RESEARCH ARTICLE

# Figures and figure supplements

RNA splicing programs define tissue compartments and cell types at single-cell resolution

**Julia Eve Olivieri *et al***

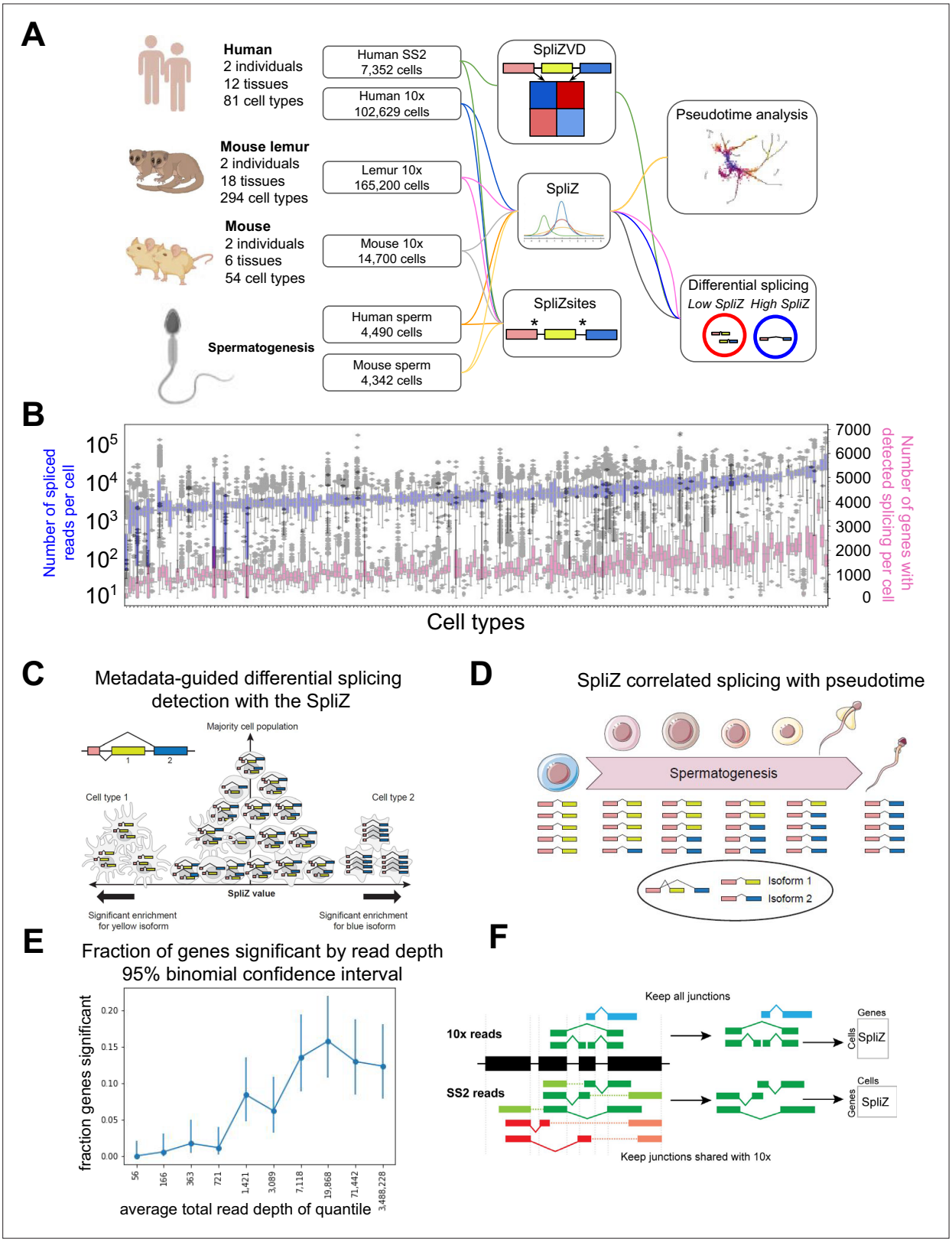


**Figure 1.** Analysis of alternative splicing in single-cell RNA-seq. (**A**) Human, mouse lemur, and mouse single-cell RNA-seq from 10X and SS2 were run through the SpliZ pipeline for differential splicing discovery. (**B**) 10X data from the first human individual contains 82 cell types with variable sequencing depth. (**C**) Given cell type annotation, SpliZ scores can be aggregated for each cell type, allowing identification of cell types with statistically deviant splicing. This figure was also included as **Figure 1C** in ***Olivieri et al., 2021***, published under the Creative Commons Attribution Non-Commercial

*Figure 1 continued on next page*

*Figure 1 continued*

No Derivatives 4.0 International License (CC BY-NC-ND 4.0; https://creativecommons.org/licenses/by-nc-nd/4.0/). (**D**) Cell-wise SpliZ values can be correlated with pseudotime to identify developmentally regulated alternative splicing. (**E**) The fraction of genes called as having significant differential alternative splicing by cell type is higher at higher sequencing depths, plateauing at around 20,000 spliced reads in the dataset, at which point around 15% of genes were called as significant. (**F**) The SpliZ is calculated independently for SS2 data restricted to junctions found in 10X data, and used to validate results from 10X data. This figure was also included as **Figure 2D** in ***Olivieri et al., 2021***, published under the Creative Commons Attribution Non-Commercial No Derivatives 4.0 International License (CC BY-NC-ND 4.0; https://creativecommons.org/licenses/by-nc-nd/4.0/).

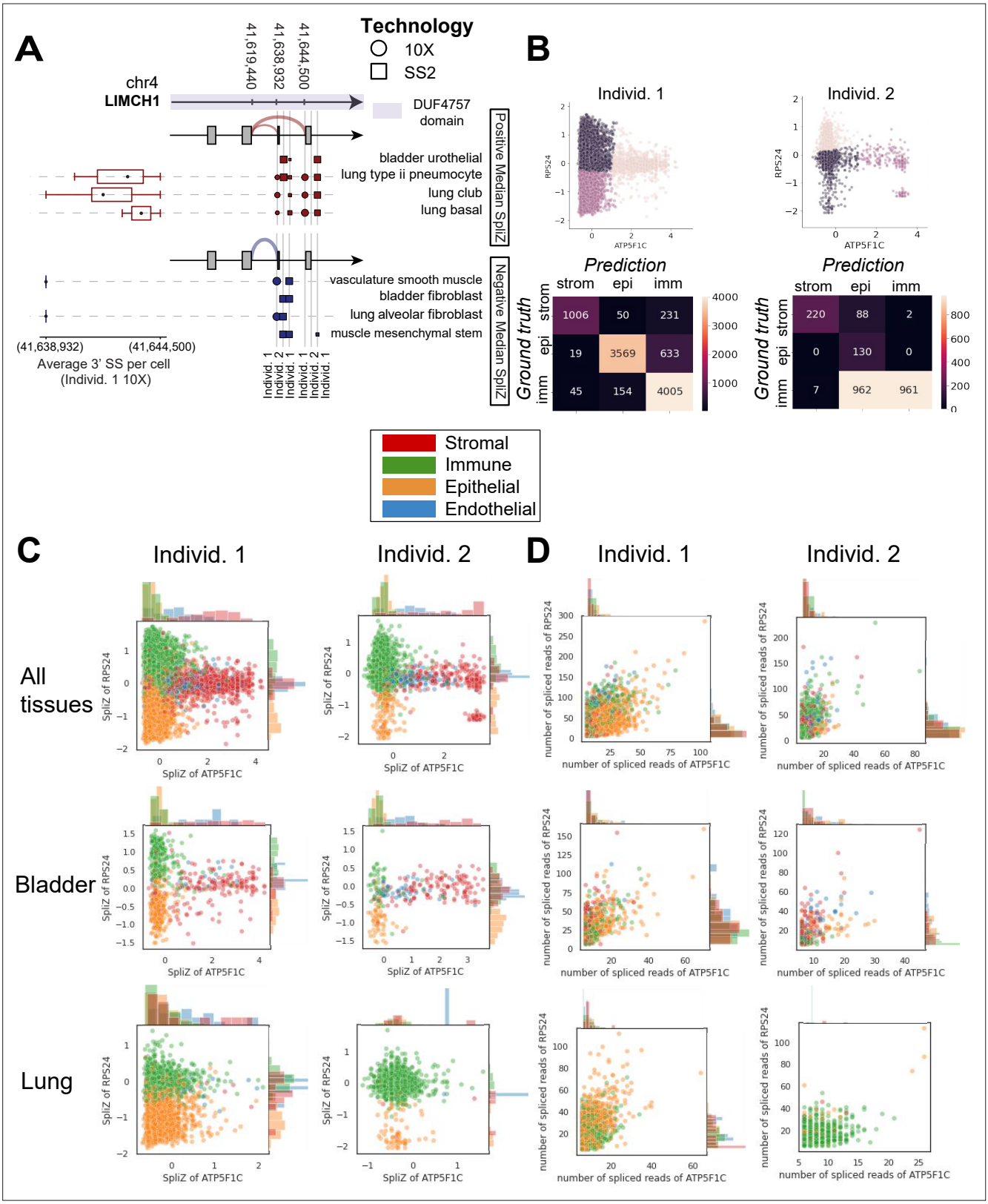


**Figure 2.** Compartment-specific alternative splicing revealed by applying the SpliZ to scRNA-seq data. (**A**) Dot and sashimi plots showing *LIMCH1* compartment-specific exon skipping (involving 5′ splice site [5′ SS] 41,619,440 and two 3′ splice sites [3′ SS] 41,638,932 and 41,644,500) impacting a protein domain of unknown function DUF4757 (shown by the purple color on the gene structure) across cell types and 10X and SS2 data from both human individuals. Each dot shows junction expression for the splice junction from the 5′ SS to one of the 3′ SS's, with dot size proportional to the

*Figure 2 continued on next page*

*Figure 2 continued*

fraction of junctional reads supporting the splice junction in that cell type and dataset. Columns of dots are biological replicates; the first column is the individual 1 10X dataset (circles) and the next two columns are SS2 datasets from individuals 1 and 2 (squares). Cell types are grouped in two sets depending on the sign of the median SpliZ score in 10X data from human individual 1. The thickness of the sashimi arcs represents the fraction of the reads mapping to each 3′ SS when all datasets and corresponding cell types for the sashimi arc are grouped together. The box plot for each cell type shows the distribution of the weighted average 3′ SS (weights being the number of reads aligning to each 3′ SS in the cell) for each cell and the reads are assigned 1 (for those aligning to the closer 3′ SS) and 2 (for those aligning to the farther 3′ SS). Stromal cells including vasculature smooth muscle cells and fibroblasts always include the exon (higher fraction of reads aligning to the splice site at 41,638,932), while epithelial cells including bladder urothelial cells skip with >50% rate. (**B**) Unsupervised k means clustering results in 78, 84, and 95% accuracy of compartment classification for the stromal, epithelial, and immune compartments, respectively, for individual 1, and 70, 100, and 49% , respectively, for individual 2. (**C**) The SpliZ scores of the genes *ATP5F1C* and *RPS24* together separate compartments in both human individuals. Each dot represents the SpliZ score in a single cell and is color coded by the compartment. (**D**) Using the spliced read counts for each gene rather than the SpliZ does not separate the compartments, showing that this separation is not driven by gene expression differences. Each dot represents the number of spliced reads in a single cell and is color coded by the compartment.

