicing events but also discover cell-type-specific splicing events through a purely statistical and annotation-free approach.

We compared the findings from the SpliZ analysis to Leafcutter[18], a method designed to detect differential splicing in bulk RNA-seq. The intersection between genes called differentially spliced by the SpliZ and Leafcutter was greater than would be expected by chance ($P < 10 \times 10^{-6}$, hypergeometric test), providing orthogonal support for the validity of SpliZ calls (Methods and Supplementary Table 5). However, in both HLCA individuals, Leafcutter called fewer genes (92 compared to 207 for individual 1, 58 compared to 219 for individual 2, Methods). Drawbacks of using Leafcutter for scRNA-seq analysis are discussed in the Supplementary Information. This supports the idea that Leafcutter and similar bulk methods are generally not as appropriate for scRNA-seq analysis: in addition to their requirement of laborious preprocessing steps for pseudobulking, they require multiple pseudobulked samples from each cell type for reliable inference, which makes them statistically unfit for a wide range of scRNA-seq studies with only one or two 10X samples (even when each containing tens of thousands of cells).

Because the SpliZ has a tractable statistical distribution and is single-cell-resolved, it enables integrative biological analysis, including clustering approaches based on the SpliZ alone or in combination with gene expression. In addition, the SpliZ can be correlated with any other numerical phenotype, such as pseudotime, cell cycle trajectory, gene expression, and spatial transcriptomics[19]. Limitations of using the SpliZ on droplet-based data include restricted discovery due to 3′ bias (Supplementary Information). Creating scRNA-seq splicing methods that can discover other splicing changes, such as intron retention, is the subject of future work. In summary, by deconvolving technical and biological noise in splicing, the SpliZ provides a framework to identify and prioritize splicing events that are regulated and functional.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01400-x.

Received: 16 April 2021; Accepted: 18 January 2022;
Published online: 3 March 2022

References

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 Biol.* **19**, 110 (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. Salzman, J., Jiang, H. & Wong, W. H. Statistical modeling of RNA-seq data. *Statistical Sci.* https://doi.org/10.1214/10-STS343 (2011).
8. 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. USA* **108**, 19867–19872 (2011).
9. Trincado, J. L. et al. Suppa2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19**, 40 (2018).
10. Shen, S. et al. rmats: robust and flexible detection of differential alternative splicing from replicate RNA-seq data. *Proc. Natl Acad. Sci. USA* **111**, E5593–E5601 (2014).
11. Najar, C. F. B. A., Yosef, N. & Lareau, L. F. Coverage-dependent bias creates the appearance of binary splicing in single cells. *eLife* **9**, e54603 (2020).
12. Westoby, J., Artemov, P., Hemberg, M. & Ferguson-Smith, A. Obstacles to detecting isoforms using full-length scRNA-seq data. *Genome Biol.* **21**, 74 (2020).
13. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
14. Dehghannasiri, R., Olivieri, J. E., Damljanovic, A. & Salzman, J. Specific splice junction detection in single cells with SICILIAN. *Genome Biol.* **22**, 219 (2021).
15. Travaglini, K. J. et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
16. Hayat, S. M. G. et al. Cd47: role in the immune system and application to cancer therapy. *Cell. Oncol.* **43**, 19–30 (2020).
17. Chao, M. P. et al. Therapeutic targeting of the macrophage immune checkpoint CD47 in myeloid malignancies. *Front. Oncol.* **9**, 1380 (2020).
18. Li, Y. I. et al. Annotation-free quantification of RNA splicing using leafcutter. *Nat. Genet.* **50**, 151–158 (2018).
19. Olivieri, J. E. et al. RNA splicing programs define tissue compartments and cell types at single cell resolution. *eLife* **10**, e70692 (2021).

Methods

Preparing splice junction input files. The scRNA-seq datasets were mapped to the reference human genome (hg38) using STAR v.2.7.5a with default parameters[13].

We used SICILIAN[14] with default settings for calling splice junctions from the STAR BAM files. 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.

SpliZ calculation. The SpliZ score computation for a gene consists of two parts: one relative to the 3′ splice sites in the gene (the 3′ splice site SpliZ) and one relative to the 5′ splice sites (the 5′ splice site SpliZ). We first explain how to calculate the 3′ splice site SpliZ for one gene (suppressing the notation specifying the gene for simplicity) and the 5′ splice site SpliZ can be computed similarly. Let i specify the 3′ splice site, j specify the 5′ splice site, 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 3′ splice site has multiple 5′ splice sites in the dataset.

The 3′ splice site SpliZ score calculation proceeds by treating each 3′ splice site separately. We will consider a plus strand gene and assume that 3′ splice site i has multiple 5′ splice sites across the whole dataset (otherwise we filter it out). We rank these 5′ splice sites in order from closest to the farthest from the 3′ splice site in question i (that is, from lowest to the highest genomic coordinate). For example, if there were four 5′ splice sites partnered with 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 5′ splice sites are ranked differently as described in the Supplementary Information. Let $r_{ijk\ell}$ denote the 5′ splice site rank for the read specified by $ijk\ell$. If $r_{ijk\ell}=1$, it means the 5′ splice site has the smallest genomic coordinate among all 5′ splice sites in the dataset.

Now, let N_i be the number of junctional reads observed across all cells for 3′ splice site i. We can compute $\overline{r_i}$, the average rank of the 5′ splice sites for 3′ splice site 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:

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

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

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

We can see that $E(S_{ijk\ell})=0$ and $\mathrm{var}(S_{ijk\ell})=1$ as we subtract the mean and divide by the standard deviation. The closer the 5′ splice site of a read is to the 3′ splice site, the smaller its corresponding $S_{ijk\ell}$ value would be. In practice, we truncate the $S_{ijk\ell}$ values at the tenth and 90th quantiles across all genes to avoid effects from extreme outliers.

We now aggregate these zero-mean and unit-variance variables for all 3′ splice sites in the gene to compute the 3′ splice site SpliZ (z^d_k) in the kth cell as:

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

where M_k is the number of junctional reads mapping to the gene in the kth cell. It is straightforward to see that $E(z_k)=0$. Note that under the alternative hypothesis, for a given cell type $E(S_{ijk\ell})=\mu\neq 0$. For a cell of this cell type,

$$E[z^d_k] = E\left[\frac{\sum_{i,j,\ell} S_{ijk\ell}}{M_k}\right] = \frac{\sum_{i,j,\ell} E[S_{ijk\ell}]}{M_k} = \frac{M_k\mu}{M_k} = \mu,$$

meaning that the SpliZ is not correlated with read depth M_k.

Knowing that the variance of the sum of independent random variables is the sum of their variances:

$$\mathrm{var}(z^d_k) \approx \frac{\sum_{ij\ell} \mathrm{var}(S_{ijk\ell})}{M_k^2} = \frac{1}{M_k}.$$

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 5′ splice site SpliZ z^a_k. We average the two scores z^a_k and z^d_k to compute the SpliZ z_k for the gene in cell k:

$$z_k = (z^d_k - z^a_k)/\sqrt{2}.$$

These values are subtracted to correct for signs, such that short introns correspond to small values and long introns correspond to high values for both. Division by $\sqrt{2}$ ensures that the variance will be comparable between averaged SpliZ values and SpliZ values for which only one of z^d_k and z^a_k is calculable (in which case that is the SpliZ value for the cell and there is no averaging).

For computing SpliZ, we only consider cells with at least six junctional reads mapping to the gene, and the junctional reads for which the 3′ splice site has more than one 5′ splice site observed in the dataset and vice versa.

Equivalence of the SpliZ and PSI. We will prove the equivalence of the SpliZ and PSI under a specific scenario: when only one of the junctional reads of the exon inclusion event is measured; this could correspond to different ending exons of the transcript.

Let n_1 be the number of reads for a given cell mapping to the junction between the 3′ splice site and the first exon, and let n_2 be the number of reads mapping between the 3′ splice site and the second exon for that cell. Then the value of PSI for the cell is $\psi = \frac{n_1}{n_1+n_2}$.