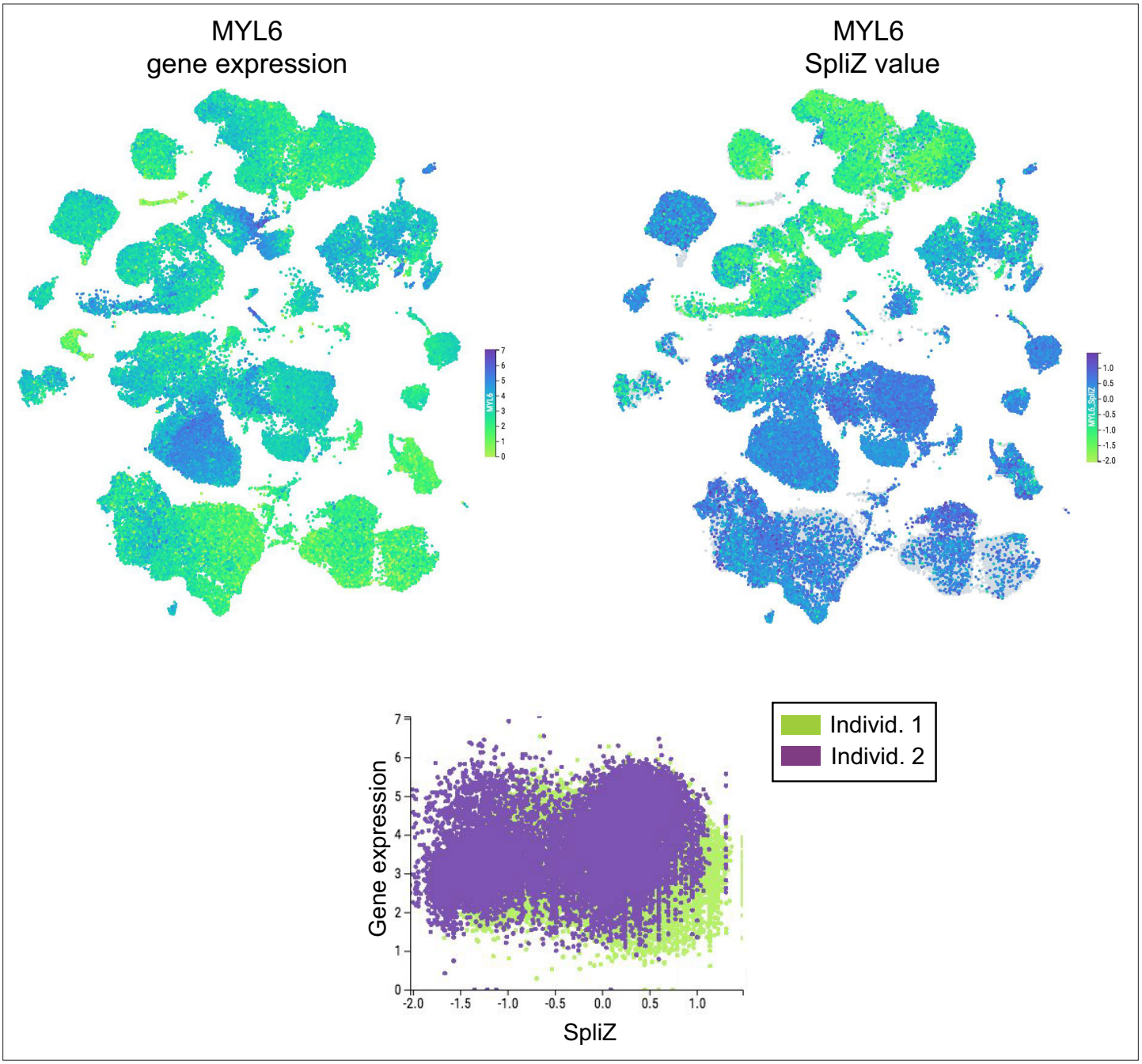


**Figure 2—figure supplement 1.** Gene expression and the SpliZ value of *MYL6* across single cells in the *Tabula Sapiens* dataset. Coloring *Tabula Sapiens* cells by both gene expression and SpliZ value shows that *MYL6* is ubiquitously expressed and that the SpliZ is independent of gene expression for these cases. Plots are obtained by using the cellxgene (***Megill et al., 2021***) visualization platform.

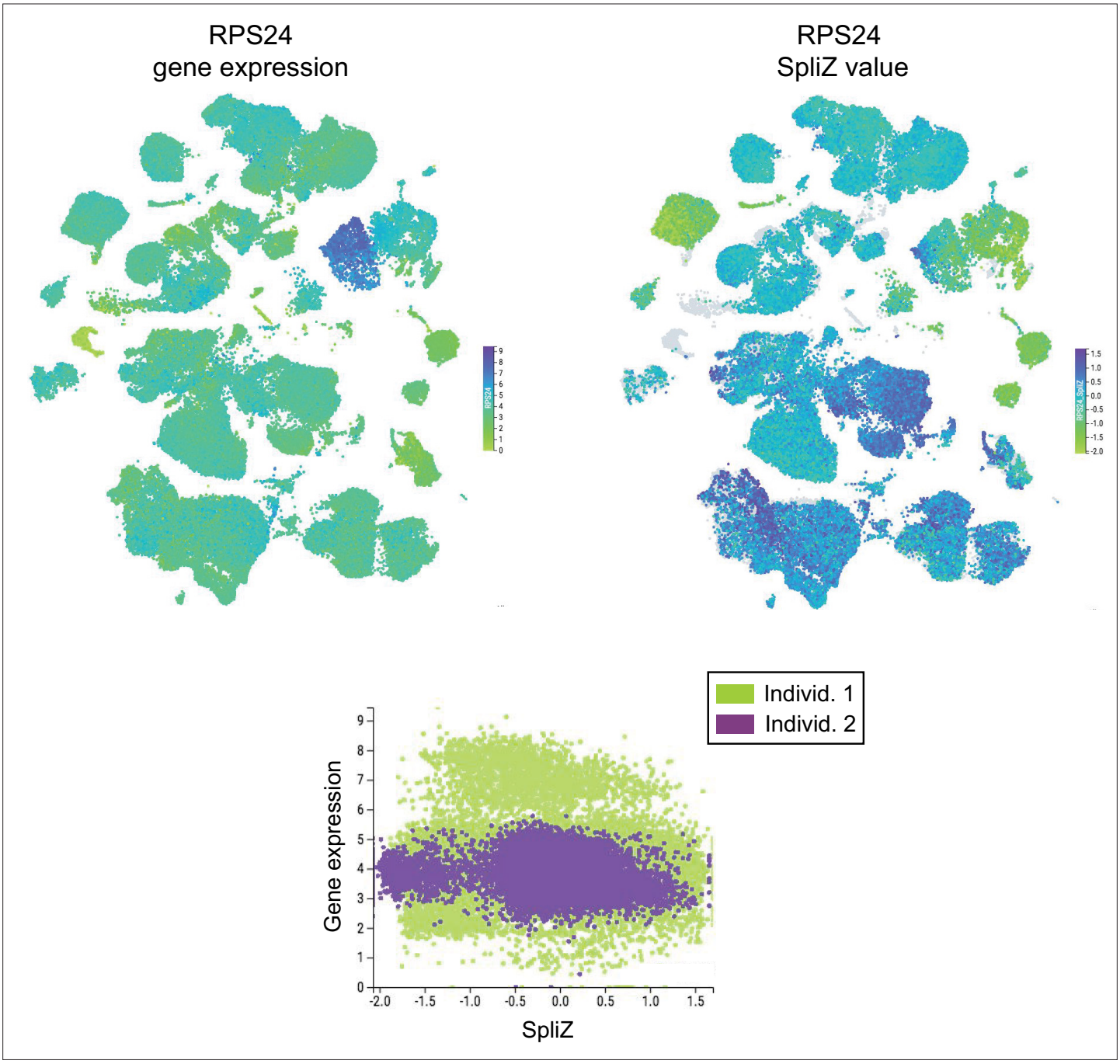


**Figure 2—figure supplement 2.** Gene expression and the SpliZ value of *RPS24* across single cells in the *Tabula Sapiens* dataset. Coloring *Tabula Sapiens* cells by both gene expression and SpliZ value shows that *RPS24* is ubiquitously expressed and that the SpliZ is independent of gene expression for these cases. Plots are obtained by using the cellxgene (***Megill et al., 2021***) visualization platform.

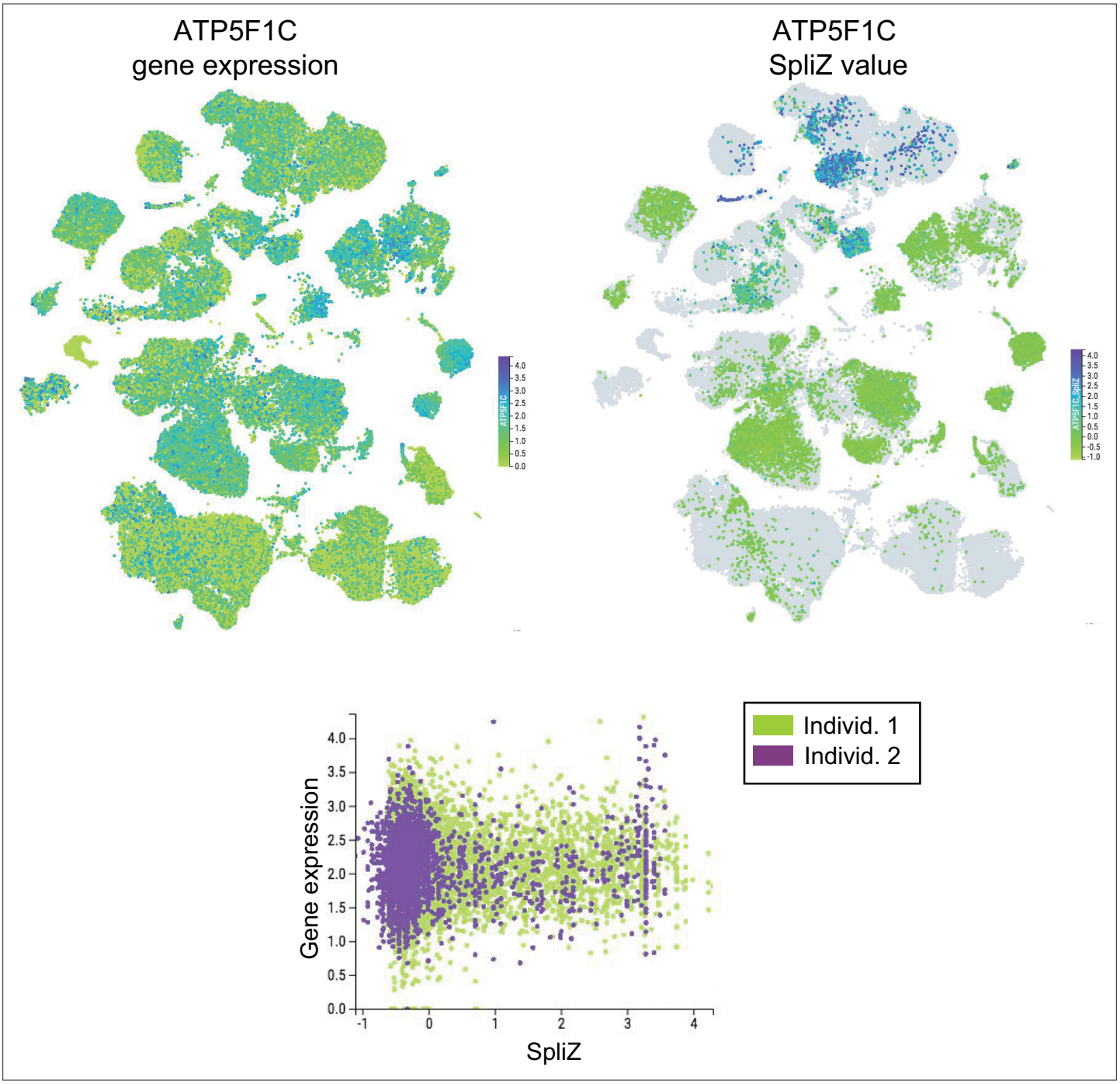


**Figure 2—figure supplement 3.** Gene expression and the SpliZ value of *ATP5F1C* across single cells in the *Tabula Sapiens* dataset. Coloring *Tabula Sapiens* cells by both gene expression and SpliZ value shows that *ATP5F1C* is ubiquitously expressed and that the SpliZ is independent of gene expression for these cases. Plots are obtained by using the cellxgene (***Megill et al., 2021***) visualization platform.

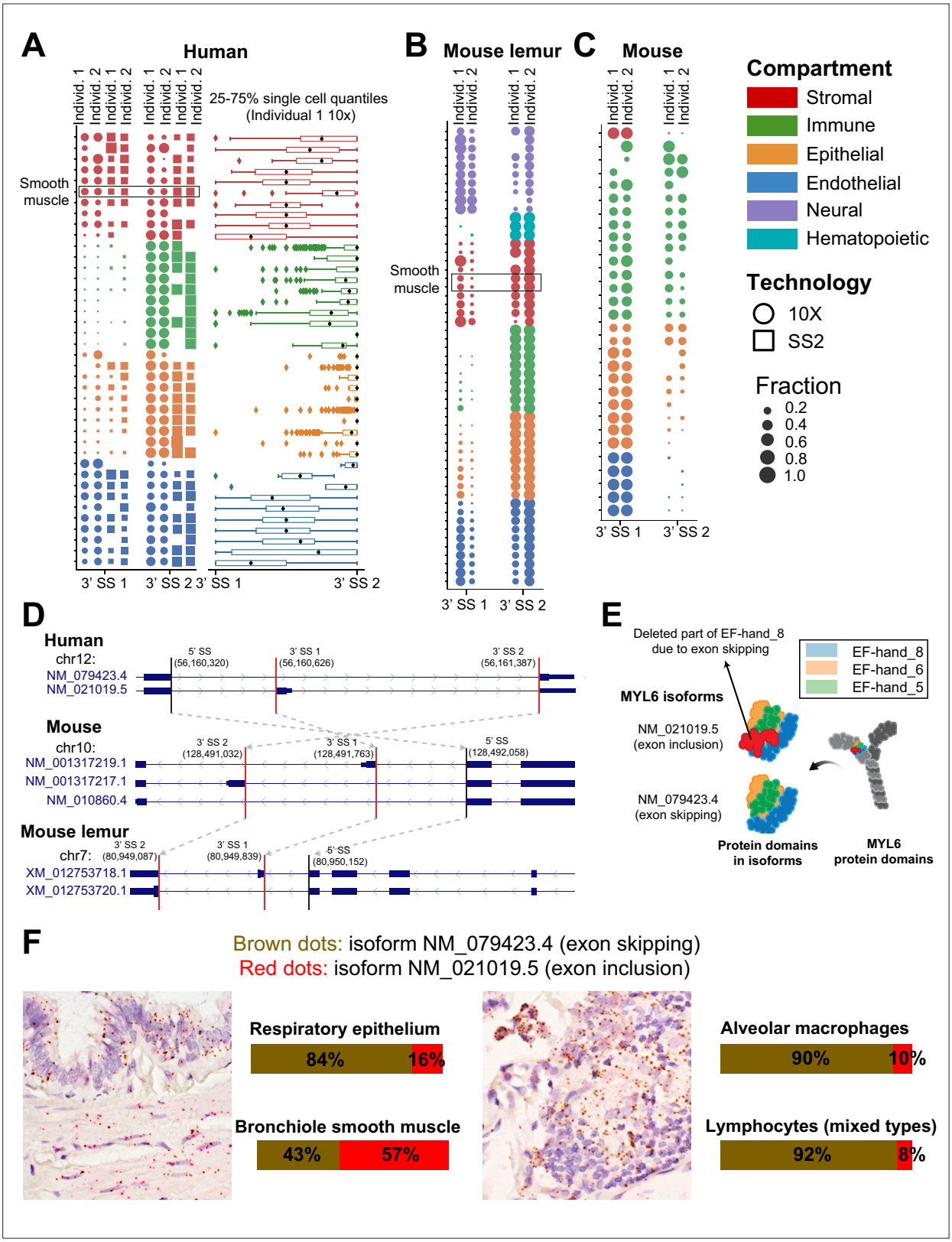


**Figure 3.** Compartment-specific alternative splicing in *MYL6*. Differential alternative splicing between compartments for *MYL6* is driven by an exon skipping event with orthologous splice sites (SS) in (**A**) human (5′ SS: 56,160,320 and two 3′ SSs: 56,161,387 and 56,160,626), (**B**) mouse lemur, and (**C**) mouse. Each dot shows the expression for the splicing to one of the 3′ SSs marked by vertical red lines on the gene annotation in (**D**) in a 10X (circles) or SS2 (squares) dataset from individuals 1 and 2. Columns of dots are biological replicates; for human data, the first two columns are 10X and the second

*Figure 3 continued on next page*

*Figure 3 continued*

two columns are SS2. Dots are colored by compartment. For mouse and mouse lemur, the two columns are 10X samples. The box plot is obtained by assigning 1 and 2 to the closer and farther 3′ SS and then computing their weighted average for each cell according to their corresponding fraction of junctional reads in the cell. Cells in the immune compartment have higher exon skipping rates than cells in the stromal compartment in all three organisms. Smooth muscle cell types are boxed within the stromal compartment. Mouse cells have the same relative proportions of exon inclusion between compartments, but express higher levels of the exon included isoform overall. The SpliZ scores (and also gene expression values) for *MYL6* across all 10X cells in human individual 1 are shown in ***Figure 2—figure supplement 1***. (**D**) Gene structures showing *MYL6* annotation in human, mouse, and mouse lemur. The gray arrows between different organisms show LiftOver mapping between human, mouse, and mouse lemur, indicating that orthologous splice sites are involved in alternative splicing in different organisms. (**E**) Protein domains in MYL6 and how they are organized in the two *MYL6* isoforms. The exon skipping leads to the deletion in the EF_hand_8 domain (shown by the red color). (**F**) RNA FISH validation in human lung: each slide is stained simultaneously with probes in red (specific to exon inclusion) and brown (specific to exon exclusion). As found from the scRNA-seq data, smooth muscle cells have a higher proportion of the included exon than the other compartments and immune cells have the lowest proportion.

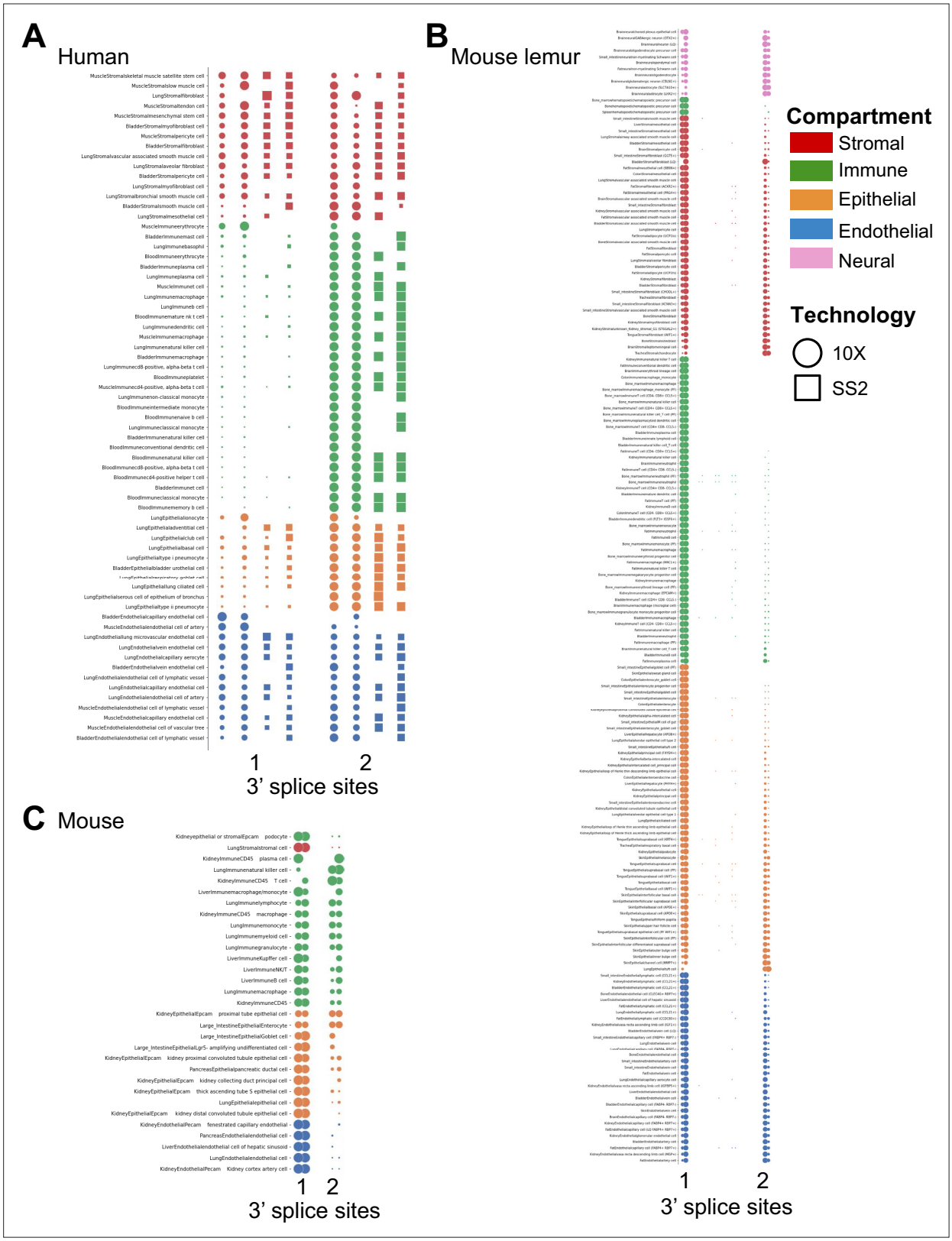


**Figure 3—figure supplement 1.** Uncropped plots of *MYL6* expression. (**A**) Human, (**B**) lemur, and (**C**) mouse for the splice sites shown in ***Figure 3***.