To calculate the SpliZ value z, we need to calculate the rank mean and standard deviation across the population of cells. Assume in the overall population the fraction of junctions including the exon is f. Then the mean rank $\overline{r} = \frac{1\times(f)N+2\times(1-f)N}{N} = 2-f$, where N is the total number of reads mapping to these two junctions across all cells. The variance σ^2 is given by

$$\sigma^2 = \frac{(f)N(1-(2-f))^2 + (1-f)N(2-(2-f))^2}{N} = f(1-f).$$

Then the SpliZ value is given by

$$z = \frac{n_1\left(\frac{1-(2-f)}{\sqrt{f(1-f)}}\right) + n_2\left(\frac{2-(2-f)}{\sqrt{f(1-f)}}\right)}{n_1+n_2} = \frac{1}{\sqrt{f(1-f)}}\left(f - \frac{n_1}{n_1+n_2}\right).$$

Therefore, z can be written in terms of ψ and f:

$$z = \frac{1}{\sqrt{f(1-f)}}(f-\psi),$$

implying that z and ψ are equivalent in this case. For example, if $f=0.5$ then $z=1-2\psi$.

SpliZVD calculation. Let M be an $n\times p$ matrix, where n is the number of cells and p is the number of splice sites for the given gene. Matrix entries are defined by:

$$M_{ki} = \frac{\sum_{\ell\in L_{ik}} \tilde{S}(\ell)}{|L_{ik}|}.$$

Here, L_{ik} is the set of reads using splice site i in cell k. $\tilde{S}(\ell)$ is the normalized residual of the rank of read ℓ, defined as follows. For a given gene, let G be the set of spliced reads that map to the gene across the dataset. Let $S(\ell)$ be the residual of read l. Then let

$$\mu = \frac{\sum_{\ell\in G} S(\ell)}{|G|},$$

and

$$\sigma^2 = \frac{\sum_{\ell\in G} (S(\ell)-\mu)^2}{|G|}.$$

then $\tilde{S}(\ell) = \frac{S(\ell)-\mu}{\sigma}$, so $\mathrm{E}[\tilde{S}(\ell)] = 0$ and $\mathrm{var}\tilde{S}(\ell) = 1$. Let $\boldsymbol{\alpha}^{(j)}$ be the jth eigenvector of M. Then the jth SpliZVD score for cell k is given by $<m_k,\boldsymbol{\alpha}^{(j)}>$ (note that for this paper we only consider the SpliZVD score based on the first eigenvector, $<m_k,\boldsymbol{\alpha}^{(1)}>$, and refer to this single score as the SpliZVD though more components can be considered). This score has mean 0 and variance $\frac{1}{|L_{ik}|}$ (j suppressed for simplicity in these calculations):

$$E[<m_k,\alpha>] = E\left[\sum_{i=1}^{p}\alpha_i\left(\frac{\sum_{\ell\in L_{ik}}\tilde{S}(\ell)}{|L_{ik}|}\right)\right] = \sum_{i=1}^{p}\alpha_i\frac{\sum_{\ell\in L_{ik}}E[\tilde{S}(\ell)]}{|L_{ik}|} = 0,$$

and

$$\mathrm{var}\,(<m_k,\alpha>) = \mathrm{var}\left(\sum_{i=1}^{p}\left(\alpha_i\frac{\sum_{\ell\in L_{ik}}\tilde{S}(\ell)}{|L_{ik}|}\right)\right)$$

$$= \sum_{i=1}^{p}\alpha_i^2\frac{\sum_{\ell\in L_{ik}}\mathrm{var}(\tilde{S}(\ell))}{|L_{ik}|^2} = \frac{1}{|L_{ik}|}\sum_{i=1}^{p}\alpha_i^2 = \frac{1}{|L_{ik}|}.$$

Note, this is assuming the residuals are uncorrelated. Under the alternative hypothesis for a cell type, as with the SpliZ, for an individual cell type if $E[\tilde{S}(\ell)] = \mu \neq 0$:

$$E[< m_k, \alpha >] = \sum_{i=1}^{p} \alpha_i \frac{\sum_{\ell \in L_{ik}} E\tilde{S}(\ell)}{|L_{ik}|} = \sum_{i=1}^{p} \alpha_i \frac{\mu |L_{ik}|}{|L_{ik}|} = \sum_{i=1}^{p} \alpha_i \mu.$$

Therefore, the SpliZVD value is not dependent on read depth.

Calling differentially alternatively spliced genes by cell type. We perform the following test independently on each gene for each individual, technology and score (SpliZ and SpliZVD). The procedure is the same for both SpliZ and SpliZVD, but we will only discuss the SpliZ here for simplicity. For a given gene, we only consider cell types with more than ten cells with SpliZ values for that gene.

Significance of alternative splicing between cell types in a gene is determined by calculating *P* values using a two-step procedure based on the work in ref. [20]. Consider the following equation from ref. [20]:

$$T_{n,1} = \sum_{i=1}^{k} \frac{n_i}{\hat{\sigma}^2_{n,i}} \left[\hat{\theta}_{n,i} - \frac{\sum_{i=1}^{k} n_i \hat{\theta}_{n,i} / \hat{\sigma}^2_{n,i}}{\sum_{i=1}^{k} n_i / \hat{\sigma}^2_{n,i}} \right]^2.$$

We can compute $T_{n,1}$ based on the SpliZ values for all cells with splicing expression for the given gene as follows: the k samples represent the k cell types with splicing expression for the given gene. Then $\hat{\theta}_{n,i} = \text{median}(X_{i,1}, \ldots, X_{i,n_i})$ is our test statistic, where $X_{i,1}, \ldots, X_{i,n_i}$ are the SpliZ values for the n_i cells from cell type i with splicing expression for the gene. $\sigma^2_{n,i}$ is the sample standard deviation of the SpliZ values for cell type i; in practice, we inflate the variance to $\sigma^2_{n,i} + 0.1$ for robustness. Our null hypothesis is that all cell types have the same median SpliZ for this gene.

Performing all permutations of cell types to cells and recalculating $T_{n,1}$ yields the permutation distribution (a subset of the permutations yields an approximation of the permutation distribution). Theorem 3.1 in ref. [20] states that under some assumptions, this permutation distribution converges to the χ^2 distribution with $k-1$ degrees of freedom. Also, if the sample distributions do not have different medians, the probability that the permutation test rejects the null hypothesis tends to nominal level α.

It is a lot quicker in practice compare to the χ^2 distribution than compute permutations; therefore, we first calculate the *P* value based on comparing to the χ^2 distribution (P_{χ^2}). Then if $P_{\chi^2} < 0.05$, we compute the permutation *P* value P_{perm} by permuting the assignments of cell types to cells for the gene and calculating $T^{(j)}_{n,1}$ based on this permuted data where j is the current permutation. This results in a permutation null distribution $T^{(1)}_{n,1}, \ldots, T^{(J)}_{n,1}$ where J is the number of permutations performed. Then

$$\text{cdf}_{\text{perm}} = \frac{\sum_{j=1}^{J} \mathbb{I}\{T^{(j)}_{n,1} < T_{n,1}\}}{J},$$

and $P_{\text{perm}} = 2\min(\text{cdf}_{\text{perm}}, 1 - \text{cdf}_{\text{perm}})$ (quantifies whether the real value is extreme in either direction). We then adjust the *P* values for multiple hypothesis testing using the Benjamini–Hochberg procedure. Genes are called significant if their adjusted *P* values are less than 0.05.

Genes with the 'most deviant SpliZ values' are defined as follows: find the largest magnitude median SpliZ value Z_i across all cell types for each gene i. Then genes with higher Z_i values have more 'deviant' SpliZ values.

Finding the SpliZsites driving the alternative splicing of a gene. To provide further biological interpretation for the SpliZ and SpliZVD scores and automatically identify the SpliZsites driving the alternative splicing in a significantly regulated gene, we take the SVD of the residual matrix and use its eigenvalues and eigenvectors to select the SpliZsites. To do so, we select up to three eigenvectors depending on the values of their corresponding eigenvalues and then in each chosen eigenvector we define up to three splice sites that correspond to up to three elements of the eigenvector with the highest absolute value as the SpliZsite. If the first eigenvalue is greater than 0.7, we select only the first eigenvector and otherwise we select the second and third eigenvectors if their corresponding eigenvalues are at least 0.2. When an eigenvector is selected, we choose the three splice sites corresponding to the top entries that have at least 10% of the eigenvector loadings based on the L^2 norm of the eigenvector.

For each driving splice site, we additionally report its annotation status, including whether it is annotated as an alternatively spliced exon. To determine whether an exon is known to be involved in alternative splicing, we extracted the splice sites from the hg38 annotation GTF file and then considered those 3′ splice sites (respectively 5′ splice sites) that are observed to be spliced to more than one distinct 5′ splice site (respectively 3′ splice site) across the extracted junctions as known alternatively spliced sites.

Simulation methods. For each simulation, two cell types are simulated with 20 cells in each. They have predefined proportions of each isoform as described in Fig. 1d,f. At each mean read depth n, 100 trials are performed, each of which proceeds as follows. First, a read depth for each cell k is sampled independently from Poisson(n). Then for each cell of cell type c, the distribution of the k reads among splice junctions is drawn from a multinomial distribution, for which each junction from isoform i has the probability $f^{(c)}_i / m_i$, where $f^{(c)}_i$ is the underlying fraction of isoform i in cell type c and m_i is the number of exons in isoform i.

Next, the SpliZ and SpliZVD are calculated as described above for this gene and this population of 40 cells. PSI is calculated on a cell-by-cell basis for each cassette exon by dividing the number of reads including that exon over all of the reads that either skip or include the exon in the given cell (note that some reads may not either skip or include the exon; these are disregarded in the PSI calculation).

Then *P* values are calculated separately for the SpliZ, SpliZVD and each PSI value as described above, except χ^2 filtering is not used and variances are not inflated. Benjamini–Hochberg multiple hypothesis testing correction is not used (there is only one score calculated for each of the SpliZ and SpliZVD; if the multiple PSI values were corrected, it would only cause them to be less significant). The gene is described as having differential alternative splicing between cell types if the calculated $P < 0.05$. SpliZsites are calculated for each by simulating the two cell types at a read depth of 20 and performing the procedure described in the methods.

For each simulation, a 'null' simulation is also included, which follows the same setup except both cell types have the same distribution among the isoforms as the first cell type in the original simulation. This allows us to estimate the type I error rate.

Smart-seq2 data processing. To make the Smart-seq2 data most suitable for validation, we subsetted the dataset to only include junctions that were ever found in 10X data and then ran it through the pipeline (Extended Data Fig. 10). To calculate the correlation between the median SpliZ scores we subsetted each 10X and Smart-seq2 dataset separately for each individual, only including junctions shared by both datasets for that individual, and only cell types shared by both technologies for that individual. Note that the proportions of cell types were not matched.

Concordance between datasets. We test whether the number of genes significant in both 10X datasets is more than expected under the null hypothesis, which is that genes found significant in both datasets are due to chance. Let s be the number genes that are called significant in both individual 1 and 2's 10X data. The probability that a given gene is significant in both datasets is:

$$P(\text{ gene significant in both }) = P(\text{ gene significant in individual 1 })$$
$$\times P(\text{ gene significant in individual 2} | \text{gene significant in individual 1 })$$

Under the null hypothesis, assume that a gene being significant in one individual is independent from it being significant in the other individual. Therefore, under the null hypothesis

$$P(\text{ gene significant in both }) = P(\text{ gene significant in individual 1 })$$
$$\times P(\text{ gene significant in individual 2 })$$

We can estimate the quantities on the right-hand side of the question for each individual:

$$P_i = (\text{ no genes called in individual } i)$$
$$/(\text{ total num genes with computable SpliZ in individual } i)$$

Therefore, under the null hypothesis, the probability that $\geq x$ genes are significant in both individuals out of n genes is given by

$$1 - \text{binom\_cdf}(x, n, P_1 P_2).$$

Correlation between datasets is calculated by considering only gene/cell-type pairs with at least ten cells in both datasets, restricting to genes called significant in both datasets and calculating the Pearson correlation.

Protein domain annotation. Protein domains for analyzed genes were determined by finding whichever protein domains overlapped with the chromosome and coordinates of the splicing in the Pfam file (location downloaded described in the Data Availability section). The *CD47* cytoplasmic domain was added based on[16,21].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

HLCA data was downloaded from the European Genome-Phenome Archive at accession number EGAS00001004344 (ref. [22]). We refer to patient 2 in HLCA as

individual 1 and patient 3 as individual 2 in our analysis. Our cell-type definition is based on concatenating the 'compartment' and 'free annotation' columns from the HLCA metadata and only considering lung cells (not blood). SpliZ scores and Leafcutter results, as well as the original data needed to reproduce these results, are available at FigShare: https://doi.org/10.6084/m9.figshare.14378819.v1. Human RefSeq hg38 annotation file was downloaded from ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz. The UCSC Pfam database for the hg38 genome assembly was downloaded from http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/ucscGenePfam.txt.gz.

Code availability

The SpliZ code along with the code used for data analysis and to create the figures are available through a GitHub repository https://github.com/juliaolivieri/SpliZ_pipeline/. This repository is archived with Zenodo under the following https://doi.org/10.5281/zenodo.5781783 (ref. [23]). The pipeline was written in Python (v.3.6.7), and installed package versions are the following (also available in an environment.yml file on github): matplotlib (v.2.2.3); numpy (v.1.18.4); pandas (v.1.0.4); pyarrow (v.0.15.1); scipy (v.1.4.1); snakemake-minimal (v.5.4.5); statsmodels (v.0.11.1) and tqdm (v.4.46.0). We used Leafcutter (https://github.com/davidaknowles/leafcutter) and regtools (https://github.com/griffithlab/regtools).

References

20. Chung, E. & Romano, J. P. Exact and asymptotically robust permutation tests. *Ann. Stat.* **41**, 484–507 (2013).
21. Li, Y. I. et al. Annotation-free quantification of RNA splicing using leafcutter. *Nat. Genet.* **50**, 151–158 (2018).
22. Travaglini, K. J. et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
23. Olivieri, J. E. juliaolivieri/SpliZ_pipeline: v1.0. *Zenodo* https://doi.org/10.5281/zenodo.5781783 (2021).

Acknowledgements

We thank P. Wang and S. Quake for insightful and instrumental comments during the development of the method, E. Meyer and R. Bierman for comments on the manuscript, J. Klein for creating parts of Fig. 1, and K. Travaglini and M. Krasnow for providing advanced access to the HLCA data before its publication. J.O. is supported by the National Science Foundation Graduate Research Fellowship under grant no. DGE-1656518, a Stanford Graduate Fellowship and a Lieberman Fellowship. R.D. is supported by the Cancer Systems Biology Scholars Program grant no. R25 CA180993 and Clinical Data Science Fellowship grant no. T15 LM7033-36. J.S. is supported by the National Institute of General Medical Sciences grant nos. R01 GM116847 and R35 GM139517 and the National Science Federation Faculty Early Career Development Program Award no. MCB1552196.

Author contributions

J.O. developed the software and analyzed the data. R.D. developed the software and analyzed the data. J.S. conceived and supervised the project. J.O., R.D. and J.S. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