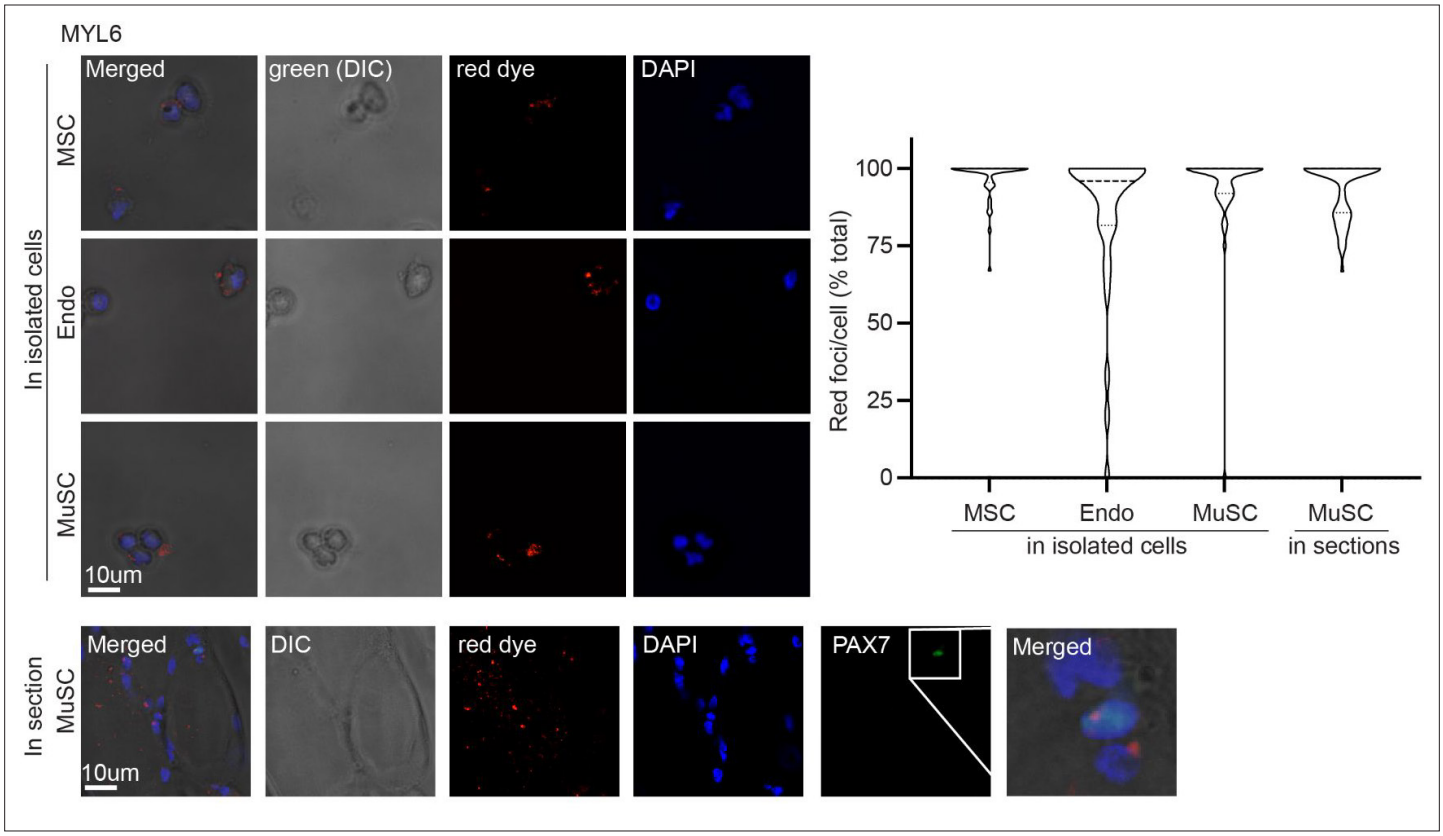


**Figure 3—figure supplement 2.** FISH validation for *MYL6* alternative splicing in cells isolated from human muscle. Indicated cell types were isolated from human muscle and stained by RNA FISH. Example images are shown on the left, with the exon 6 isoform shown in red, the 5–7 isoform shown in gray in the DIC channel, and DAPI shown in blue. Graph depicting the quantifications is shown on the right.

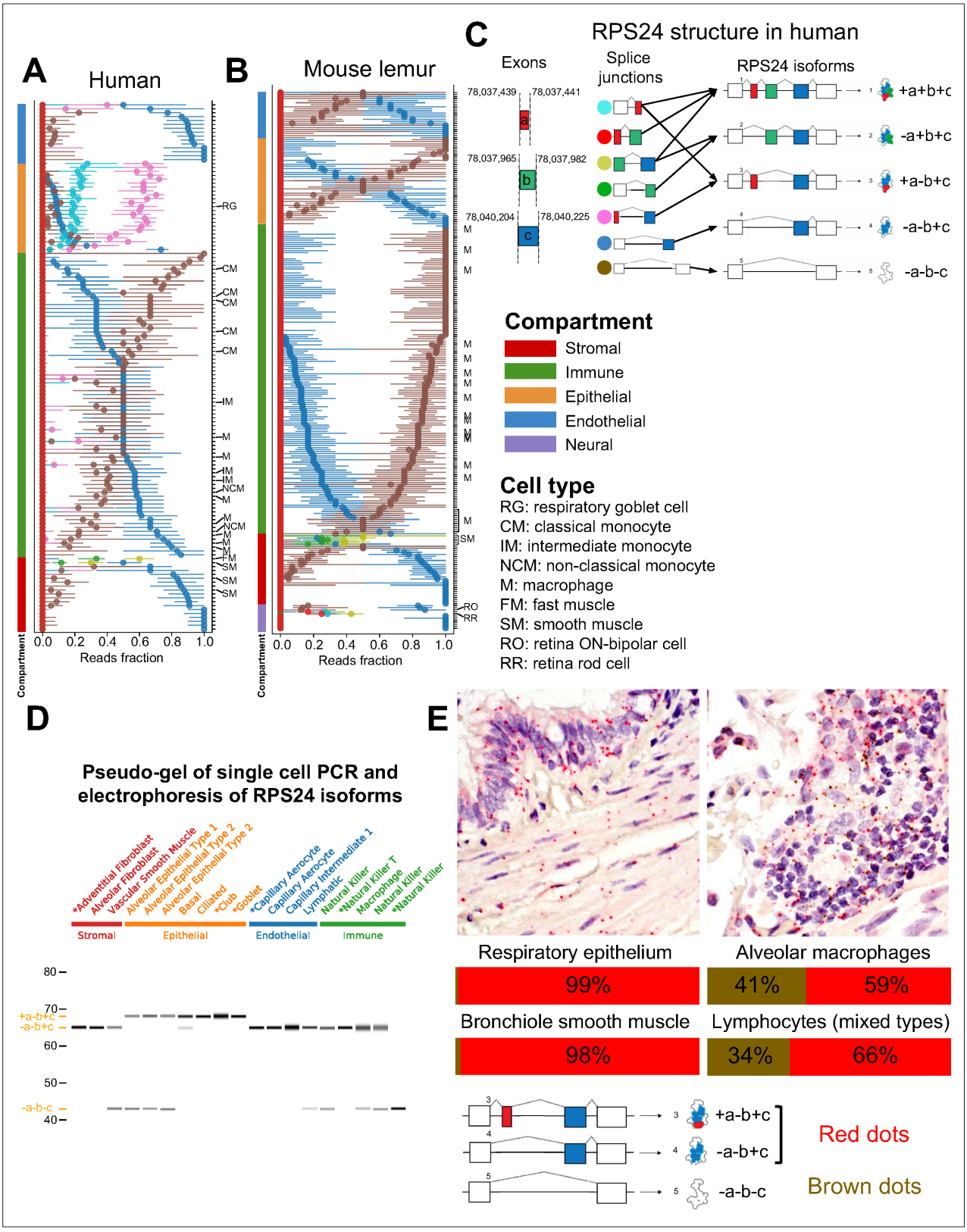


**Figure 4.** RPS24 has striking compartment-specific alternative splicing. (**A**) Each colored circle in the plot represents one *RPS24* junction that uniquely identifies an isoform (junction with green circle represents two isoforms). For each cell type (y axis), the median of all single-cell point estimates of junction fraction in the cell type is plotted on the x axis, with bars representing the 25th and 75th quantiles of single-cell junction fractions. Cell types are sorted by compartment. Black arrows show the splice junctions that can be used for identifying each isoform. Within the immune compartment, the

*Figure 4 continued on next page*

*Figure 4 continued*

fraction of the blue junction increases from classical monocytes to intermediate monocytes to non-classical monocytes. (**B**) The isoform with epithelial-specific splicing in human is not expressed in mouse lemur. However, the same isoform is expressed in smooth muscle as in human. Retinal cells are the only cells to express the *+a+b+c* isoform. (**C**) *RPS24* isoform structure in human shows alternative inclusion of three cassette exons *a*, *b*, and *c* create five annotated isoforms. (**D**) Single-cell PCR validates the prediction that the *+a-b+c* isoform is epithelial-specific. All the epithelial cells contain the isoform with the 3-nt exon *a* , while none of the cells from other compartments do. PCR products from the cells prefixed by asterisks were Sanger-sequenced and matched the expected isoform without evidence of mixture. (**E**) *RPS24* FISH in human lung validates scRNA-seq computational predictions. Slides were simultaneously stained with probes in red and brown, specific for alternative splice junctions. As found from the scRNA-seq data, respiratory epithelium and bronchiole smooth muscle in the epithelial and stromal compartments, respectively, have a low proportion of the *-a-b-c* isoform compared to alveolar macrophages and lymphocytes, both of which are in the immune compartment.

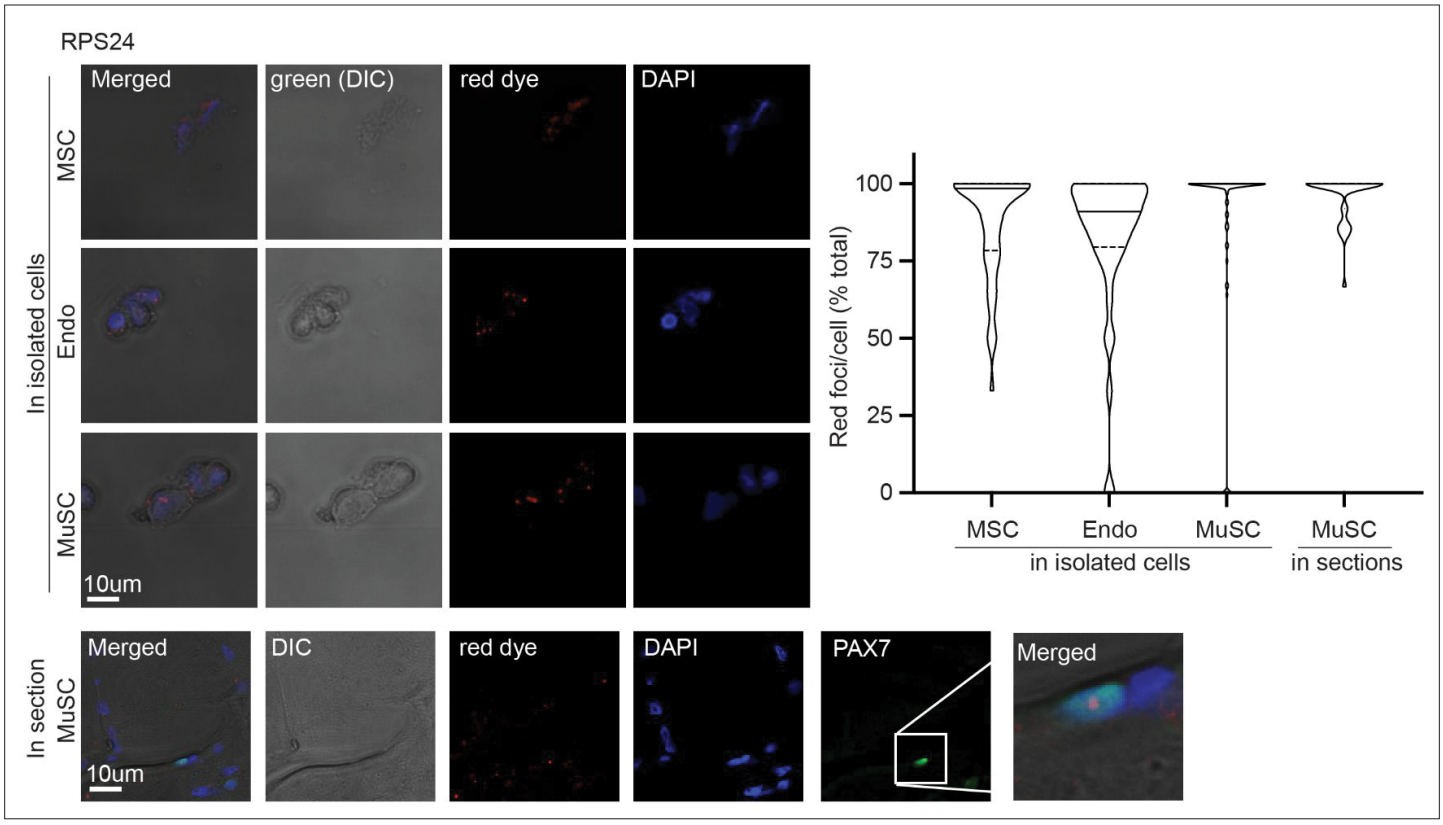


**Figure 4—figure supplement 1.** FISH validation for *RPS24* alternative splicing in cells isolated from human muscle. Indicated cell types were isolated from human muscle and stained by RNA FISH. Example images are shown on the left, with the *-a-b+c* isoform shown in red, the *-a-b-c* isoform shown in gray in the DIC channel, and DAPI shown in blue. Graphs depicting the quantifications are shown on the right. Graph depicting the quantifications is shown on the right.