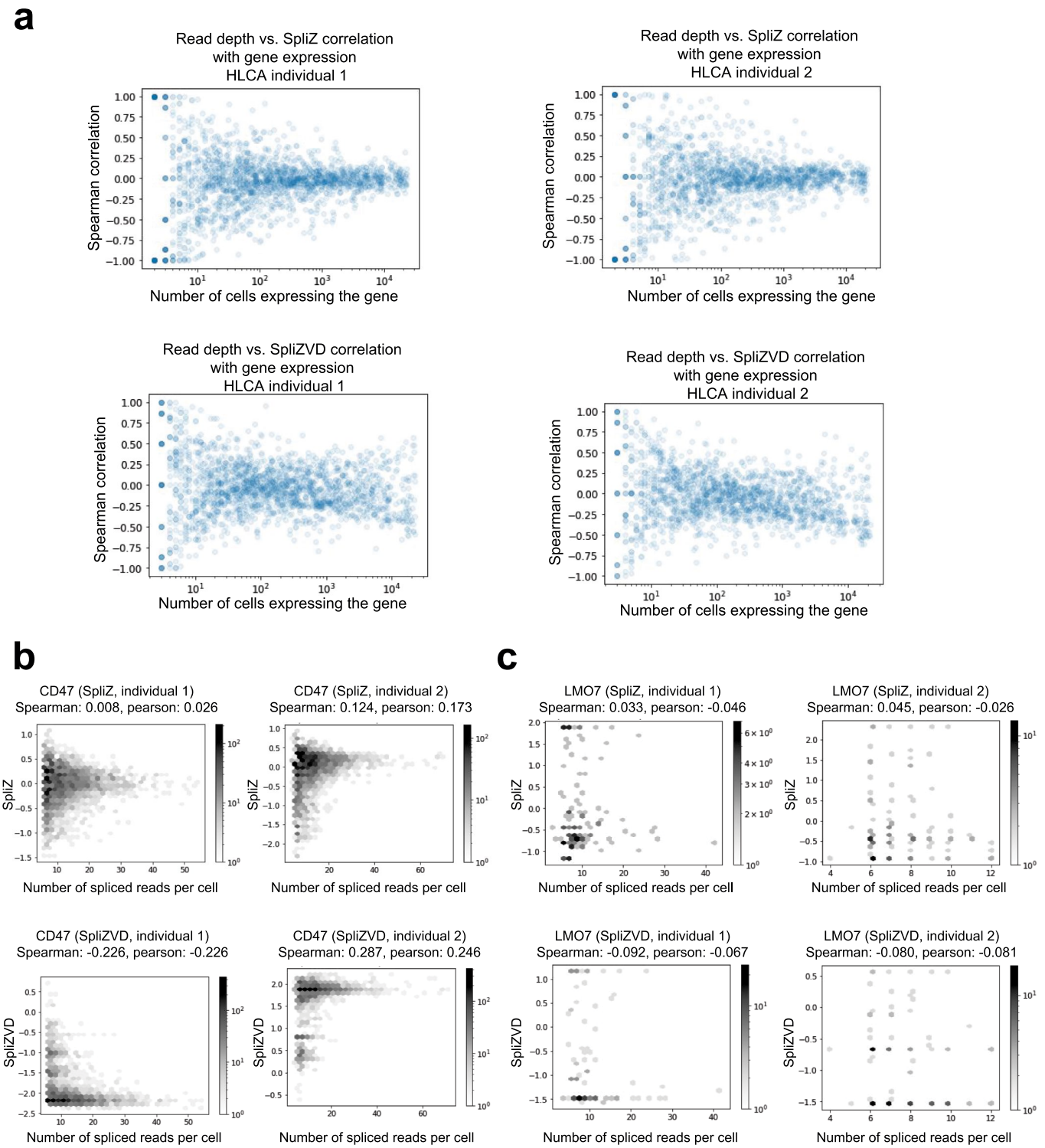
Extended data are available for this paper at https://doi.org/10.1038/s41592-022-01400-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01400-x.

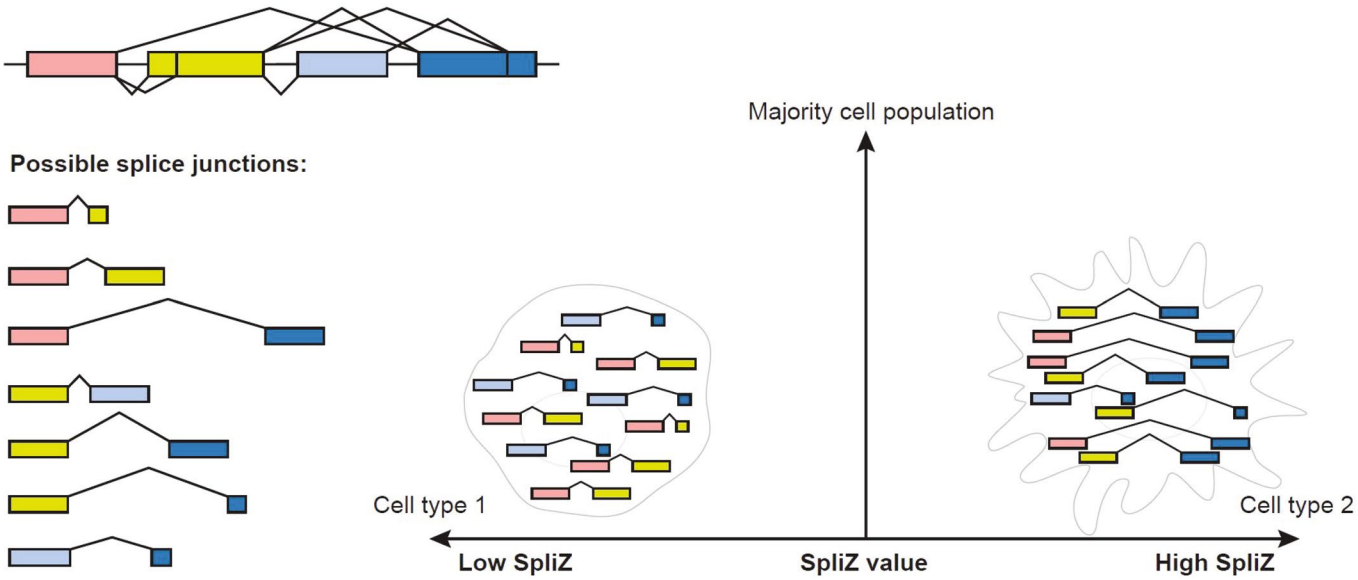
Correspondence and requests for materials should be addressed to Julia Salzman.

Peer review information *Nature Methods* thanks Ángeles Arzalluz-Luque, Yang I. Li and the other, anonymous, reviewer for their contribution to the peer review of this work. Lin Tang, in collaboration with the *Nature Methods* team, was the Primary Handling Editor. Peer reviewer reports are available.

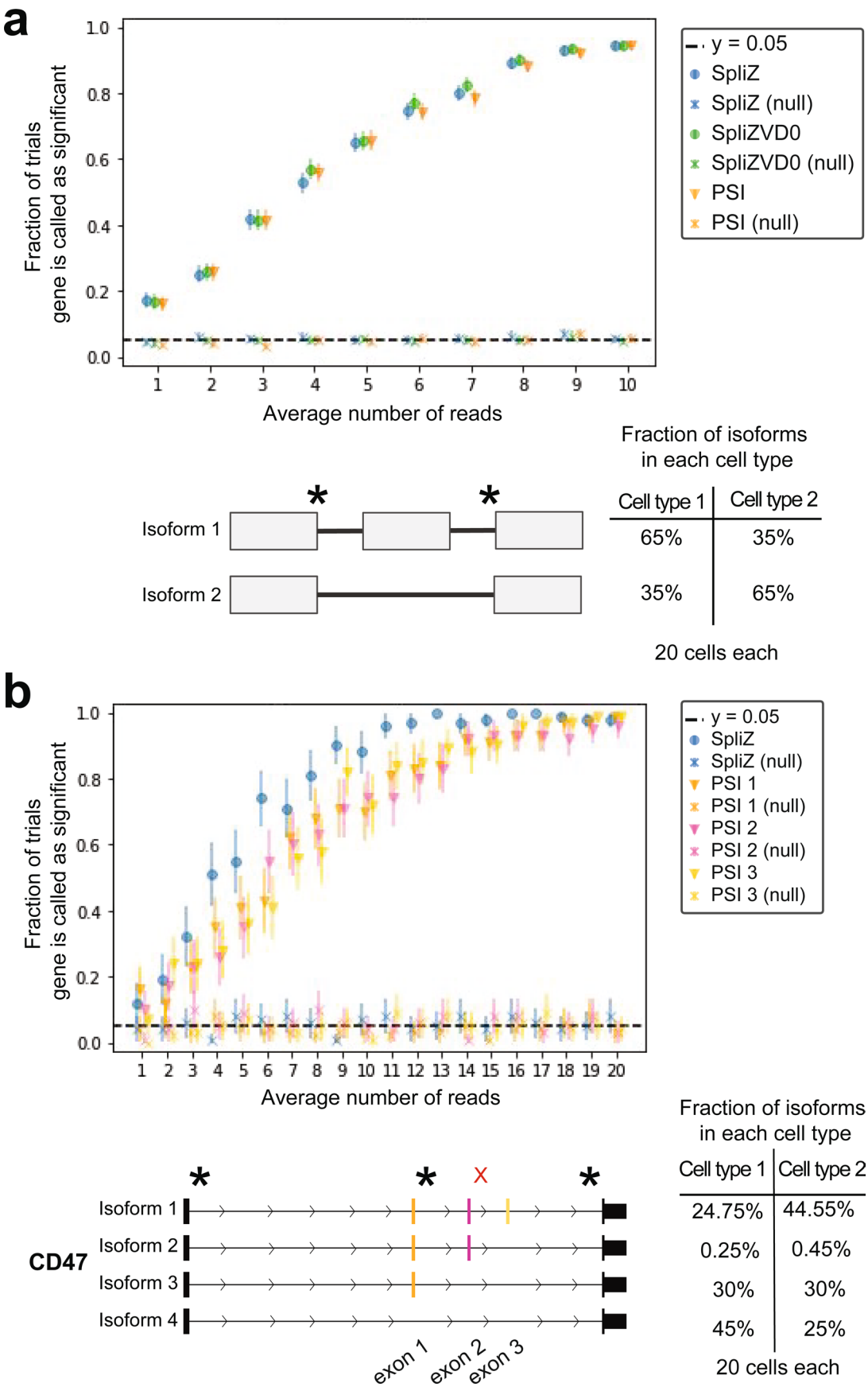
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | The SpliZ is not correlated with gene expression. a. There is no consistent correlation between either the SpliZ or SpliZVD and gene expression. b. Plots of number of spliced reads vs SpliZ and SpliZVD show that there is no significant correlation between gene expression and SpliZ or SpliZVD. c. *LMO7* shows no evidence of correlation between number of spliced reads and SpliZ or SpliZVD.

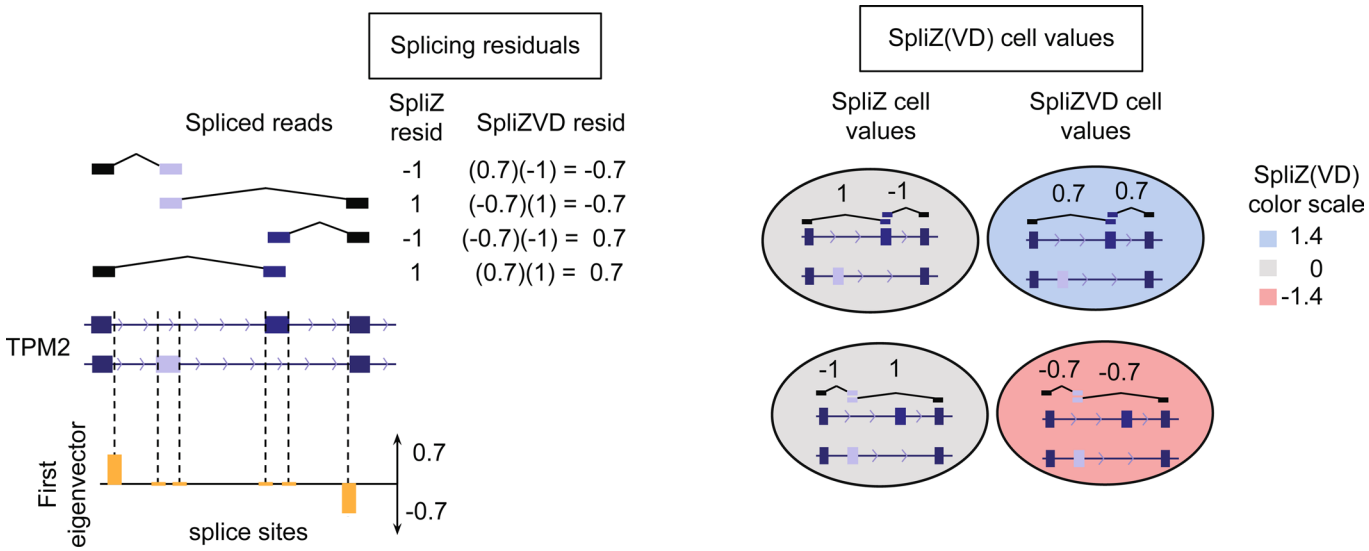


Extended Data Fig. 2 | SpliZ toy example. Cell on the left has short average introns vs the cell on the right, giving it a lower SpliZ.

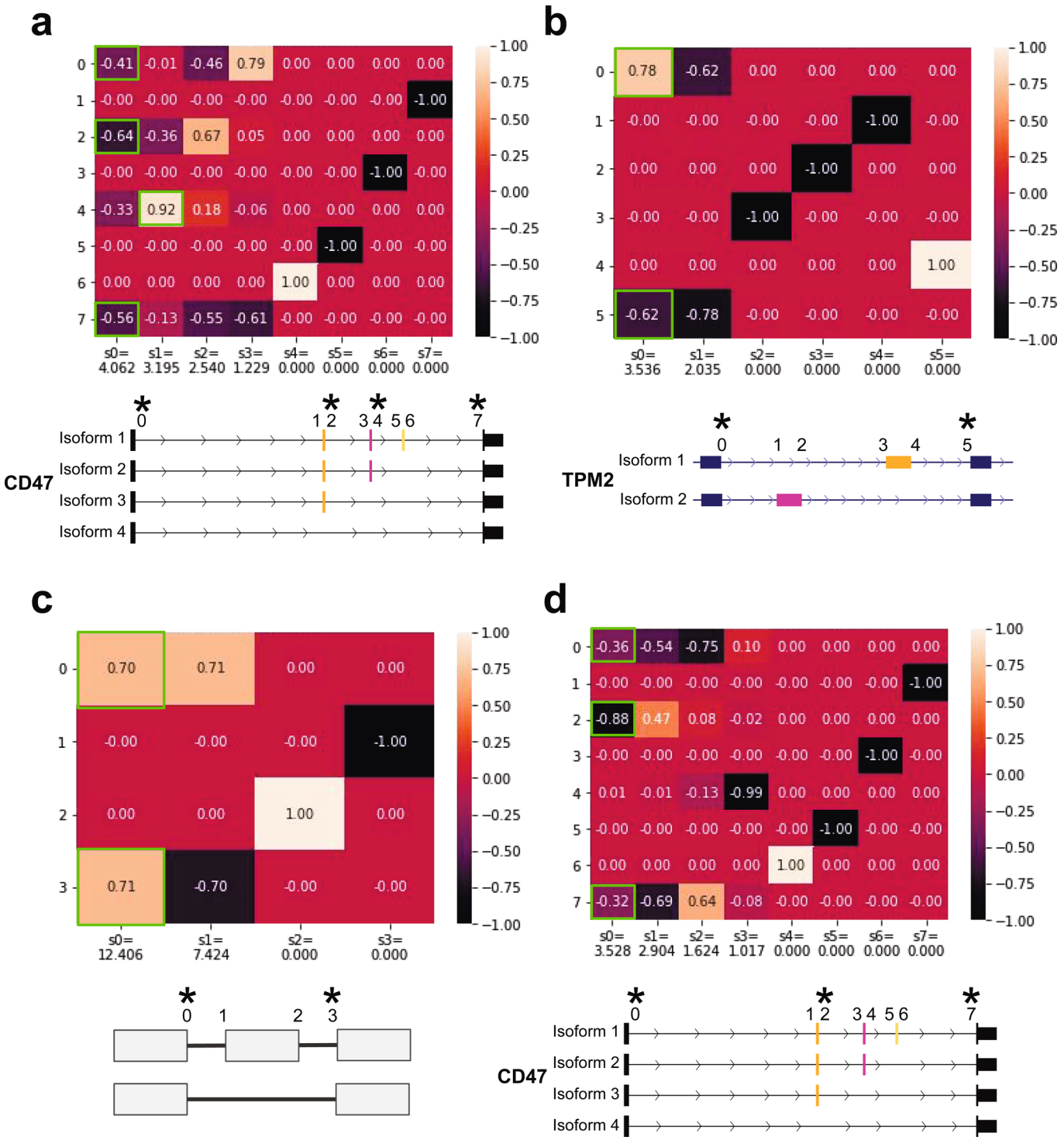


Extended Data Fig. 3 | See next page for caption.