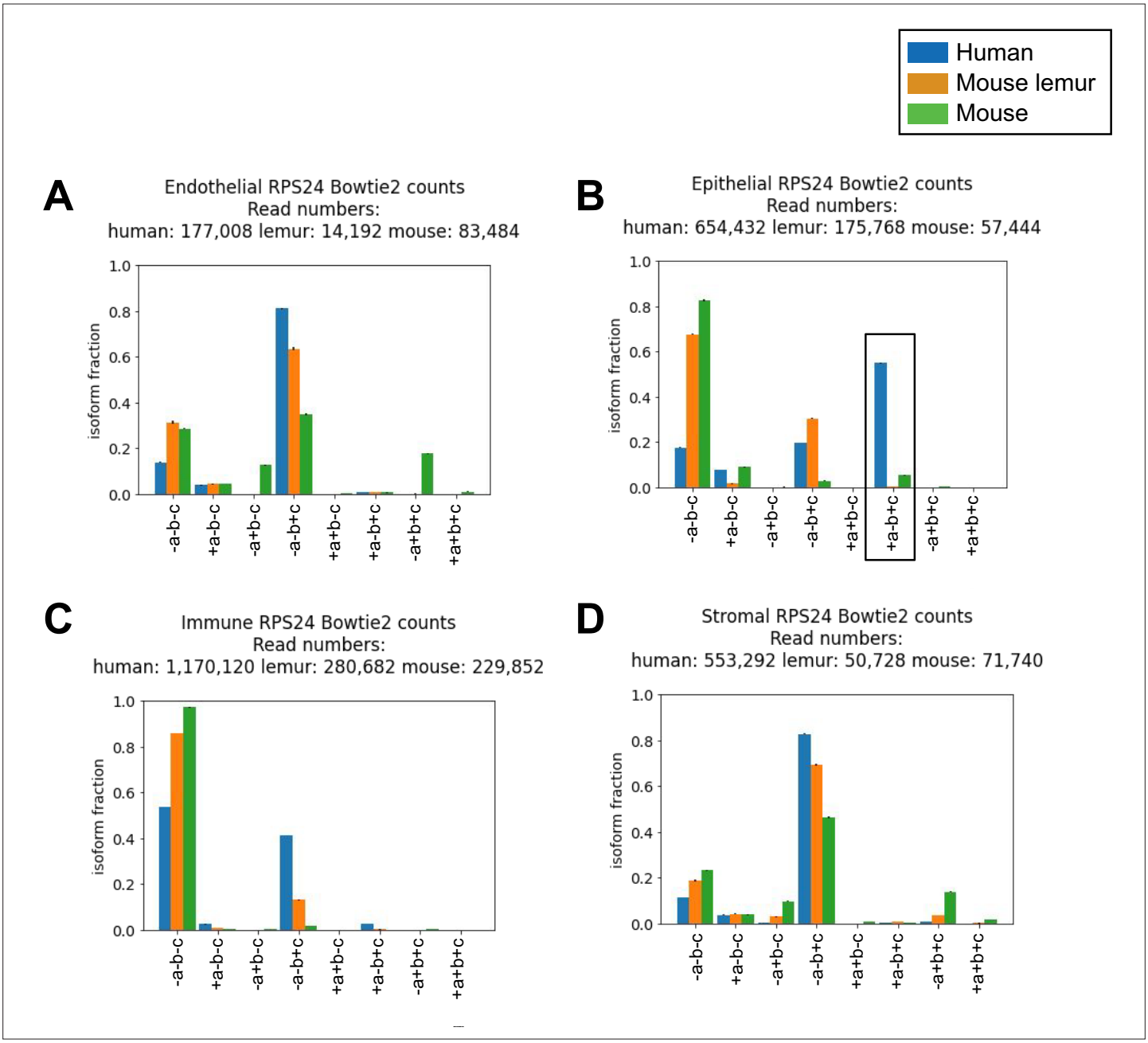


**Figure 4—figure supplement 2.** *RPS24* isoforms quantified by the Bowtie2 aligner validate STAR-based discoveries. Bar plots for the (**A**) endothelial, (**B**) epithelial, (**C**) immune, and (**D**) stromal compartments show proportions of each isoform for each species, with error bars corresponding to 95% binomial confidence intervals. Within epithelial cells the *+a-b+c* isoform was expressed at the highest fraction in human, while only at a small fraction in mouse and mouse lemur, confirming that this isoform's epithelial specificity is not conserved in mouse and mouse lemur.

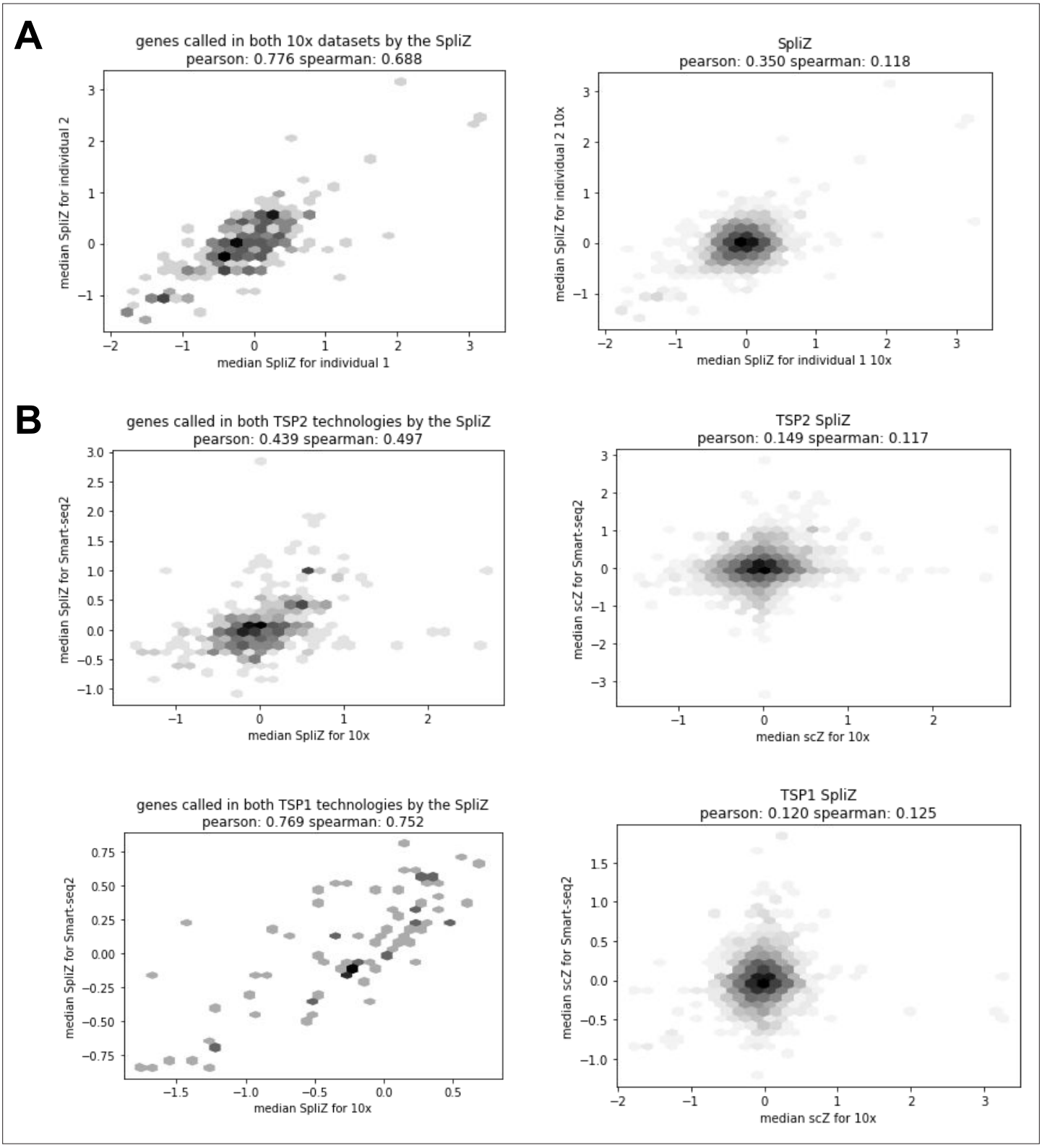


**Figure 5.** Correlations show high concordance of the SpliZ values for significant genes across biological replicates. (**A**) When subsetted to only shared junctions and shared cell types, the SpliZ values for significant genes for both 10X datasets are highly concordant (Pearson correlation of 0.776). (**B**) Comparing datasets from the 10X and SS2 technologies.

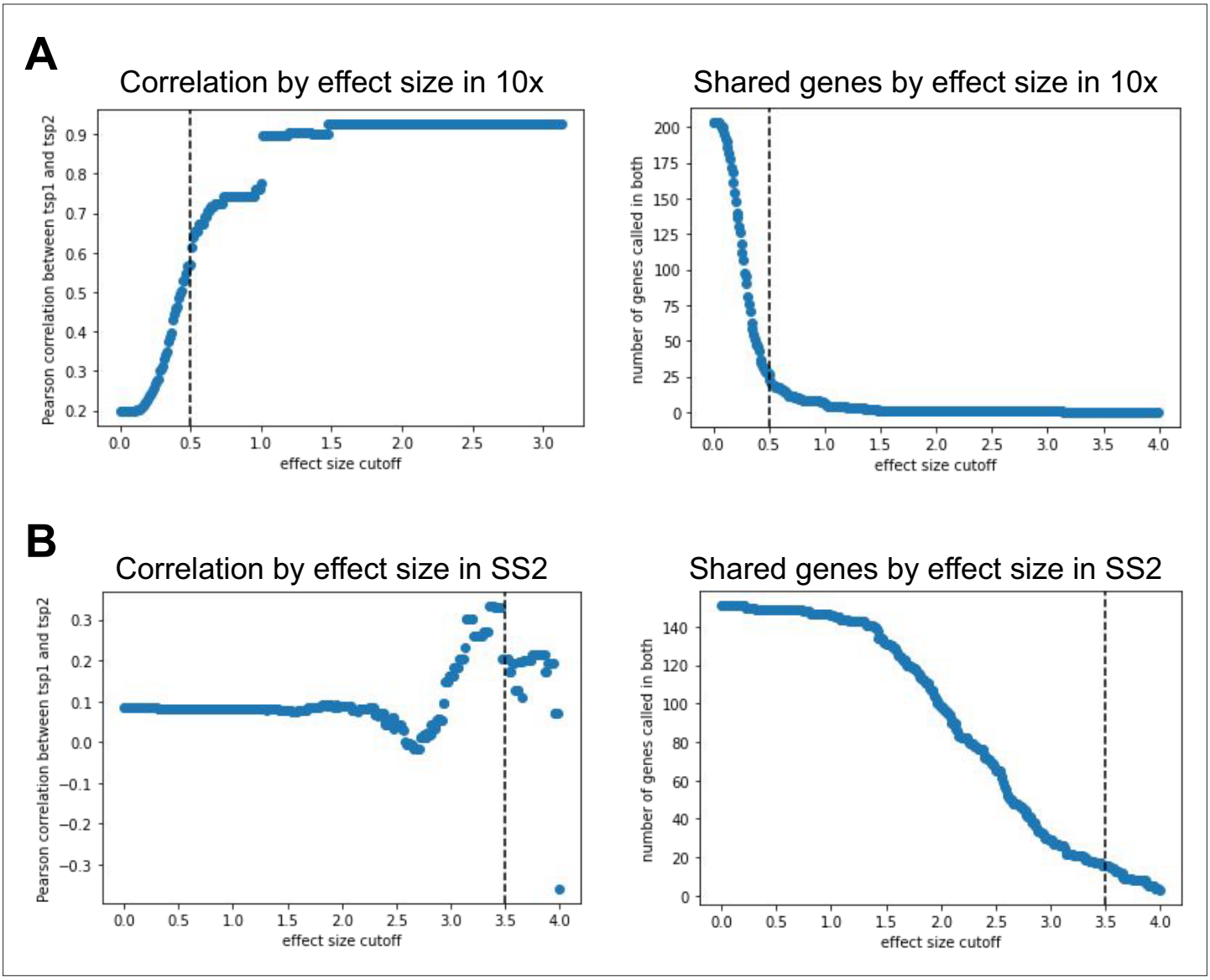


**Figure 5—figure supplement 1.** Choosing effect size filters. Effect size filters were chosen based on correlation analysis of (**A**) Both human 10X datasets and (**B**) separately for SpliZVD by correlation of both human SS2 datasets.

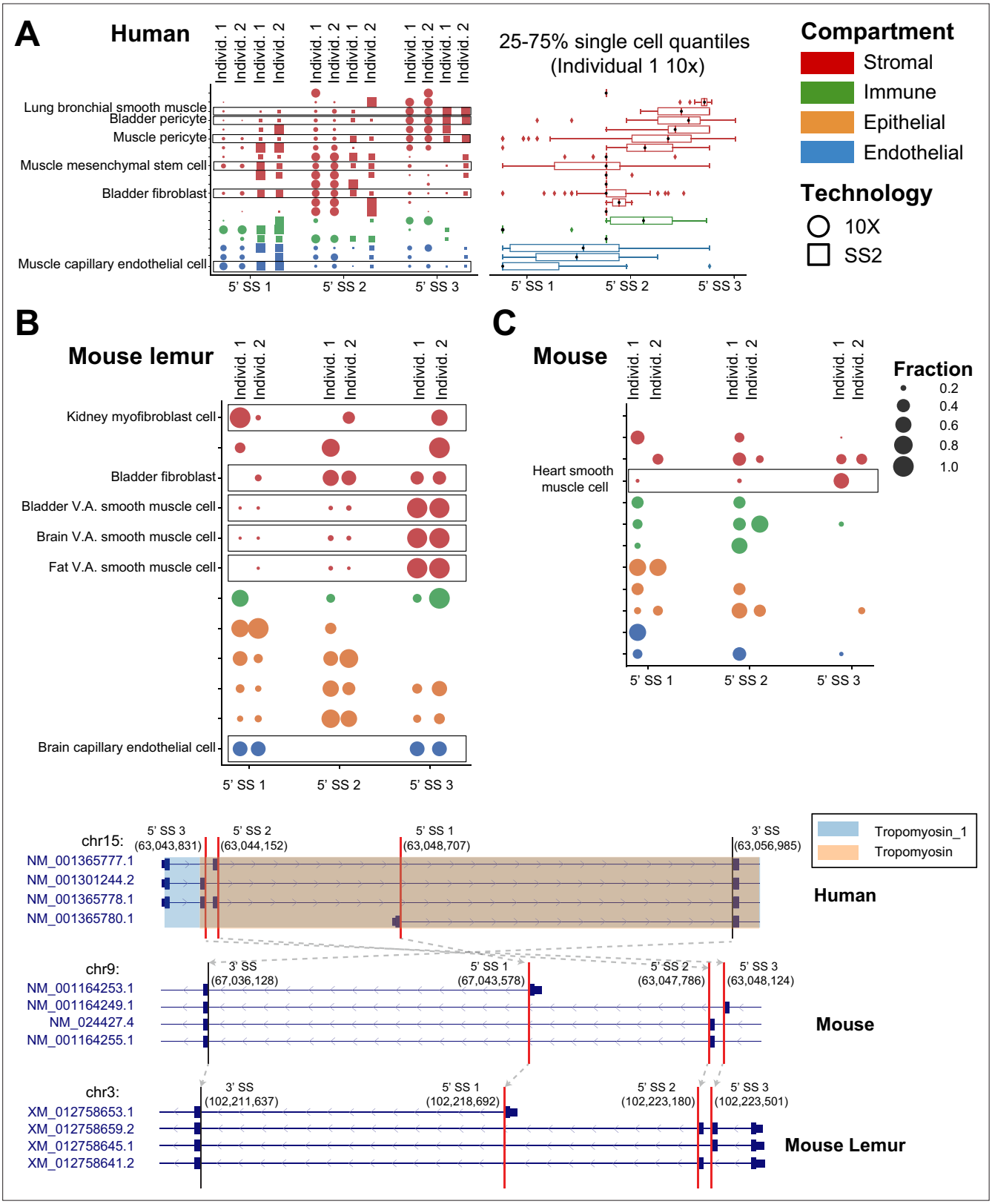


**Figure 6.** Cell-type-specific and conserved alternative splicing in *TPM1*. Conserved splicing in *TPM1* is recovered in (**A**) human, (**B**) mouse lemur, and (**C**) mouse. *TPM1* has a pattern of differential splicing involving two cassette exons and an alternate 5′ end (as shown by the gene structures at the bottom of the figure). Capillary endothelial cells mostly express the isoform with the alternate 5′ end (5′ splice site [5′ SS] 1), while smooth muscle almost exclusively expresses the isoform with the 5′-most domain (boxed in the figure). The box plot shows the distribution of the average 5′ SS (obtained as

*Figure 6 continued on next page*

*Figure 6 continued*

the weighted average of 5′ SS when ranked from 1 to 3 from the closest to the farthest according to their fraction of junctional reads) for the cells within a cell type (see ***Figures 2 and 3*** for more explanation of dot and box plots). There is differential isoform usage within the stromal compartment as well, for example, human bladder stromal fibroblasts and bladder stromal pericytes each express a different dominant cassette exon. Both lemur and mouse similarly express cell-type-specific differences in *TPM1* isoform usage. Orthologous SpliZsites in human, mouse, and mouse lemur are involved in alternative splicing based on the LiftOver mapping, as shown by gray arrows on the gene structures.

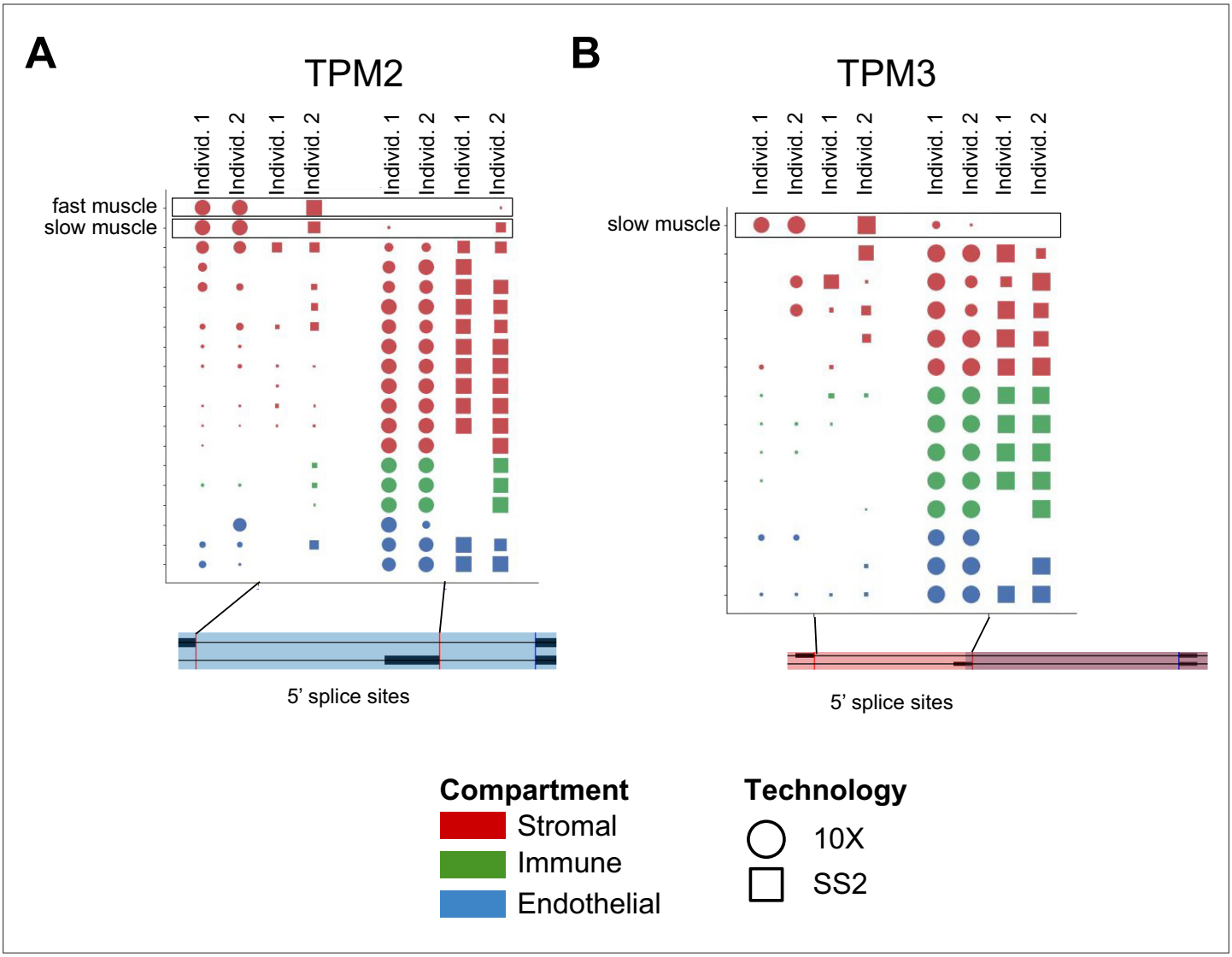


**Figure 6—figure supplement 1.** Cell-type-specific splicing in other members of the tropomyosin family. Both *TPM2* (**A**) and *TPM3* (**B**) exhibit cell-type-specific splicing patterns in human.

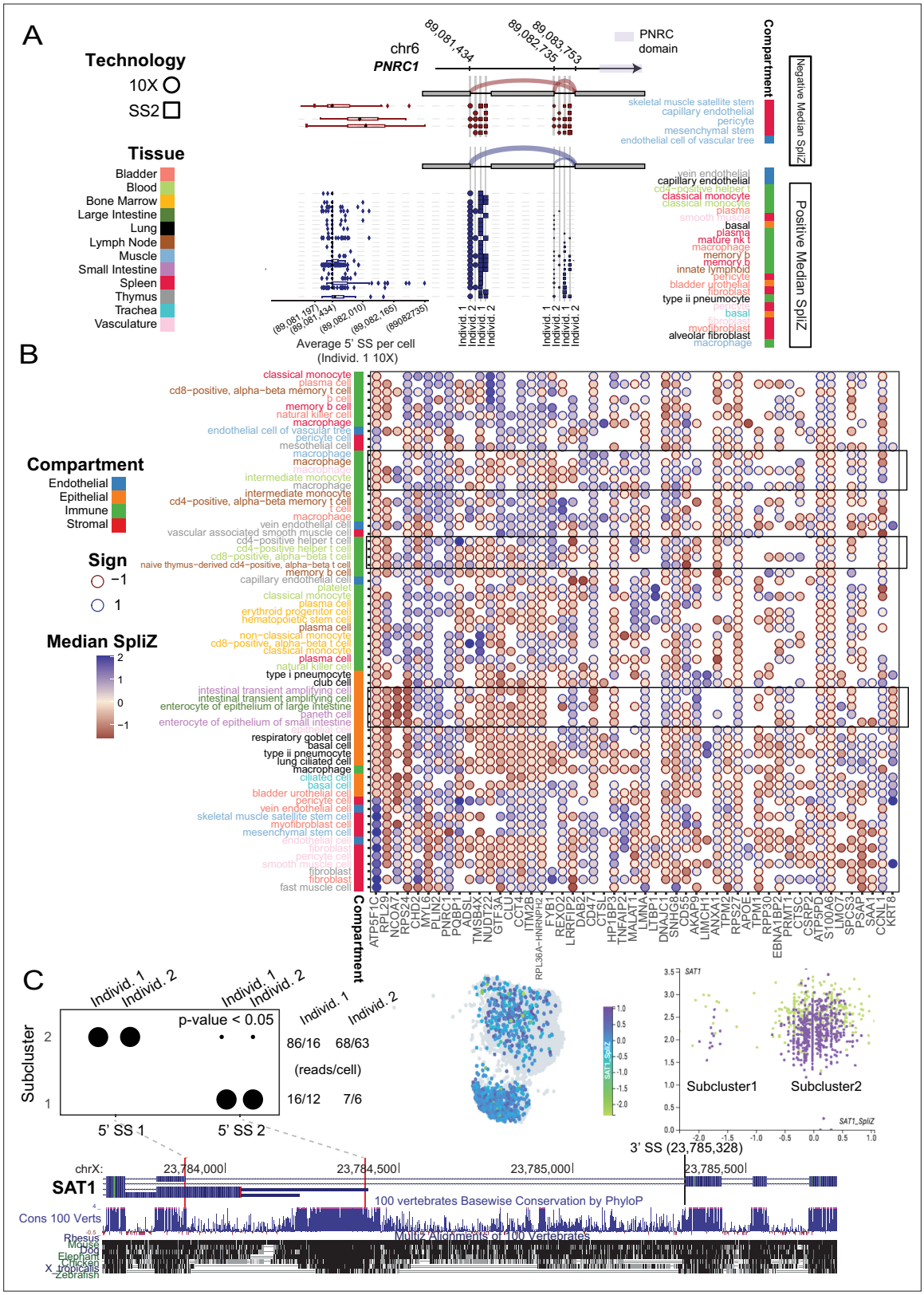


**Figure 7.** Cell-type-specific alternative splicing is prevalent in human genes and can reveal novel cell subpopulations. (**A**) Cell-type-specific exon skipping in *PNRC1* (involving one 3′ SS and two 5′ SS's) is replicated across the four human datasets. Skeletal muscle satellite stem cells include the exon about 50% of the time, whereas vein endothelial cells in the thymus never include the exon. Cell types with negative median SpliZ values are on the top panel, and those with positive median SpliZ values are on the bottom panel. Each of the four human datasets is plotted, with circles representing

*Figure 7 continued on next page*

*Figure 7 continued*

10X data and squares representing SS2 data. The gene annotation is shown above the dots, with sashimi arcs indicating the mean expression of each junction for the given cell types and datasets in the corresponding panel (similar to ***Figure 2***). The known protein domain is marked on the gene structure. Box plots for each cell type are based on weighted 3′ usage (based on the fraction of junctional reads to each 3′ SS) of the 5′ SS for each cell in the cell type in individual 1 10X data (**L**). Each box shows 25–75% quantiles of average 5′ SS per cell. (**B**) Unsupervised clustering analysis with the SpliZ identified clusters of cell types and compartments independent of tissue. Dots show the median SpliZ (effect size) for genes found to be significantly regulated across cell types. Only 50 significant genes with the highest effect size and cell types with >25 significant genes are shown. Hierarchical clustering was performed on both genes and cell types based on median SpliZ values. Cell type names are color-coded based on their tissue (same tissue colors as in **A**) and the side bar shows the compartment for each cell type. (**C**) Alternative splicing of gene *SAT1* distinguishes two populations of cells within blood classical monocytes and involves an ultraconserved exon. The dot plot shows the differential inclusion of the 5′ SSs for the 3′ SS at 23,785,300 for cells grouped based on their assigned subclusters. The number of reads (X) and cells (Y) containing the splice junctions involving the 3′ SS in each individual are shown at right. Clustering based on gene expression as shown by cellxgene visualization (middle panel) and scatter plot (right panel) does not distinguish cell populations with distinct splice profiles. In the scatter plot, the x and y axes represent the gene expression and SpliZ values for *SAT1* in each cell, respectively. Cells are colored according to their human individual number. Visualization of the gene expression value for *SAT1* does not distinguish the populations; both subclusters contain classical monocytes from both human individuals (right scatter plot).

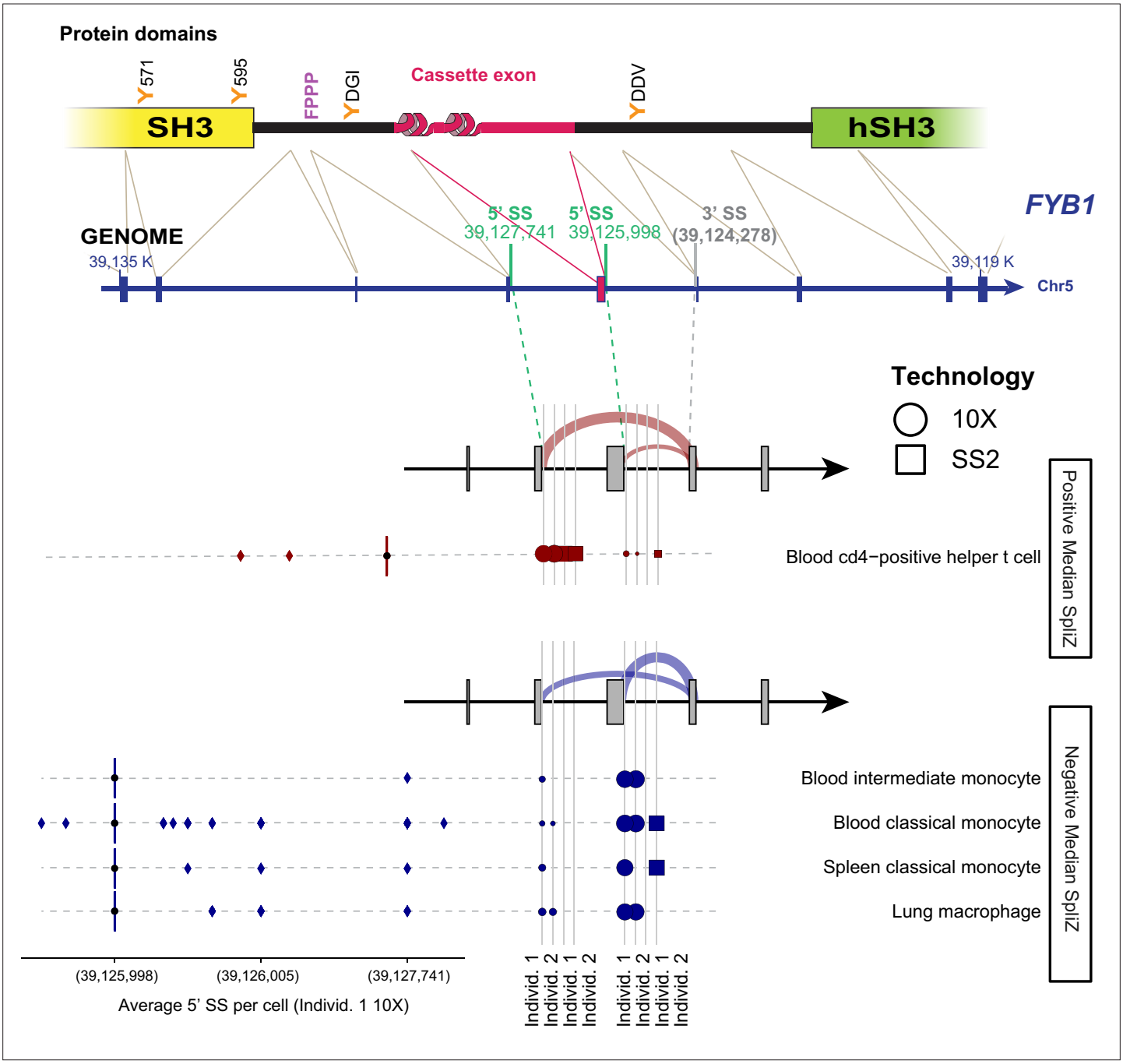


**Figure 7—figure supplement 1.** Differential alternative splicing of *FYB1* within the immune compartment. The pattern is replicated in 10X and SS2 data from both human individuals and technologies, with monocytes and macrophages tending to include an exon (at position 39125998) that is usually skipped in T cells. (See ***Figure 3C*** for top-panel plot explanation.) At the bottom are diagrams of *FYB1* genomic organization and protein features. The exon structure is completely conserved between mouse and human, including the cassette exon; the conserved mapping of exons to protein sequence is shown for the C-terminal portion. Most of the protein is predicted to be unstructured, except for the distal SH3 and hSH3 domains. Numerous tyrosine (Y) phosphorylation sites are conserved and annotated in the expanded view of the protein. The cassette exon contains two predicted short alpha-helical stretches; it may allow additional protein interactions and/or may modulate accessibility of the two important phosphorylation sites that flank it.

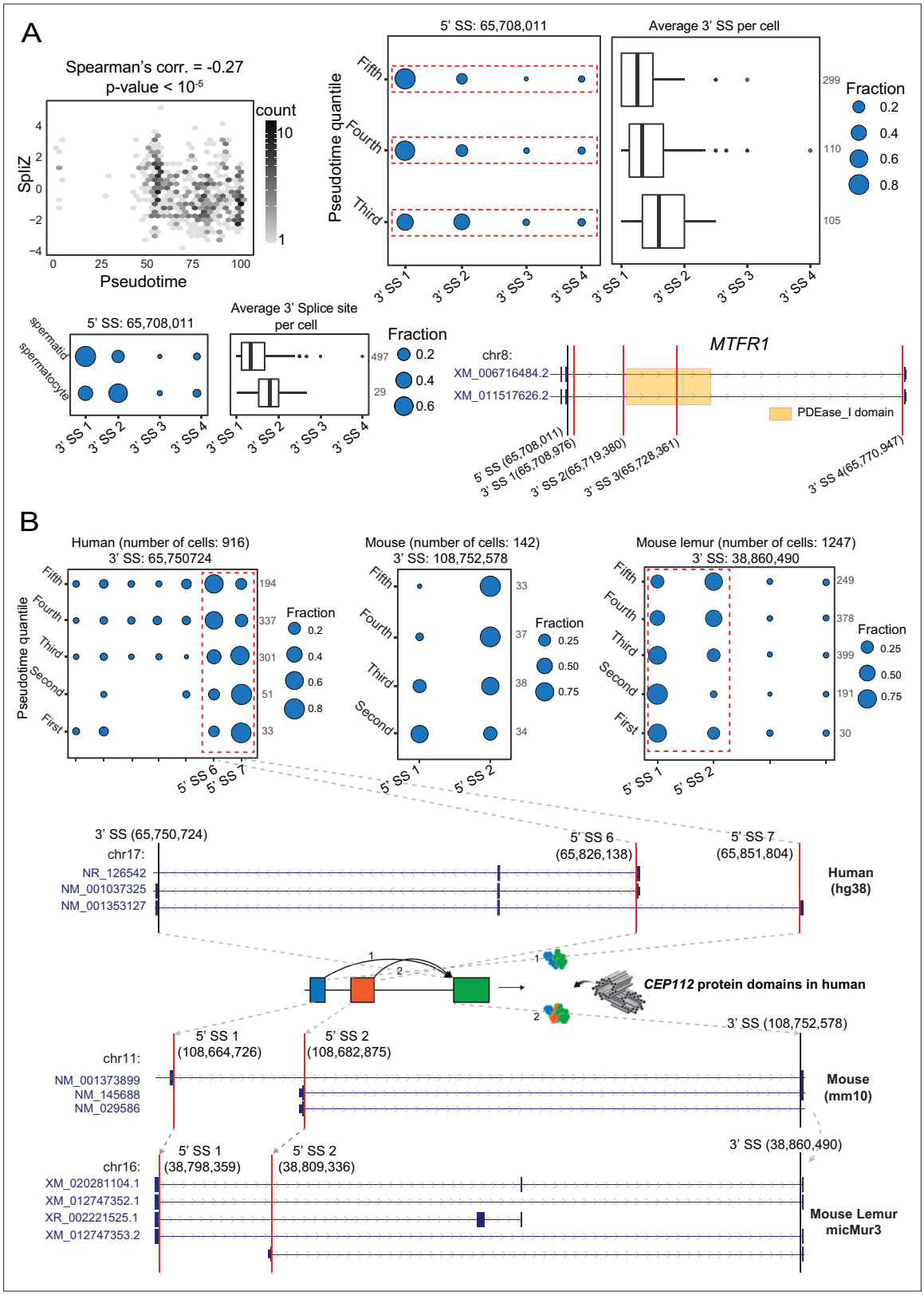


**Figure 8.** Developmentally regulated alternative splicing in mammalian spermatogenesis. (**A**) Regulated alternative splicing of *MTFR1* during sperm development. Significant negative correlation (Spearman's correlation = –0.27, p-value = 1.23e-7) between the SpliZ score for gene *MTFR1* and pseudotime in human sperm cells (top left). Dot plot and box plot show increasing use of a downstream 3′ splice site driving the *MTFR1* SpliZ in equal pseudotime quantiles (top right) with the same trend in immature (spermatocyte) and mature (spermatid) cells (bottom left). The gene structure for

*Figure 8 continued on next page*

*Figure 8 continued*

*MTFR1* according to the human RefSeq annotation database is shown in the bottom-right panel. The orange box on the gene structure represents the PDEase_I domain and how it is affected by the alternative splicing. (**B**) Dot plots showing the developmentally regulated alternative splicing of gene *CEP112* in testis cells from human, mouse, and mouse lemur. Cells are grouped according to pseudotime quantiles. The alternative splicing is conserved (i.e., involves the same set of 5′ and 3′ splice sites in human, mouse, and mouse lemur data as shown by the gray arrows on the gene structures) and involves 5′ splice sites 65,826,138 and 65,851,804 in human, 5′ splice sites 108,664,726 and 108,682,875 in mouse, and 5′ splice sites 38,798,359 and 38,809,336 in mouse lemur. The 3′ splice site and the two 5′ splice sites involved in alternative splicing are shown by black and red vertical lines, respectively, on the gene structures. *CEP112* is on the minus strand in the human genome but is on the plus strand in mouse and mouse lemur genomes, leading to a negative correlation in splice site usage. Gray arrows show the LiftOver mapping between the 3′ splice site and two 5′ splice sites of the exon skipping event (indicating that the alternative splicing is conserved) and gray dashed lines for the human plot show the location of the 5′ splice sites and how splicing changes the apolipoprotein protein domain.

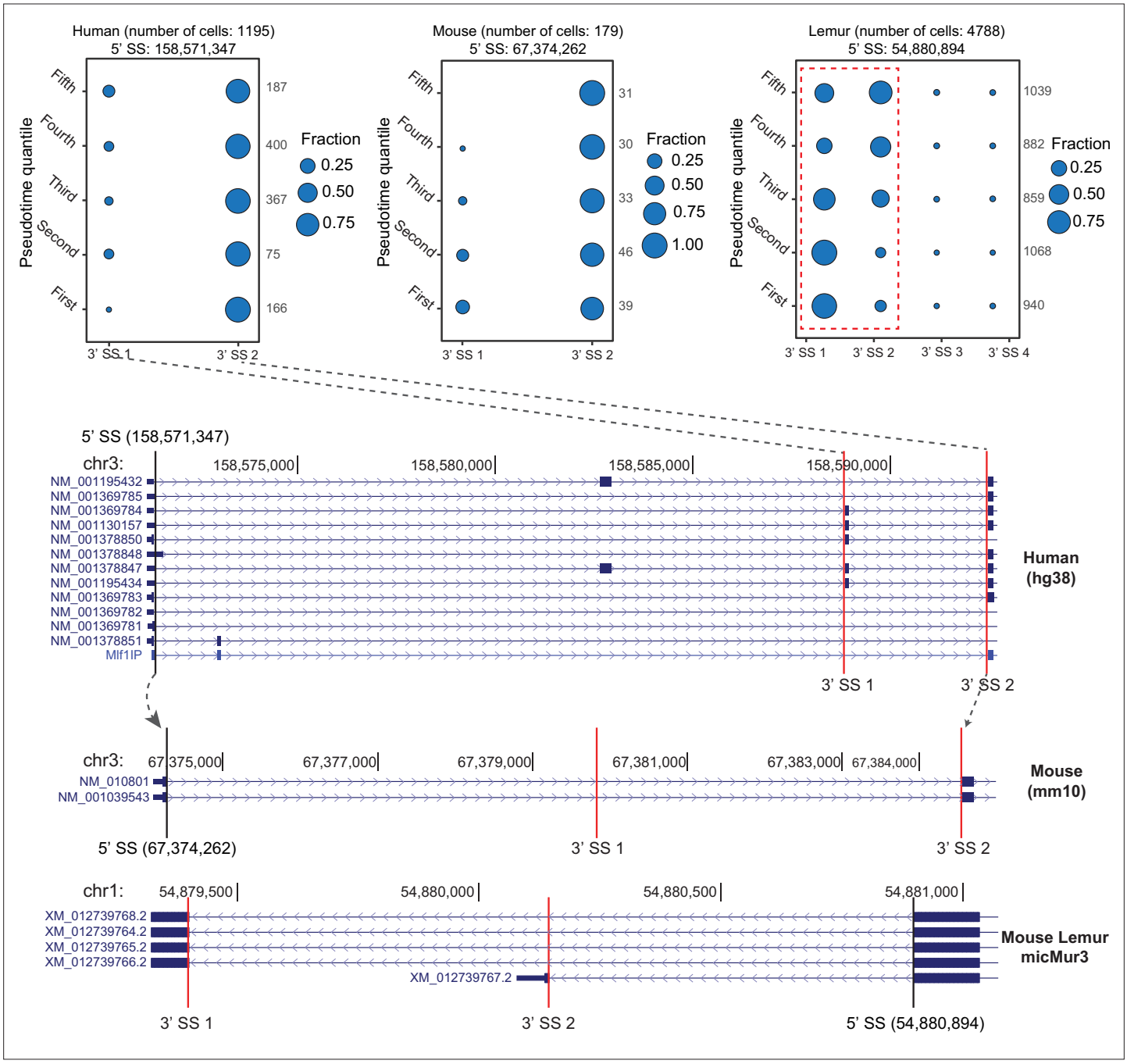


**Figure 8—figure supplement 1.** Regulated alternative splicing of *MLF1* during sperm development in human, mouse, and mouse lemur. The same 5' splice site drives the alternative splicing in human and mouse. The gene structures for *MLF1* according to the RefSeq database for human, mouse, and mouse lemur are shown.

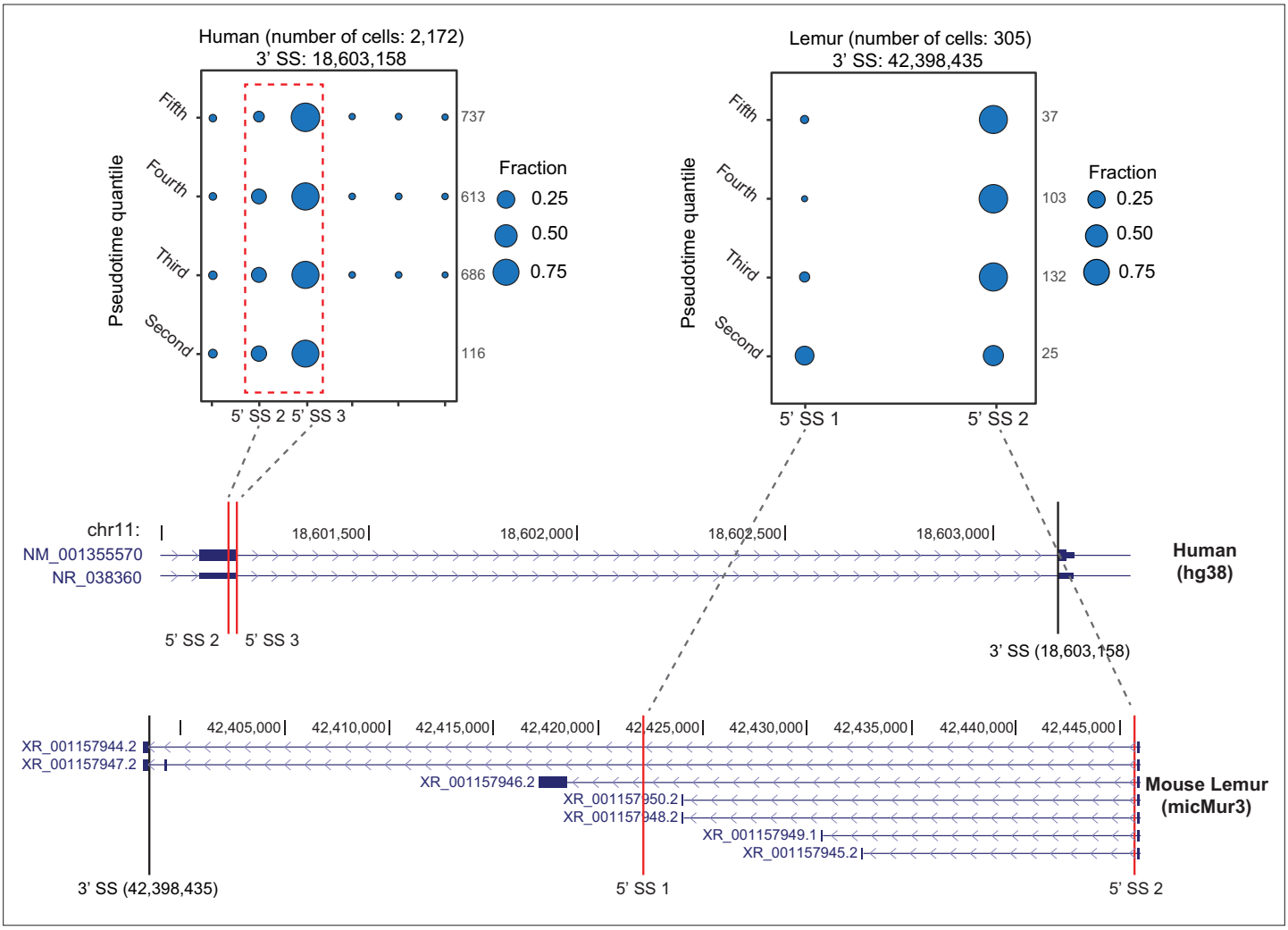


**Figure 8—figure supplement 2.** Regulated alternative splicing of *SPTY2D1OS* during sperm development in human and mouse lemur. Regulated splicing in both human and mouse lemur involves one unannotated 5′ splice site. The gene structures according to the RefSeq database for human and mouse lemur are shown.