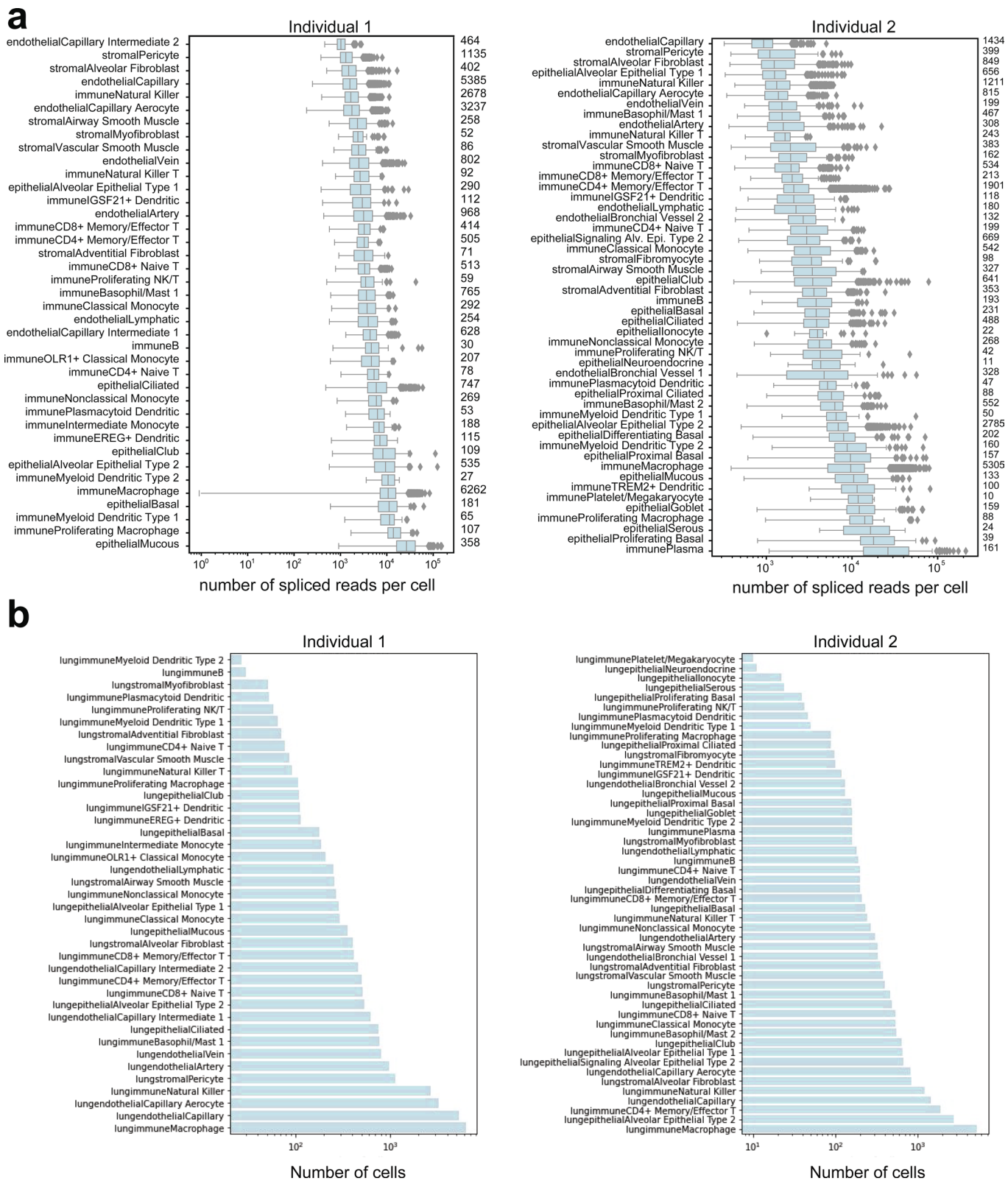
Extended Data Fig. 3 | Simulations of the SpliZ, SpliZVD, and PSI. In both simulations, two cell types with 20 cells respectively are simulated, each type having a different proportion of isoforms (1000 trials for a, 100 trials for b). At each read depth, Poisson(*n*) reads are sampled in proportion to isoform abundances. Null values are calculated from cell populations with identical isoform expression. SpliZsites based on the SVD described in the methods are starred by asterisks, and coincide with all splice sites with differential partner exon usage between the two cell types. a. The SpliZ, SpliZVD, and PSI all have the same power in the case of exon skipping. b. The simulation from Fig. 1d was modified, changing the fractions of isoform abundance as described in the figure such that when exon 2 is present, exon 3 is included 99% of the time. While this splice site was identified as a SpliZsite in Fig. 1d, here it is not identified (red X), as would be expected for a splice junction with 99% constitutive splicing. This simulation setting shows that the SpliZ can correctly identify only alternative exons as SpliZsites.



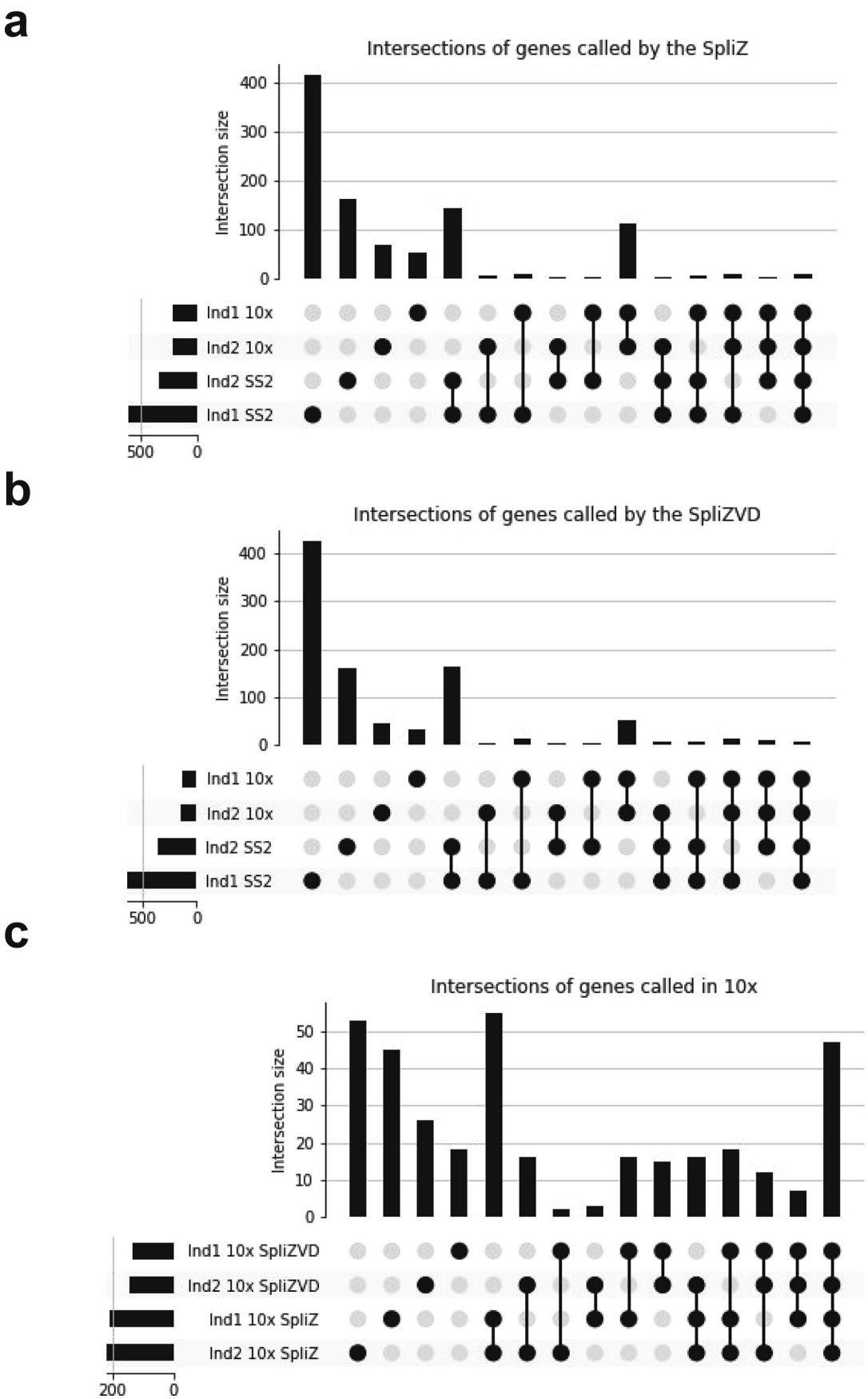
Extended Data Fig. 4 | SpliZVD calculation example. The SpliZVD is the projection of the matrix of splicing residuals onto its first eigenvector. The residual matrix's SVD is used to identify the most variably alternatively spliced sites. The top left shows a gene structure of *TPM2*, with reads aligning to different junctions above. Each read is assigned a SpliZ residual and SpliZVD residual, the latter of which is based on the first eigenvector (shown below the gene annotations). Ovals representing different cells are colored based on the sign of their SpliZ or SpliZVD values, showing that in this case the SpliZVD is able to distinguish differences in splicing where the SpliZ cannot.



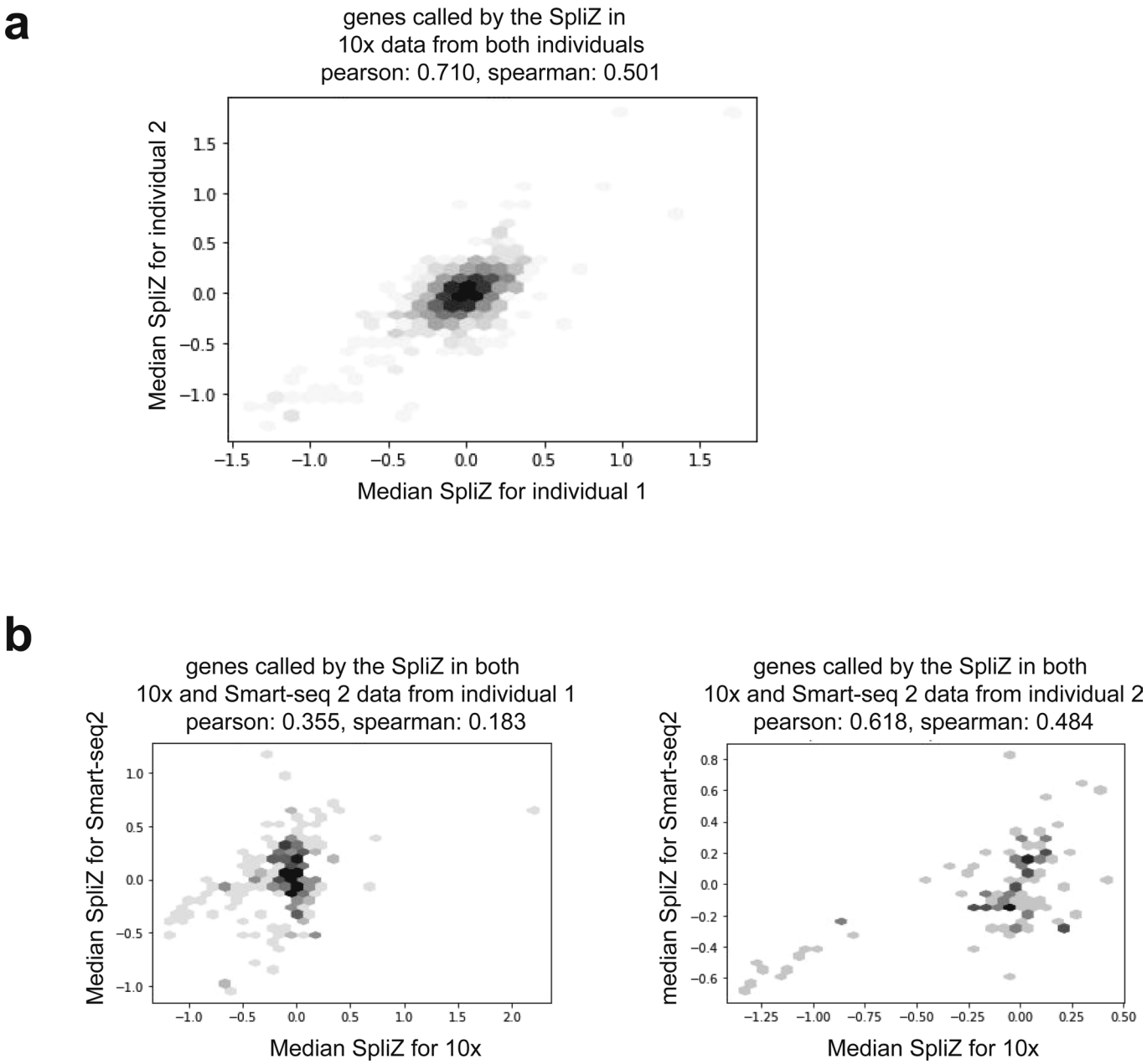
Extended Data Fig. 5 | Splicing residual matrices. SVD decompositions of the splicing residual matrix based on the simulations in Fig. 1d,f, and Extended Data Fig. 3 at average read depth 20. Rows correspond to splice sites and columns correspond to eigenvectors. For each splice site, the value from the SVD that causes the splice site to be picked as SpliZsite is boxed in green. a,b,c. As expected, all non-constitutive splice sites are selected as SpliZsites. d. As expected, the three most variable splice sites are chosen as SpliZsites, while the 99% constitutive splice site is not chosen as SpliZsite.



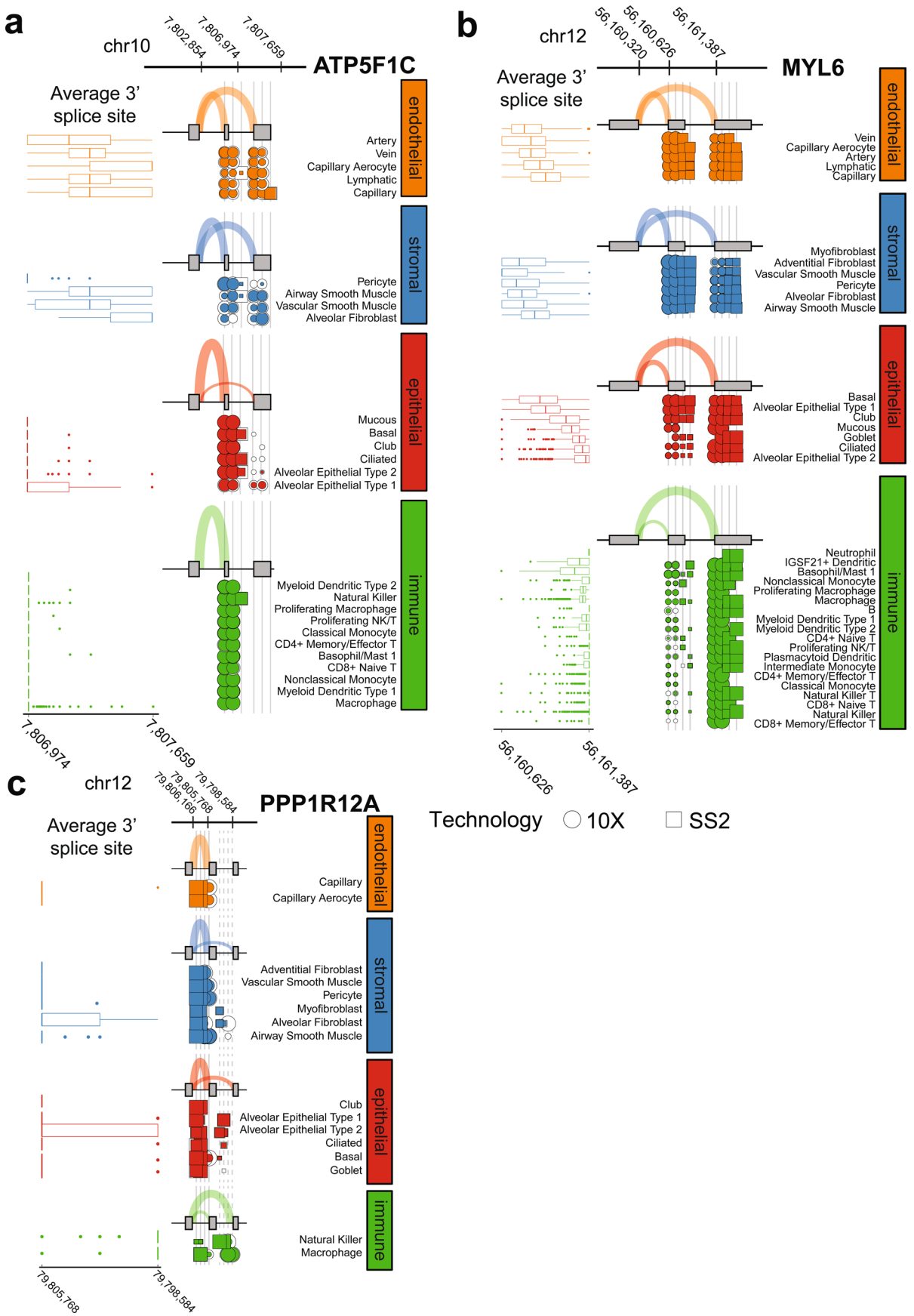
Extended Data Fig. 6 | Cell type summary in HLCA dataset. a. Box plot of the number of reads per cell for each cell type in both individuals' 10x data. Numbers to the right of the plot indicate the number of cells plotted for each cell type over the given experiment. The middle line of each box is the median, and each box extends from the first to third quartile. Whiskers extend to 1.5 times the interquartile range. All points outside of this range are plotted individually. b. Bar plots of the number of cells per cell type for each individuals' 10x data.



Extended Data Fig. 7 | Called gene intersections. a. Intersection of genes called by the SpliZ between 10X and subsetted Smart-seq2 data. b. Intersection of genes called by the SpliZVD between 10X and subsetted Smart-seq2 data. c. Intersection between genes called by the SpliZ and SpliZVD in 10x data.

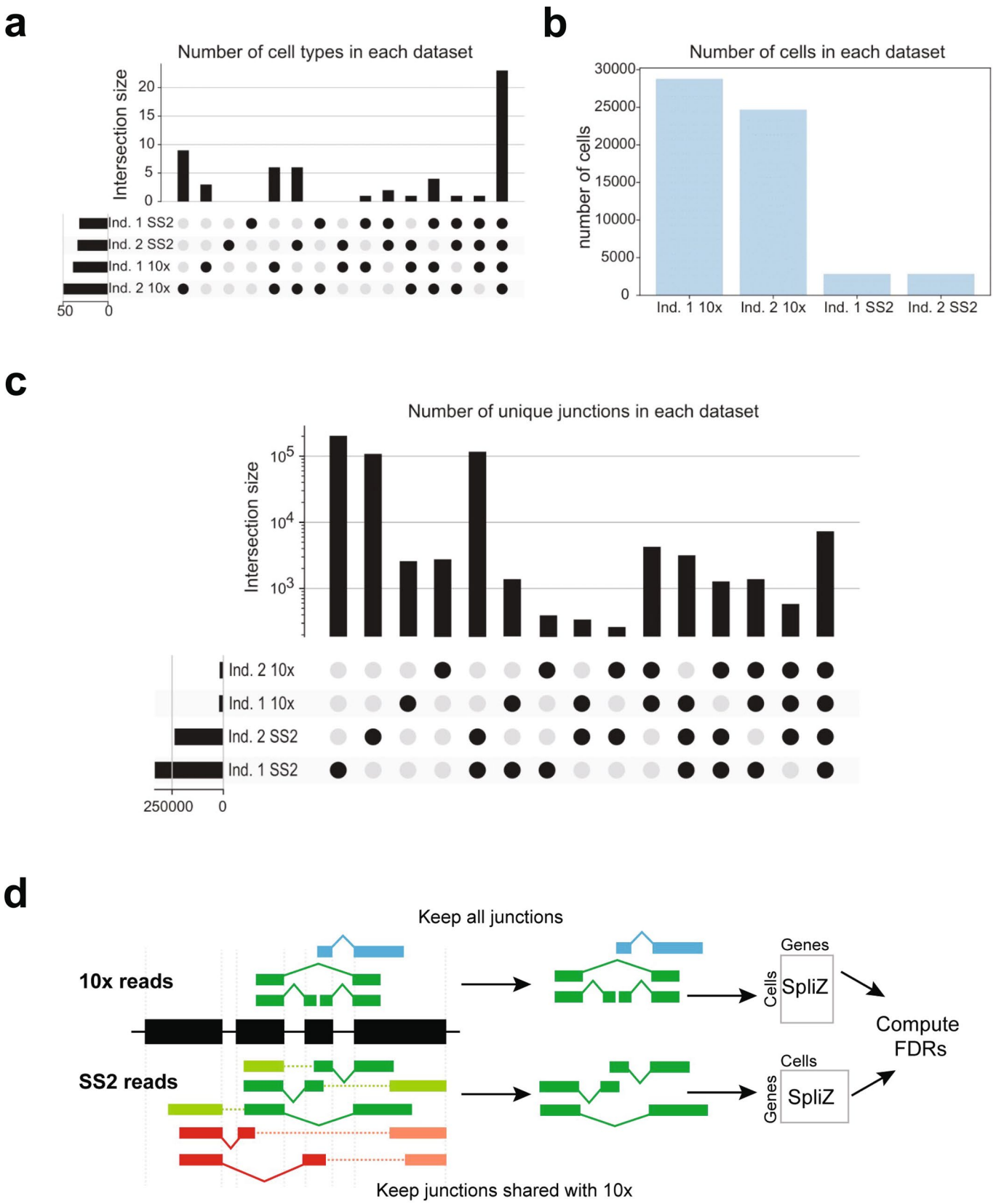


Extended Data Fig. 8 | SpliZ correlations. Correlation between median SpliZ values per matched gene and cell type a. Between individuals for 10X data. b. Between technologies for each HLCA individual (datasets subsetted to shared junctions and cell types per individual before running).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Compartment-specific splicing. The compartment-specific regulated alternative splicing of (a) *ATP5FC1*, (b) *MYL6*, and, (c) *PPP1R12A*. For *ATP5F1C*, endothelial and stromal cells have a higher fraction of junctional reads for the exon exclusion event compared to other compartments. For *MYL6*, immune cells have lower fraction of junctional reads for the exon inclusion event compared to other compartments. For *PPP1R12A*, immune cells have a higher fraction of junctional reads for the exon inclusion event. The plots include the cell types in which the splice site has at least 20 junctional reads in at least 10 cells in at least one of the 4 datasets (two individuals and two technologies: 10X and SS2) were chosen. Dots represent the fraction of junctional reads for each alternative splice site in the celltype. For each dot, the outer ring (in white) shows the upper CI and the inner ring (color-coded) shows the lower CI. Box plots show the average alternative splice site for each cell across all technologies and individuals. Each box shows 25-75% quantiles of average splice site per cell. Arcs represent the average fractions at the compartment level. There is no dot when the alternative splice site has 0 junctional reads.



Extended Data Fig. 10 | 10X vs Smart-seq2 comparison. a. Upset plot showing the number of cell types sequenced in each HLCA individual and technology. b. Bar plot showing the number of cells sequenced for each HLCA individual and technology. c. Upset plot showing the number of alternative junctions (junctions for which at least one splice site has at least one other partner in the dataset) for each HLCA individual and technology. d. The SpliZ is calculated independently for Smart-seq2 data restricted to junctions detected by 10X to measure technology-dependence of results.

nature portfolio

Corresponding author(s): Julia Salzman

Last updated by author(s): Dec 17, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The data used in our analysis was based on a published paper (https://www.nature.com/articles/s41586-020-2922-4), and the authors in that paper provided a thorough description of their procedure for generating the raw sequencing files used in our analysis. We did not use any software for data analysis.
Data analysis	We used SpliZ version 1.0 for our analysis in the manuscript (DOI: 10.5281/zenodo.5781783), SICILIAN v1.0.0 (DOI: 10.5281/zenodo.5081832), STAR v2.7.5 for identifying splice junctions in the data, matplotlib (2.2.3); numpy (1.18.4); pandas (1.0.4); pyarrow (0.15.1); scipy (1.4.1); snakemake-minimal (5.4.5); statsmodels (0.11.1); tqdm (4.46.0), Leafcutter (https://github.com/davidaknowles/leafcutter), and regtools (https://github.com/griffithlab/regtools)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

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. The cell types we use here are based on concatenating the "compartment" and "free annotation" columns from the

HLCA metadata and only considering lung cells (not blood). For the purposes of review, Tables 1, 3, and 4, as well as input data for the pipeline, are available at the following FigShare DOI: https://doi.org/10.6084/m9.figshare.14378819.v1.

Human RefSeq hg38 annotation file was downloaded from: ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz

The UCSC Pfam database for the hg38 genome assembly was downloaded from: http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/ucscGenePfam.txt.gz

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A complete scRNA-seq human lung dataset from two individuals was used in our analysis. Technical replicates were chosen based on available material.
Data exclusions	We excluded all cells without a cell type analysis and all cells that weren't assigned to the lung tissue (e.g. blood cells).
Replication	The primary dataset of discovery was 10x data in individual 1. We used 10x data from individual 2, Smart-seq2 data from individual 1, and Smart-seq2 data from individual 2 as replicates. There was a positive correlation in SpliZ scores for genes significant by both 10x and Smart-seq 2 data, as well as genes called in both individuals 10x data, as evidence of reproducibility.
Randomization	There was no allocation into sample groups.
Blinding	Blinding was not performed in this study because analysis was performed after data had been collected and published, and this was exploratory analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging