Unsupervised reference-free inference reveals unrecognized regulated transcriptomic complexity in human single cells

Roozbeh Dehghannasiri[1,2,*], George Henderson[1,*], Rob Bierman[2,1], Kaitlin Chaung[1], Tavor Baharav[3], Peter Wang[1], Julia Salzman[1,2,4,†]

Author affiliation
[1]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305
[2]Department of Biochemistry, Stanford University, Stanford, CA 94305
[3]Department of Electrical Engineering, Stanford University, Stanford, CA 94305
[4]Department of Statistics and of Biology (by Courtesy), Stanford University, Stanford, CA 94305

[†]Corresponding author: julia.salzman@stanford.edu
*These authors contributed equally to this work.

Abstract

Myriad mechanisms diversify the sequence content of eukaryotic transcripts at both the DNA and RNA levels, leading to profound functional consequences. Examples of this diversity include RNA splicing and V(D)J recombination. Currently, these mechanisms are detected using fragmented bioinformatic tools that require predefining a form of transcript diversification and rely on alignment to an incomplete reference genome, filtering out unaligned sequences, potentially crucial for novel discoveries. Here, we develop SPLASH+, a new analytic method that performs unified, reference-free statistical inference directly on raw sequencing reads. By integrating a micro-assembly and biological interpretation framework with the recently developed SPLASH algorithm, SPLASH+ discovers broad and novel examples of transcript diversification in single cells *de novo*, without the need for genome alignment and cell type metadata, which is impossible with current algorithms. Applied to 10,326 primary human single cells across 19 tissues profiled with SmartSeq2, SPLASH+ discovers a set of splicing and histone regulators with highly conserved intronic regions that are themselves subject to targets of complex splicing regulation. Additionally, it reveals unreported transcript diversity in the heat shock protein *HSP90AA1*, as well as diversification in centromeric RNA expression, V(D)J recombination, RNA editing, and repeat expansion, all missed by existing methods. SPLASH+ is unbiased and highly efficient, enabling the discovery of an unprecedented breadth of RNA regulation and diversification in single cells through a new paradigm of transcriptomic analysis.

Introduction

The diversity of transcripts in eukaryotes results from mechanisms such as alternative splicing, RNA editing, and alternative 5' and 3' UTR use as well as genetic changes in single cells including insertion of mobile elements, repeat expansions, or segmental duplications. The genome's potential to create more than 10^{13} genetic variants through V(D)J recombination (Schroeder 2006) in the adaptive immune system is imperative for determining the specificity and efficacy of defense mechanisms against pathogens. Transcript diversification can have significant functional consequences, including causal links to various diseases from cancer to neurodegeneration (Kung, Maggi, and Weber 2018; Yum, Wang, and Kalsotra 2017; Bonnal, López-Oreja, and Valcárcel 2020; Ma et al. 2021)

Despite its importance to cell specialization, the extent to which transcript diversity is regulated in single cells remains a significant open question. Current computational approaches to detect transcript diversity in single cells are specialized for certain events such as alternative splicing (Olivieri, Dehghannasiri, and Salzman 2022; Buen Abad Najar et al. 2022) or V(D)J recombination reconstruction (Lindeman et al. 2018) and also rely heavily on references, beginning with alignment of reads to a reference genome, thus censoring unmapped reads and introducing mapping biases. Quantifying diversity in single cells is also hindered by low read counts and dropouts (Westoby et al. 2020). Other tasks, such as identifying somatically acquired repeats or RNA

editing are not even attempted in scRNA-seq. Further, it is unclear whether these methods are sensitive to all events they aim to detect. For example, they may fail to align highly edited or spliced transcripts (Eisenberg and Levanon 2018) and most significantly, they are incapable of detecting sequences that are absent from a reference genome.

Statistics is at the core of inference for single-cell genomics. Yet, current genomic inference is conditional on the results of partially heuristic alignment algorithms, which often discard reads that do not map to the reference genome (Figure 1A). Moreover, reference-first approaches heavily favor sequences from European ancestries, further adding to the bias (Sherman et al. 2019). The statistical tests downstream of alignment are typically parametric or require randomized resampling, leading to inaccurate or inefficient p-values. De novo assembly approaches (Cmero et al. 2021; Swanson et al. 2013), while attempting to create more accurate references, also introduce fundamental biases and unknown false positive and negative rates as shown by previous studies (Freedman, Clamp, and Sackton 2021). They also face computational biases from the aligner as their downstream inference still depends on alignment information to the de novo assembly. Together, there is a strong argument to bypass reference alignment to find regulated sequence diversity through an unbiased and unified framework.

To overcome these challenges, we recently introduced SPLASH (Chaung et al. 2023), a unified reference-free algorithm that performs statistical inference directly on raw sequencing reads to detect a wide range of biological processes that diversify transcripts, including but not limited to RNA splicing, mutations, RNA editing, and V(D)J recombination (Figure 1A). Additionally, SPLASH can identify variation in repetitive regions of the genome, e.g., in multicopy non-coding RNA loci or centromeres that are difficult to map due to long arrays of near-identical repeats.

SPLASH's core includes a novel statistical test to detect sample-specific sequence variation, filling a gap in existing methods for transcriptome analysis. Classical and parametric tests struggle to prioritize biologically important variation because they are overpowered in the context of noise generated from biochemical sampling, and may report inaccurate p-values. SPLASH's test provides finite-sample valid p-value bounds, unlike Pearson's chi-squared test, better controls false positive calls under commonly used models such as negative binomial for scRNA-seq (Buen Abad Najar, Yosef, and Lareau 2020; Baharav, Tse, and Salzman 2023), and performs inference independent of any cell metadata such as cell type.

In this manuscript, we build on this core to introduce SPLASH+, which includes new approaches to analyze SPLASH's output, including a new, reference-free statistical approach for de novo assembly as well as a framework to interpret and annotate its results (Figures 1B,C). We should note that by "reference-free" we mean that statistical inference for detecting sequences with sample-specific variation is completely independent of assembly and the potential biases that could be introduced through alignment. Here, alignment to the reference genome is used only after inference of biological sequence variation, and is done to annotate and interpret such sequences.

To systematically analyze transcript diversity in human cells, we applied SPLASH+ to 10,326 human SmartSeq2 (SS2) cells across 346 cell types and 12 donors from *Tabula Sapiens* dataset (Suppl. Figure 1) (Tabula Sapiens Consortium* et al. 2022). SPLASH+ reveals new insights into the regulation of transcript diversification in human single cells, including features of RNA splicing, RNA editing, and non-coding RNA expression missed by specialized, domain-specific bioinformatic pipelines. Novel findings include (i) regulated expression of repetitive loci such as *RNU6* variants and higher order repeats in centromeres; (ii) complex splicing programs, including un-annotated variants in genes like *CD47*, a major cancer immunotherapy target, and *RPS24*; (iii) pan-tissue regulation of splicing in splicing factors, histone modifications, and in the heat shock protein *HSP90AA1*; (iv) de novo rediscovery of immunoglobulin loci as the most transcriptionally diverse human loci with higher sensitivity compared to the state-of-the-art for V(D)J rearrangement detection; and (v) the detection of single cells with transcribed repeat expansion and high levels of RNA editing. SPLASH+ makes unbiased discoveries by not relying on cell metadata or reference genomes. This is particularly important for scRNA-seq analysis, as cell types can be difficult to generate and remain imprecise (Zeng 2022) and can miss important variation within cell types, such as B cell receptor variation (Watson and Breden 2012).

Our results demonstrate that SPLASH+ is a single algorithm that could replace numerous custom bioinformatic approaches for detecting different types of RNA variation and significantly expands our understanding of the transcriptome's diversity.

An integrated, reference free pipeline to discover regulated RNA expression

SPLASH is an unbiased reference-free algorithm that operates directly on raw sequencing data to identify differentially diversified sequences (Chaung et al. 2023)(Figure 1A), characteristic of important transcript regulation such as splicing, RNA editing, and V(D)J recombination. SPLASH parses reads from all input samples and extracts specific *k*-mers (substrings of length *k*) called anchors. Each anchor is followed by a set of downstream diverse *k*-mers, called targets (Methods, Figure 1A). SPLASH then performs a statistical test for each anchor under the null hypothesis that target count distribution is the same across all samples, providing a closed-form p-value for each anchor. Anchors with significant p-value have sample-specific target expression, reflecting inter-cell expression variation. This variation can be a signature of alternative RNA splicing, RNA editing, among others.

Both anchors and targets (27-mers by default), are too short for biological interpretation of the underlying mechanisms generating sequence diversity which could be complex or span long loci (e.g., V(D)J rearrangement). To address this, we developed SPLASH+ by integrating SPLASH with a new seed-based statistical de novo assembly called *compactors* to establish a longer sequence context for the short anchors called by SPLASH (Figure 1B, Methods). Compactors distinguish sequencing errors from inherent biological sequence variations, offering a straightforward and interpretable framework for dissecting transcriptome complexity and deriving meaningful biological insights.

Significant anchors called by SPLASH are used as seeds in the compactors step. First, FASTQ files are parsed for each called anchor and reads containing the anchors are collected (Figure 1B). Using the collected reads for each anchor, a recursive branching step is performed. At each position downstream of the anchor, where multiple nucleotides are observed across reads, the assembly employs a statistical criterion. It branches into a separate sequence for each nucleotide whose frequency exceeds the threshold determined by the sequencing error assumtion (Figure 1B; Methods). Our statistical analysis has shown that generating a new branch due to sequencing error is highly unlikely, when applying a threshold of >10% (default threshold) for nucleotide frequency (Methods). This rule is applied recursively, resulting in extended sequences called "compactors" for each SPLASH-called anchor.

Compactors denoise input reads, enabling the differentiation between different types of transcript diversification, such as splice isoforms, RNA editing, and V(D)J recombination. To our knowledge, unlike any other de novo transcript assembly used for scRNA-seq, compactors can be statistically characterized to quantify the probability of generating an artifactual compactor due to sequencing error (Methods). Furthermore, compactors reduce the computational burden of downstream analysis, such as genome alignment. For example, in this study, compactors reduced the number of sequences 120 fold: from 183,471,175 raw reads to 1,515,555 compactors 57.8% (1339/2315). We tested if compactors could precisely reconstruct gene fusions using 5 simulated datasets from a benchmarking study (Haas et al. 2019). We took the 27-mer immediately upstream of each fusion breakpoint as seeds and generated compactors (Methods). We then tested if compactor sequences showed evidence for a fusion based on whether the compactor contained the fusion breakpoint. Compactors identified evidence for 57.8% (1339/2315) of total fusions (Suppl. Figure 2A; Suppl. Table 1). Compactor sensitivity surpassed that of two de novo assembly fusion detection methods (JAFFA-Assembly and TrinityFusion-D) in (Haas et al. 2019) and was comparable to the other two methods (TrinityFusion-C and TrinityFusion-UC) that exclusively utilized chimeric and unaligned reads biased for fusion transcripts and are highly computationally intensive (Figure S3 panel A in (Haas et al. 2019)). This result supports the precision and generality of the use of compactors, which are designed to detect general classes of sequence variation including but far from limited to fusion detection. Notably, compactors detected the majority of sufficiently-expressed fusions as 98% of undetected fusions had TPM < 100 and only 2% of 965 fusions with TPM > 100 were undetected (Suppl. Figure 2B).

After generating compactors, SPLASH+ classifies them into biologically-meaningful categories (Methods, Figure 1C, Suppl. Figure 3), enhancing interpretability of SPLASH calls and facilitating targeted or integrative downstream analysis on the anchors within and across categories. Each anchor is classified into one of 6 different categories (Methods, Suppl. Figure 3): splicing, internal splicing, base pair change, 3'UTR, centromere, and repeat. The classification is based on string metrics (Hamming and Levenshtein distances) between the top two compactors, as well as their alignment to the T2T genome using a spliced aligner (e.g., STAR (Dobin et al. 2013)). An anchor is classified as splicing if STAR reports spliced alignment for at least one of the two compactors (Methods). When both compactors lack splice junctions, the anchor is classified as internal splicing or base pair change based on the string distances between the compactors (Methods). The mapping positions for the remaining unclassified anchors are further intersected with annotation databases for 3' UTRs, centromeric repeats, and repeats to classify them into one of these categories accordingly. Through our classification, each compactor, even if it fails to map, can still be *annotated-by-association* with the annotation of the most abundant compactor for the anchor. By using alignment and references only after statistical inference for interpretability, SPLASH obtains statistically-valid and unbiased inference. SPLASH+ facilitates the interpretability of these findings, allowing for a direct comparison of the improvements it offers over existing algorithms.

SPLASH+ detects transcript diversity in repetitive RNA loci including centromeres and U6 spliceosomal RNA

We ran SPLASH+ on 10,326 cells profiled with SmartSeq2 from 19 tissues and 12 donors (29 donor-tissue pairs) and 346 cell types from the Tabula Sapiens Dataset (Tabula Sapiens Consortium* et al. 2022) (Suppl. Figure 1), including 10 tissues (e.g., blood, muscle, lung) with at least two donors, allowing us to analyze reproducibility, as each donor-tissue was run separately. SPLASH is considerably more efficient than other approaches as it utilizes k-mers instead of reference alignment and employs closed-form statistics. SPLASH is fully parallelized and implemented in a dockerized Nextflow pipeline, making it suitable for massive analyses on high-performance computing clusters. We show the run time and required memory for each parallelized step to call significant anchors for multiple donor-tissues (Suppl. Figure 4), demonstrating that SPLASH calls significant anchors in <40 minutes for a donor-tissue with 400 SS2 cells (provided there is sufficient capacity for full parallelization).

Among called anchors, those classified as splicing had the highest average number of compactors per anchor (Suppl. Figure 5A), consistent with the fact that most genes have only few highly expressed splice variants per tissue (Ezkurdia et al. 2015).

SPLASH+ classified 5.75% (20,891) of anchors as centromeric (19,989,187 reads) (Suppl. Figure 5B; Suppl. Table 2). The human centromere was assembled for the first time in the T2T genome (Altemose et al. 2022); however, as T2T is based on a single cell line, it lacks population-level or single-cell variation. Supporting the limitations of genome alignment, 86% (46,348) of the compactors for centromeric anchors comprising 14% (2,800,038) of their reads failed to align to T2T. Pericentromeric DNA, including human satellite repeat families HSat1-3, is known to be transcribed in certain in vitro and in vivo contexts, but have not yet been studied at single-cell resolution (Altemose et al. 2022). SPLASH+ detected 6,418 anchors containing two consecutive CCATT repeats (or its reverse complement) which defines HSat2. We analyzed the compactors generated for a single anchor containing 2 consecutive CCATT repeats, 166 compactors in donors 1, 2, 4, 7, 8, and 12 (Figure 2A). Compactor sequence diversity for this anchor is extensive as illustrated in the multiway alignment (Figure 2A). 73 compactors (53% of anchor reads) did not map to T2T by STAR but all compactors BLAST perfectly to the T2T genome, supporting the idea that the compactors precisely recover true biological sequences. SPLASH+ detected substantial expression variation in multiple cell types including skeletal muscle satellite, mesenchymal, and tongue basal cells that possess proliferative potential (Figure 2A). Basal cells in donor 4 tongue express 23 of the 26 compactors that were found in this donor and tissue (Suppl. Figure 6). The enrichment in proliferative cell populations suggests the hypothesis that expression levels of

pericentromeric repeats and replication are linked, as has been explored in limited previous studies (Lu and Gilbert 2007; Probst et al. 2010).

We also used SPLASH+ to identify expression variation in non-coding RNA. These loci are extremely difficult to map due to their high copy number and repetitive structure. We bypassed genome alignment and queried compactors against the Rfam (Kalvari et al. 2018), a database of non-coding and structured RNA families, using infernal cmscan (Mistry et al. 2013). The most highly expressed non-coding families were ribosomal RNA in Eukaryotes (6.9M reads, 66% of reads assigned to Rfam-annotated compactors), Prokaryotes (2.1M reads, 20.64%), Archaea (871K reads; 8.28%), and Microsporidia (432K reads; 4.11%) (Suppl. Figure 7, Suppl. Table 1). Some detected rRNA could represent contamination or microbiome composition, as has also been reported by a recent microbial analysis of human single cells (Mahmoudabadi, Tabula Sapiens Consortium, and Quake 2023). The most abundant non-ribosomal noncoding RNA was U6-snRNA (*RNU6*) (22K reads; 0.22% of all Rfam-annotated reads), a small nuclear RNA involved in the spliceosome. *RNU6* has recently been shown to have high cytoplasmic representation (Mabin et al. 2021), suggesting this abundance may be expected due to polyadenylation. More than 75% of compactors assigned to *RNU6* failed to map by STAR, due to multimapping to the 1000+ annotated *RNU6* loci (Figure 2B). We employed the annotation-by-association approach for unaligned compactors associated with anchors with at least one *RNU6*-matching compactor, referring to them as U6-by-association. We computed the minimum Hamming distance for each U6-by-association compactor and U6-directly-annotated compactor (directly matched an *RNU6* variant by Rfam mapping) relative to the reference set of *RNU6* variants. Both groups had comparable hamming distance to the reference *RNU6* variants (Figure 2C), suggesting that U6-by-association compactors are potentially false negatives missed by Rfam annotation. There were eight distinct annotated *RNU6* or *RNU6*-pseudogene variants that had uniquely-matching compactors with hamming distance 0, including *RNU6-6P* whose compactors had non-uniform single-cell expression in donor 2 skin (Figure 2D).

We also observed that 30% of *RNU6* compactors had aligned past the 3' end of the gene, while only 3% mapped upstream of the gene. For example, compactors mapping to *RNU6-8* (659 supporting reads) spanned 45 bps downstream of its 3' end (Figure 2E). This strongly supports the expression of *RNU6* variants with extended 3' end, to our knowledge for the first time. There were also *RNU6* variants with compactors mapping upstream of 5' end (*RNU6-6P, 315* supporting reads) or completely within the annotated locus (*RNU6-116P*) (Figure 2E).

SPLASH+ improves precision of spliced calls and identifies extensive splicing in *CD47* including novel isoforms

Over 95% of human genes undergo alternative splicing (Pan et al. 2008), but the number of dominant expressed isoforms is mainly based on bulk tissue-level analyses and remains a topic of ongoing debate (Ezkurdia et al. 2015; Arzalluz-Luque and Conesa 2018). This debate arises from alignment-based reference-first approaches with approximate statistical inference due to problems associated with mapping to multiple isoforms (Zheng, Ma, and Kingsford 2022).

SPLASH+ reported 20,385 anchor calls classified as splicing across all donors and tissues (11,995 unique anchor sequences from 3,700 genes), representing single-cell-regulated alternative splicing (Suppl. Table 3). 73.2% of splicing anchors corresponded to annotated alternative splicing junctions (Figure 3A). 7.5% of anchors (1,537) had compactor sequences mapping to an unannotated junction. 1,387 anchors had >10% reads mapping to the unannotated junction of which 706 anchors were found in more than one donor-tissue pair. We compared SPLASH+ against SpliZ (Olivieri, Dehghannasiri, and Salzman 2022), the current best performing algorithm for predicting single-cell-regulated splicing. Given that Tabula Sapiens is a real dataset with no ground truth, for evaluation, we utilized reproducibility of the called splicing across replicates, a standard practice for evaluating discoveries on such datasets. Focusing on the lung, blood, and muscle where >1 donor was available, SPLASH+ achieved higher concordance for the same tissue between different donors (Figure 3B, Suppl. Figure 8). In both blood and lung, SPLASH+ showed significantly higher reproducibility for genes with splicing anchors compared to SpliZ. For example in blood, SPLASH+ found 15.4% (77/501) of

splicing genes in both donors, whereas SpliZ found only 3.3% (9/272) in both. Similarly for lung, SPLASH+ found 16.2% (202/1250) compared to SpliZ's 5.8% (75/1289) in both donors (Figure 3B). For muscle, both SPLASH+ and SpliZ called almost the same number of unique genes across three donors. Yet, 111 genes by SPLASH+ were shared among all three donors, while SpliZ only had 17 shared genes (Suppl. Figure 8). This improvement by SPLASH+ is noteworthy given that it did not use any cell type metadata for inference. The most highly expressed anchor missed by SpliZ but found by SPLASH+ in all three muscle replicates was *GAS5*, a noncoding RNA regulating apoptosis and growth (Mourtada-Maarabouni et al. 2009; Kino et al. 2010). *GAS5* shows reproducible cell type- and compartment-specific alternative splicing, where CD8+ t cells in all replicates consistently show higher rates for the isoform with shorter intron (Figure 3C).

Recent reference-based metadata-guided studies on human cells and experimental validations have found that *MYL6* and genes within the TPM family (*TPM1*, *TPM2*, *TPM3*) undergo highly cell-type-specific alternative splicing (Olivieri et al. 2021). All these genes were also found by SPLASH+, highlighting its power for detecting cell-type-specific patterns even without utilizing cell type metadata. In muscle, true positives *MYL6* and *TPM1* were significant in all three donors; in contrast, SpliZ only called *TPM1* and *MYL6* in donors 1 and 2. Both SPLASH+ and SpliZ called *TPM2* and *TPM3* in only donor 4.

We also investigated *CD47*, a clinical target for both cardiovascular events (Kojima et al. 2016) and cancer immunotherapy (Gordon et al. 2017) as our previous work showed *CD47* isoform expression was compartment-specific (Olivieri et al. 2021). Among all *CD47* anchors classified as splicing, SPLASH+ detected 10 distinct spliced isoforms, including 5 novel isoforms. One of the *CD47* anchors reveals expression of 8 distinct isoforms including 2 novel isoforms (Figure 3D), all impacting either the cytoplasmic or transmembrane domains. Compartments prefer different isoforms: endothelial and stromal compartments predominantly express E7-F2-3'UT isoform, while immune and epithelial cells in addition to this isoform also express F2-F3-F4-3'UT and E7-F2-F3-3'UT isoforms, respectively. One of the isoforms detected is intron-F2-3'UT (red, Figure 3D); if this isoform indeed represents full intron retention, it would also result in a stop codon after the first transmembrane domain, similar to E7-3'UT isoform. We also detected two novel isoforms: E3-3'UT and E7-F4-3'UT, being the only detected isoforms in 4 cells (Figure 3D). We independently tested for the existence of the novel junctions in these isoforms across all human RNA-seq data deposited in the SRA database (~285K datasets) via the Pebblescout, a fast k-mer tool that queries RNA-seq studies for the existence of user-provided kmers (Methods). This analysis showed significantly higher prevalence of the two detected novel junctions (E3-3'UT and E7-F4-3'UT) in SRA compared to the control set of unannotataed junctions generated by concatenating all exon pairs absent from annotated junctions in *CD47* (Suppl. Figure 9): Pebblescout reported 131K, 9,540, and 2,983 studies on average for the annotated junctions, the two SPLASH-detected unannotated junctions, and the control set, respectively.

SPLASH+ revealed new insights into splicing of *RPS24*, a highly conserved essential component of the ribosome. *RPS24* has 5 annotated isoforms that include ultraconserved intronic sequence and microexons (Olivieri et al. 2021). SPLASH+ detected 4 *RPS24* isoforms in donor 2's lung cells, consistent with our previous findings of compartment specificity for this gene. Moreover, SPLASH+ identified a novel isoform containing only a 3bp microexon which both STAR and BLAT missed and was detected only through multiple sequence alignment (Figure 3E). Given *RPS24* is highly expressed, we were able to further validate its regulated splicing by analyzing 10x Visium spatial transcriptomic data from lung (SRR14851100). We used *RPS24* anchors detected in scRNA-seq as seeds for compactor generation in the Visium data. We found significant spatial organization for the two compactors corresponding to the two most abundant *RPS*24 isoforms in scRNA-seq. The isoform containing the microexon had higher expression in the two bronchiolar structures marked by red ellipsis in the histology image (Figure 3F), comprising mostly club and ciliated cell types (Travaglini et al. 2020). Both cell types exhibited concordant enrichment of microexon containing isoform in both lung replicates in scRNA-seq, confirming our spatial predictions (Suppl. Figure 10). Together, this analysis shows that without using any cell metadata or reference, SPLASH+ can generate biologically consistent predictions for splicing in human tissue.

Genes with pan-tissue, single-cell-regulated splicing are enriched for splicing factors and histone regulation

2,118 genes with splicing anchors called by SPLASH+ in more than one tissue, including 10 genes found in 18/19 tissues, being referred to as *core genes* (Figure 4A). We performed GO enrichment analysis on 61 genes with splicing anchors detected in at least 15 tissues. Remarkably, enriched pathways with the highest log-fold change were all involved in mRNA processing and splicing regulation (Fisher test, FDR p-value < 0.05, Figure 4B). These results imply that splicing factors and histone modifications themselves are under tight splicing regulatory mechanisms in diverse tissues, possibly co-regulating their expression.

SPLASH+ reveals diverse splicing regulation in the 10 core genes, detecting 71 isoform variants, including four unannotated isoforms, one in each of the *HNRNPC*, *KMT2E*, *SRSF7*, and *SRSF11* genes. While each of the core genes are known to have significant regulatory roles, the extent and complexity of their splicing regulation, as revealed by SPLASH+, has been underappreciated.

HSP90AA1 was the sole core gene found in all 19 tissues (Figure 4C). *HSP90AA1* is one of two isoforms of the HSP90 heat shock protein functioning in myriad cellular processes including a chaperone of protein folding (Hoter, El-Sabban, and Naim 2018) and transcriptionally regulated under cell stress (Zuehlke et al. 2015). We detected anchors with 12 distinct differential intron retention events for 7 introns of this gene including unannotated intron retention events for introns 1 through 4 and intron 7, and a novel splicing between the first and fourth exons (Figures 4C,D). Detected intron retentions are highly compartment-specific, with higher expression fraction for immune and stromal cells compared to epithelial and endothelial cells (Figures 4C,D). For 9 anchors, immune cells had the highest intron retention fraction compared to other cells. Anchor 6 had the strongest differential pattern between compartments with 44%, 22%, 17% and 0% for intron retention in immune, epithelial, stromal, and endothelial cells, respectively. Due to its known transcriptional regulation upon stress, we cannot exclude the possibility that the regulated intron retention is due to differential compartment-specific response to dissociation stress. However, compartment-specificity and abundance of intron retention forms suggests *HSP90AA1* has previously-unknown post-transcriptional regulation, even if part of the detected signal is differentially regulated physiological response to dissociation. We also queried the novel junction (between exons E1 and E3) via Pebblescout, which indicated a 6-fold increase in prevalence compared to control set of unannotated junctions (Figure 4E). To our knowledge, splicing regulation in *HSP90AA1*, is a potentially novel mechanism to tune the protein levels of this critical molecular chaperone.

Five core genes (50%) are themselves splicing factors, including nuclear ribonucleoproteins (hnRNPs) *HNRNPC*, *HNRNPDL* (Suppl. Figure 11), and SR family members *SRSF5*, *SRSF7*, *SRSF11*. The detected compactors represent complex isoforms, some un-annotated, and some including splicing into ultraconserved intronic regions known to create poison exons in *SRSF5*, *SRSF7*, *SRSF11* and *HNRNPDL* (Lareau et al. 2007; Königs et al. 2020; Raihan et al. 2019; Ni et al. 2007). The remaining four core genes are involved in histone regulation or nuclear co-repression: *KMT2E* a histone methyltransferase with known mutations in neurodevelopmental disorders (O'Donnell-Luria et al. 2019), *PCMT1*, another histone methyltransferase (Biterge et al. 2014), *HMGN3*, a high mobility group nucleosome binding protein and transcriptional repressor and *NCOR1*, the nuclear co-repressor (Perissi et al. 2010). SPLASH+ detects a poison exon and intron retention event in *NCOR1* (Figure 4F) and *KMT2E,* respectively, predicted to trigger nonsense mediated decay. Core genes all have portions of highly conserved intronic sequences, suggesting a mechanism for splicing regulation. Together, these results support the idea that alternative isoforms play a critical regulatory role that includes use of premature stop codons and complex alternative splicing in the 5' UTR, gene body, and 3' UTRs. While these isoforms had been predicted by analysis of EST data and in cell culture, to our knowledge, direct evidence of regulated splicing patterns of any of these regulators in single cells has been previously missing (Ding et al. 2022; Lareau et al. 2007).

We also used SPLASH+ to identify genes classified as splicing with high numbers of variants, which are known to drive organization of complex tissues (Schmucker et al. 2000; Yagi 2008). We identified genes with the most unique splice junctions across the entire dataset (Figure 4G): *IL32*, *GAS5*, and *RBM39* (33, 28, and 28 junctions, respectively). Additionally, 49 genes, including *PRPF38B*, *TACC1*, *CCDC66*, and *TAX1BP3*,

had at least 15 distinct splice junctions (Figure 4G). Among these genes are known oncogenes *TACC1* (Cully et al. 2005) and *CDC37* (Gray et al. 2008) with 19 and 11 splice variants, respectively. Splicing factors *SRSF10* and *RBM39* (each found in 17 tissues) were also highly ranked, having 16 and 28 splice variants, respectively, and are all associated with tumor initiation or growth (Kim et al. 2015; Xu et al. 2021; Shkreta et al. 2021). For *PRPF38B*, a splicing factor with prognostic biomarker potential in breast cancer (Abdel-Fatah et al. 2017), 17 distinct splice junctions were detected (across all of its anchors) in 17 tissues. One of its anchors shows compactors with complex alternative splicing involving skipping of two cassette exons, alternative 5' splice sites, intron retention, and a novel splice junction, which is the dominant isoform in 4 immune and stromal cells (Figure 4H). SPLASH+ reveals complexity of splicing regulated at the single cell level missed by current methods, and supports the idea that many human genes have single-cell-specific splicing patterns, rather than exclusively favoring a dominant form.

SPLASH+ detects Alu element insertion polymorphisms *de novo*

We used SPLASH+ to find potentially novel exonized sequences by prioritizing compactors with stringent criteria that include partial mapping to both the human genome and known transposable elements (Methods). The majority of these compactors (24/32, 74%) contained soft clipped reads that map to Alu repeats. Alu insertions are known to be both individual-specific and single-cell-specific polymorphic and mobilized and have significant impact on gene expression (Payer et al. 2021). SPLASH+ detected an Alu insertion for *PSPH*, a Phosphoserine phosphatase, in two donors: donor 2's small interestine, and mammary and muscle from donor 4 (Suppl. Figure 12A). In donor 2 muscle, other compactors that map to *PSPH* support a novel unannotated exon skipping event (Suppl. Figure 12A), which was also found in 8 other donor-tissues. These 3 compactors all map to the exon containing the translation start site and the 5'UTR, accounting for 22-25% of the reads from this anchor (Suppl. Figure 12B). The rate of Alu insertion was cell-type-specific (Suppl. Figure 12 C). For example, in donor 4 muscle, the Alu-containing isoform comprises 100% (14/14) and 81.25% (78/96) of reads in macrophages and skeletal muscle satellite stem cells, respectively. Other events were represented in one donor-tissue pair: in the lung of donor 2, >15% of reads originating in gene *AFMID* show exonization of an Alu element after an exon (Suppl. Figure 12D).

SPLASH+ rediscovers and expands the scope of V(D)J transcript diversity

Single cells can somatically acquire copy number variation, SNPs, or repeat expansions. Detection of genetic diversity in single cells has required custom experimental and computational workflows. Our statistical analysis demonstrates that under the null assumption where all cells in a donor express only two alleles of any given splice variant, the probability of observing many counts for compactors beyond the two genuine ones decays rapidly (Methods). Positive controls expected to violate the null include mitochondria where genomes are polyploid (Barrett et al. 1983) or the the rearrangement of immunoglobulin loci in B cells and T cell receptor loci which undergo V(D)J recombination.

To investigate global trends, for each anchor, we collected all distinct compactors across donor-tissues (Suppl. Table 5). Anchors mapping to immunoglobulin genes had the highest number of distinct compactors and differentiated from other genes on this purely numerical criterion (Figure 5A). For example, an anchor mapping to immunoglobulin kappa chain (*IGKC*) had the highest number of compactors (140) across all donors and tissues. As expected, these anchors were observed only in immune tissues. Centromeric anchors defined as containing CCATTCCATT repeats had also substantial compactor diversity (Figure 5A), consistent with the fact that centromeres have substantial sequence diversity within and across individuals.

Mitochondrial genes also had substantial diversity, notably in an anchor from gene *MT-ND5* in donor 1 lung that had 24 compactors. This was the highest number of compactors for a single anchor mapping to an annotated gene within a single donor-tissue. *MT-ND5* is a component of the transmembrane electron transport chain in mitochondria with previously reported recurrent mutations with clinical significance (Jaberi et al. 2020; Wang et al. 2022).

While current approaches require mapping to the genome on a read-by-read basis, SPLASH+ enables a statistics-first micro-assembly to detect variants in the B cell receptor (BCR) locus avoiding genome alignment. We compared SPLASH+ power to detect BCR rearrangement in B cells from spleen of donors 2 and 7 against the state-of-the-art custom pipeline BraCeR (Lindeman et al. 2018). We name a compactor as an "IG-compactor" if STAR maps it to an immunoglobulin gene (Methods). As a more stringent criterion, we further annotated IG-compactors using IgBLAST (Ye et al. 2013) and considered only those IG-compactors that had both variable (V) and joining (J) immunoglobulin gene segments identified thorugh IgBLAST (Methods). In donor 2 spleen, not only did SPLASH+ detect an IG-compactor with annotated V and J genes in every cell for which BraCeR reported a BCR contig, but it also found BCR evidence in 53 additional B cells (41 memory B cell and 12 plasma cell) which BraCeR missed (Figure 5B; Suppl. Table 6). Similarly, in donor 7 spleen both BraCeR and SPLASH+ found evidence of BCR rearrangement in the same 47 plasma B cells, but SPLASH+ also found BCR evidence in 2 additional plasma B cells (Figure 5B). SPLASH misses one cell among memory B cells and naive B cells from donor 7, which BraCeR called. We further investigated SPLASH+ calls missed by BraCeR to see if they had evidence of fully functional in-frame transcripts. 100% of plasma cells in both donors and 93% and 100% of memory B cells in donors 2 and 7, respectively, that were called only by SPLASH+ had at least one in-frame IG-compactor as predicted by IgBLAST (Figure 5B), suggesting that higher sensitivity of SPLASH+ can result in the identification of high confidence fully-functional immunoglobulin transcripts with implications in adaptive immune response.

Of the cells analyzed by both algorithms, SPLASH+ called IG-compactors in 136 cells from donor 2 spleen, and 142 and 123 in donor 7 spleen and lymph node (Suppl. Table 6). We tested if SPLASH+'s IG-compactors were concordant with BraCeR's calls in these cells by computing the minimum Hamming distance between the two sets of BraceR contigs and SPLASH+ IG-compactors for each cell (Figure 5C). A high fraction of cells had perfect matches between SPLASH+'s and BraCeR's sequences: 58.1%, 65.8%, and 64.1% for donor 2 spleen, donor 7 spleen, and donor 7 lymph node, respectively, with increasing concordance as more relaxed minimum Hamming distance criterion is considered between IG-compactors and BracerR contigs (Figure 5C). To identify IG-compactors that were missed by BraCeR, compactors with Hamming distance of greater than 30 to all BraCeR contigs were collected. 416 anchors had more than 3 of such compactors. One of these anchors had 8 compactors in donor 2 spleen where each compactor had partial alignment to one or both of two different IGHV loci, likely representing distinct V segment inclusion (Figure 5D). In summary, SPLASH+ automatically detects V(D)J rearrangement agreeing with, but extending what detected by BraCeR in expected B cell subtypes, with implications for downstream biological inference and opportunities to explore other sequences nominated by SPLASH+ that do not meet the stringent criteria used here.

Cell type-specific hypermutation or RNA editing including in intronic regions of *AGO2*, UTRs of *ANAPC16* and the 5' and translational start of *ARPC2*

We investigated anchors that had either comparable or more distinct compactors across donor-tissues than those associated with V(D)J. This list includes anchors with compactors showing abundant canonical Adenosine-to-inosine (A-to-I) RNA editing in *ANAPC16* (Figure 6A), a regulator of anaphase, and *AGO2* (Figure 6B), the argonaut protein involved in miRNA targeting.

In *AGO2*, an enzyme critical for RNA interference, SPLASH+ generated 18 compactors across donor 2 skin, lung, and donor 4 muscle from a single anchor (Figure 6B). As with *ANAPC16*, the majority of reads are assigned to edited variants constituting 84%, 64%, and 85% of reads in donor 2 skin, lung, and donor 4 muscle, respectively. These compactors support A-to-I canonical intronic editing in an Alu element at 8 positions including one editing, supported by 36% of reads across three donor-tissues, that is not reported in the REDIPortal database of RNA editing events in humans (Picardi et al. 2017). Also, a compactor representing 15% of reads in donor 2 skin was both edited and showed circular RNA backsplice junction (Figure 6B), suggesting that splicing precedes editing. The extent of editing in these loci and extremely low support from a comprehensive, ultra-deep reference database (each editing event being supported by at most 14 studies out of 9,642 studies curated in REDI) suggest reads at this locus would be unmapped or

mismapped with conventional pipelines (Eisenberg and Levanon 2018). In fact, as shown in the BLAT alignment of these compactors (Figure 6B), for the first and fifth compactors in donor 2 skin, BLAT missed few editing events and instead reported a split alignment, illustrating the challenges of reference-based alignment method for detecting these events.

Extensive RNA editing diversity was also found in *ARPC2*, the actin-related protein 2/3 complex subunit 2. We focused our analysis on the single anchor from this gene with the most compactors in a single donor (16 compactors). Compactors for this anchor represent prevalent base pair changes with respect to the reference, with consistent editing rates (~36%) in bone marrow cells from donors 11 and 13 (Figure 6C). These editing events lack a known mechanistic explanation as they are neither canonical A-to-I editing nor show signatures of reverse-transcriptase induced base pair changes due to RNA base modifications (Werner et al. 2020). Intriguingly, the changes are concentrated in the start codon and would likely affect translation initiation. *ARPC2* has other known non-canonical translation regulation: an internal ribosome entry site in its 5' UTR (Al-Zeer et al. 2019), and un-annotated splicing in its 5' UTR, suggesting the possibility of non-canonical translation initiation. Because of its surprising nature, we tested if apparent editing in *ARPC2* existed in other donors, but SPLASH+ was underpowered to detect it. We generated compactors using the above anchor for all cells in the study: 15% of all reads in the dataset containing this anchor had discrepant bases in the 17-bp window (chr2: 218,703,167–218,703,183), and 11% had base pair changes in the start codon, much higher than expected by chance under sequencing error rate (median error rate .01% for the Illumina NovaSeq 6000) (Stoler and Nekrutenko 2021).

We further used Pebblescout to test the frequency of editing events detected in *ARPC2* (Methods). Editing events detected by SPLASH+ were reproducibly found in human RNA-seq samples (Suppl. Figure 13A). To investigate if the detection rate is higher than base pair changes due to sequencing error, we also constructed and queried decoy sequences containing base pair changes not predicted by SPLASH+. The detection rate was significantly higher for the editing locations predicted by SPLASH+ (Suppl. Figure 13B) and the prevalence difference between editing and decoy sequences grew with the number of base pair changes. For example, k-mers with 2 detected edits by SPLASH+ were present in an average of 26K samples whereas decoys were detected in only 163 (~160-fold change). No decoys with 3 edits were detected whereas SPLASH+ kmers were detected on average in ~3K samples .

The reproducibility across donors, tissue specificity, stereotyped positions, and the level of diversity are strong evidence against these base-pair changes being an artifact or arising in DNA. In summary, SPLASH+'s automatic statistical inference identifies extensive and novel editing in single cells. To our knowledge, these events have not been and cannot be detected with current custom workflows (Cohen-Fultheim and Levanon 2021).

Evidence for repeat polymorphism including in *BGN* and *VSNL1*

Other anchors with high compactor diversity show evidence of repeat polymorphism. For example, for an anchor mapping to 3' UTR of *BGN*, compactors show multiple AG dinucleotide repeat lengths (Figure 6D). *BGN* codes for biglycan and has roles in metabolic pathways and cell proliferation (Ying et al. 2018; Morimoto et al. 2021). Dinucleotide repeats are known to be polymorphic, but repeat length variation could also be generated during PCR, through a process called slippage (Shinde 2003). Thus, we investigated if polymerase slippage could explain the repeat polymorphisms found by SPLASH+. The profile of *in vitro* Taq PCR slippage has been studied, showing that after each PCR cycle the repeat element is contracted with a non-negligible probability p (Shinde 2003). Under this error model, the contraction process is equivalent to a binomial process and therefore we should observe a unimodal probability distribution of repeat lengths shorter than the reference repeat in the genome.

The top two dominant repeats in the vast majority of cells in donor 2 lung were of different lengths and distinct from the reference repeat (Figure 6D; Methods): 9 cells and 10 cells had contractions of 5 and 6 as their primary and secondary repeat, respectively. We tested if observed repeat variants in *BGN* are consistent with the slippage error model. In donor 2 lung, the dominant repeat lengths were 15 (reference repeat length)

and 10 (20% and 53% of reads, respectively), inconsistent with being generated *in vitro* by PCR slippage, which would generate a unimodal probability distribution of repeats ranging from 15 to 10. Instead, this supports a model that donor 2 has two *BGN* alleles with 15 and 10 repeats. To further test if these variable repeat lengths could be explained by PCR artifacts, we generated compactors in donor 1 lung where SPLASH+ did not call the *BGN* anchor. The expression patterns for two repeat lengths in Donor 1 were compatible with being generated due to PCR artifacts as the reference allele comprised majority of of reads (78%), while contractions of 1 or 2 AG dinucleotides accounted only for 22% of reads which is in contrast to what was observed in donor 2 lung. This, together with the donor-specific repeat polymorphism is strong evidence that SPLASH+ calls *BGN* because single cells express different allelic repeat lengths, perhaps due to allelic imbalance, rather than being due to PCR artifacts.

Another repeat polymorphism found by SPLASH+ was in *VSNL1* (Visinin like 1 protein) a neuronal sensor calcium protein (Figure 6E). Prior literature shows that repeat polymorphism in *VSNL1*, is highly conserved in vertebrates and implicated in dendritic targeting (Ola et al. 2012; Riley and Krieger 2009). Contractions of 6 and 7 repeats were most abundant (together 75% of reads, Figure 6E); Contractions of 0, 1 and 2 repeats had 8%, 7%, and 10% of reads, respectively, with no intermediate repeats between these and contractions of 6 and 7. Highlighting the importance of avoiding cell type metadata for testing, *VSNL1* polymorphism was detected predominantly in one cell type: tongue basal cells, which are thought to be stem cell progenitors (Iwai et al. 2008). If these alleles were due to polymerase slippage, error models predict observing other contractions such as -8; however, none were observed. This, along with the diverse single-cell expression of non-allelic repeat variants, suggests donor-specific somatic diversification of repeat numbers, likely originating from somatic variation within a donor rather than PCR.

Conclusion

Alignment to the human reference genome is thought to be a prerequisite for the analysis of RNA-sequencing, and great efforts have been made to provide complete and curated reference genomes and transcript annotation. Similarly, cell type metadata is currently considered a critical starting or ending point for the analysis of single-cell sequencing experiments. In this work, we showed that novel biology of transcriptome regulation can be discovered using a direct statistical approach to analyze sequencing data without cell type metadata and using reference genomes for post-inferential interpretation.

In addition to computational and conceptual unification, SPLASH+'s reference-free approach predicts biology in single cells that has been missed by customized bioinformatics methods in multiple domains. SPLASH+ enables automatic discovery of cell-specific diversity in non-coding RNA such as *RNU6* and centromeric repeats, among others which are, to our knowledge, unavailable with current bioinformatic approaches. In domains where custom algorithms exist, such as detection of RNA splicing, V(D)J recombination, or RNA editing, SPLASH+ unifies and extends discovery. SPLASH+'s findings of complex splicing regulation in splicing factors provide direct evidence of and extend previous predictions of such regulation from EST databases and DNA sequence. We also uncovered novel cell-regulated splicing in diverse genes, including noncoding RNAs such as *GAS5*, together suggesting new candidates to prioritize for functional studies. The extent of splicing diversity uncovered by SPLASH+ provides further evidence that transcriptome complexity in primary single cells is extensive. This implies the power and scale of a data science driven approach will be needed to predict regulation and function for this transcriptome diversity, as experimental approaches cannot be scaled to the throughput required to study each isoform.

This manuscript also showed how SPLASH+'s approach unifies disparate areas beyond splicing discovery, including variation in noncoding RNA loci, centromeres, detecting genome insertions such as Alus, V(D)J recombination, RNA editing, and repeat polymorphisms. This suggests further avenues for discovery of human disease biology in both RNA-seq and DNA-seq where SPLASH+ allows repeat polymorphisms to be further scrutinized. For example, dinucleotide repeats detected in this study are predicted to be bound by CUG-binding protein (*MBNL1*) and TDP-43 (Takahashi et al. 2000; Buratti and Baralle 2001). The repeat polymorphisms identified by SPLASH+ further suggest the potential for predicting cell-specific impacts of

repeat expansions, including their contribution to stress granule formation and disease (Sproviero et al. 2017; Estany et al. 2007). To focus this work, we did not include a discussion of other dimensions of transcript diversity found by SPLASH+, including alternative polyadenylation within human transcripts, cell-level variation in indels, potential structural rearrangements within the human genome, and even non-human sequences found by SPLASH+ in this dataset, which include an enrichment of bacteriophages that may reflect prokaryotic contribution to the human metatranscriptome.

In this first unbiased systematic analysis of human transcription diversity in single cells, SPLASH+ establishes a unified statistics-first approach to sequence analysis, which reveals prevalent transcript diversity variations overlooked by current bioinformatics. While examples discussed in this study provide a glimpse into its complexity due to subsampling of human cells and tissues, analysis of larger single-cell datasets, as well as DNA sequencing, hold potential for a new generation of genetic and transcriptomic diversity mechanisms as predictors of cellular phenotype or disease. Indeed, SPLASH+ is versatile and applicable to any RNA-seq or DNA-seq study, offering opportunities for massive, statistically driven study of transcriptomes. Finally, we should note that while SPLASH+ is applicable to diverse genomics problems, it may not be suitable for studies where the main focus is on differential gene expression analysis. Also, providing better interpretation for those SPLASH+'s calls that do not have BLAST hit is part of the ongoing work.

Acknowledgments
We thank all members of the Salzman lab for comments, Aaron Straight, Pragya Sidhwani and Charles Limouse for reading and interpretation of the results on centromeric repeats. Lu Chen for discussion of RNU-6 variation, Liana Lareau for comments. We also thank Serana Tan for her help with interpreting Visium data. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results. J.S. is supported by the Stanford University Discovery Innovation Award, National Institute of General Medical Sciences grants R35 GM139517 and the National Science Foundation Faculty Early Career Development Program Award no. MCB1552196. RB was funded by the NCI grant 5F31CA243170-02. TZB was funded in part by the Stanford Graduate Fellowship and the NSF GRFP.

Competing Interests
K.C., T.Z.B., and J.S. are inventors on provisional patents related to original implementation of SPLASH. The authors declare no other competing interests.

Figures:

Figure 1. A general reference-free alignment-free approach unifies diverse transcrptomic diversity analysis. (A) Current RNA-seq analysis methods, relying on read alignment to a reference genome, introduce biases and blindspots, and are tailored for specific applications. SPLASH provides a unified statistical solution, bypasses alignment, for detecting myriad transcriptome diversifying mechanisms. SPLASH parses reads from all input samples (cells) to extract constant kmers (anchors) that are followed by a set of diverse kmers (targets). The core of SPLASH is to test if anchors have non-uniform expression for their targets across cells. For each anchor, SPLASH first creates a contingency table of target counts across cells, and then computes a closed-form p-value without resampling, identifying anchors with significant p-values indicative of having single-cell-dependent target distribution evidencing regulation and can subsequently be used for downstream analysis for a wide range of applications. (B) SPLASH-called anchors are extended to longer sequences or "compactors" through a statistically-valid local assembly approach. The process involves collecting reads containing the anchor, recursively extending the sequence beyond the anchor by comparing nucleotide frequencies at each position against a threshold based on sequencing error, resulting in branches corresponding to compactors for the anchor that are subsequently used for biological interpretation and classification. (C) SPLASH+ integrates SPLASH with local assembly ("compactors") and biological interpretation to categorize calls into biologically-relevant events such as splicing, SNP, internal splicing, V(D)J recombination or antibody diversity and repeat expansion. Unmapped compactors are annotated by the identity of other compactors linked to the same anchor are classified through *annotation by association.* Also,

additional annotation information can be obtained by mapping compactors to databases like Rfam, Pfam, and BLAST.

Figure 2. Transcript diversity variation in repetitive RNA loci. (A) SPLASH+ detects centromeric repeat array variation as illustrated for the anchor ATTCCATTCCATTCCATTCCATTCCAC, containing 5 contiguous repeats of ATTCC the characteristic sequence for canonical pericentromeric. Among the anchors containing ATTCCATTCC, this anchor has the most diverse compactors (147 compactors) ordered concordantly through multiway alignment. The heatmap shows the cell-type-specific expression of compactors in logarithmic scale by showing the read counts assigned to each compactor per cell type (collapsed across the donors and tissues having the celltype as shown next to each cell type name). (B) The 71 detected *RNU6* compactors had high conservation with *RNU6* reference sequences which has 1,281 gene and pseudo-gene loci dispersed throughout the human genome. (C) Direct and annotated-by-association *RNU6* compactors have comparable abundances and sequence similarity to *RNU6* reference genes, implying that the annotated-by-association compactors are in fact missing from the reference annotation. (D) Heatmap showing the differential compactor count per cell (skin cells from donor 2) for a compactor mapped to pseudo-gene *RNU6-6P* and other compactors from the same anchor but with no unique alignment. (E) Perfect multiway alignments of compactors to the annotated *RNU6* loci show compactors align both upstream and downstream of the annotated boundaries.

Figure 3. SPLASH+ enables de-novo analysis of alternative splicing in single cells. (A) Dot plot showing the number of distinct splicing anchors per donor-tissue pair, with >55% involving annotated alternative splicing events (based on T2T). (B) Upset plots compare the concordance of significant splicing genes called by SPLASH+ (in red) and SpliZ (Olivieri, Dehghannasiri, and Salzman 2022) in lung (left) and blood (right) replicates. SPLASH+ achieves higher concordance in both tissues despite not using cell metadata, which have different distributions and or composition in the replicates. (C) The plot shows reproducible compartment-specific alternative splicing of *GAS5* in muscle cells from 3 donors (intervals are 95% binomial confidence interval). CD8+, alpha-beta t cells consistently show a higher inclusion rate for the isoform with shorter intron (green isoform). (D) SPLASH+ detects extensive single-cell expression variation of eight *CD47* isoforms, including novel isoforms, for a single anchor across 10 donors and 14 tissues. Cell with >5 reads are included and horizontal side bars show the donor, tissue, and compartment identity for each cell. The heatmap, sorted by hierarchical clustering, includes a novel *CD47* isoform (yellow) with annotated junctions (isoform coordinates: chr3:110767091:110771730--chr3:110771761:110775311). (E) Four alternative splice variants of RPS24 in lung (from donor 2) involve inclusion/exclusion of ultraconserved region cassette exons, including a 3-bp microexon. SPLASH+ detects a novel isoform with only the microexon, missed by both STAR and BLAT which was confirmed through multiway alignment. The heatmap shows the compartment-specific usage of four *RPS24* isoforms across lung cells from donor 2, with the isoform containing the 3-nt and 21-nt exon predominantly expressed in epithelial cells and the novel isoform predominantly expressed in type ii pneumocytes. (F) Spatial validation of alternative splicing for *RPS24* using 10x Visium data of lung tissue. Top panels show the fraction of each isoform (left: isoform 1 excluding microexon, right: isoform 2 including microexon). The histology image (bottom) shows bronchial structures (marked by red ellipses), where the isoform with microexon inclusion exhibits much higher expression.

Figure 4. Pan-tissue alternative splicing regulation of splicing factors and histone modifications. (A) 57% of genes (2,118) with significant splicing in at least 2 tissues, including 10 genes found in >= 18 tissues. Notably, 8 splicing factors (green) and histone modifications (brown) are present. (B) GO enrichment analysis of genes in >15 tissues reveals pathways related to mRNA processing and splicing regulation (Fisher test, FDR corrected p-value< 0.05). (C, D) Pan-tissue intron retention diversity in *HSP90AA1*, the only SPLASH+ core gene found in all 19 tissues. Twelve anchors detect differential intron retention events in 7/8 of *HSP90AA1* introns. Compartment-specific expression of compactors corresponding to intron retention (pink) and splice junctions (green); the barplot on the right shows total read counts for intron retention and splicing isoform for each anchor. Fractional isoform distribution is depicted for the anchor with the most compactors (anchor 1) and the anchor found in the most tissues (anchor 3). (E) Number of SRA studies containing each *HSP90AA1* junction reported by Pebblescout supports SPLASH+'s novel splice prediction. Junctions are grouped as annotated, unannotated, and the one detected unannotated junction (between exons E1 and E4) for *HSP90AA1*. (F) Single-cell dependent alternative splicing of gene *NCOR1,* involving 3 different splicing isoforms for an *NCOR1* anchor found in the most tissues. (G) Plot shows the number of tissues and unique

splice junctions for each gene with a splicing anchor. *IL32*, *RBM39*, and *IGKC* have the most unique splice junctions. (H) Alternative splicing of gene *PRPF38B* involving an intron retention and six splicing isoforms. This gene has one of the most diverse isoform structures, with significant alternative splicing in 11 donors and 17 tissues.

Figure 5. SPLASH+ detects V(D)J recombination with highest levels of transcriptome diversity and achieves higher sensitivity than BraCeR. (A) The plot showing the number of anchors and number of donors for each anchor, indicating that anchors with the most compactor diversity extend known biology of V(D)J recombination. The top marginal histogram shows the probability that each compactor category falls into a specific range of number of compactors per anchor. The right marginal histogram shows the probability of each category as a function of number of donors. Multiple sequence alignment of the anchor mapping to an immunoglobulin gene that have the most distinct compactors. (B) Comparison of SPLASH+ and BraCeR for spleen cells (that are analyzed by both methods) from donors 2 and 7 indicates higher sensitivity for SPLASH+. Each cell is called by BraCeR or SPLASH+, based on the presence of at least one BraCeR contig or IG-compactor, respectively. The x-axis represents the highest count for an IG-compactor in each cell. We also present the percentage of cells called only by SPLASH+ (green) within each cell type that have at least one in-frame V(D)J transcript as confirmed by IgBLAST. (C) Shared cells by SPLASH+ and BraCeR increase as a function of minimum Hamming distance between IG-compactors and BraCeR contigs for each cell. (D) Alignment of multiple compactors from the same anchor to *IGHV3* shows split mapping to *IGHV3-43-201* and *IGHV3-15-201*.

Figure 6. Single-cell-regulated RNA editing and repeat polymorphism detected *de novo* by SPLASH+. (A) Multiple sequence alignment of compactors generated *de novo* by SPLASH+ reveals extensive cell-type regulated editing in 5' UTR of *ANAPC16*. Red indicates A-to-I edits, and compactors matching the reference allele for each donor-tissue are highlighted in orange boxes. The side bar plots show the total count for each compactor in each donor tissue. The bars for the top 4 compactors in each donor tissue are color coded to show shared sequences in different donor tissues. Predicted miRNA binding site (Chen and Wang 2020) in *ANAPC16* is disrupted by observed edits in all four donor-tissues. (B) Similar analysis for *AGO2* identifies RNA editing within an Alu element region including a circular RNA in the third compactor from donor 2 skin. (C) *ARPC2* sequence diversity, with 20 distinct variants occurring in a 17-bp region (positions 58-74 in the compactor sequences). Multiple sequence alignment of 20 17mers in *ARPC2* exon 2 reveals base-pair changes between the 5' UTR and translational start (T2T chr2: 218,703,167–218,703,183). All kmer counts from donor bone marrow exceed the expected sequencing error rate by orders of magnitude. (D) Sankey plot displays cell counts for each primary repeat length (repeat with the highest count in cell) and secondary repeat length (repeat with the second highest count in cell) in donor 2 lung (where the anchor was called). The reference repeat length is indicated as 0 and repeat contractions are indicated relative to this reference length. Bar plot shows cell counts with reference or contraction of one repeat. Each cell is counted for a given repeat length if either its primary or secondary repeat matches that specific length. (E) The same analysis for the repeat polymorphism in *VSNL1*, showing the number of cells for each combination of secondary and primary repeat length.

Suppl. Figure 1. SPLASH+ was run on 10,326 SmartSeq2 cells from 12 donors and 19 tissues (29 donor-tissue pairs), including replicates for several tissues from multiple donors, enabling reproducibility analysis.

Suppl. Figure 2. (A) Number of detected fusions per simulated datasets. In total 56.7% (1313/2315) of fusions were detected in compactor sequences. (B) The empirical cumulative distribution function (ecdf) of the expression values (TPM) for the detected and undetected fusions by compactors, suggesting that compactors provided evidence for the majority of sufficiently-expressed fusions.

Suppl. Figure 3. Post-facto classification assigns distinct biologically-relevant RNA events to each anchor based on its compactors' characteristics. Six categories are considered: splicing, internal splicing, base pair change, 3' UTR, centromere, and repeat (ordered by priority). An anchor that cannot be assigned to any of these categories is referred to as unclassified. Reference-free categories include internal splicing and base pair change; and splicing, 3' UTR, centromere, and repeats are assigned based on compactors' alignment to the genome.

Suppl. Figure 4. (A) Run time and (B) memory usage per step (extracting and counting anchor and target sequences from input sample in parallel for each input sample, stratification of anchor sequences to allow for parallelization of p-value calculation, p-value calculation in parallel for each set of stratified anchors, and multiple testing correction) in SPLASH across eight donor tissues. SPLASH, being fully parallelized, allows the initial step of parsing input reads and extracting anchor and target sequences to run concurrently for each cell. Thus, the run time for this step is presented for individual cells. Memory requirements are displayed for the two steps consuming the most memory.

Suppl. Figure 5. (A) Compactor count per anchor in each tissue, categorized by anchor class, with splicing anchors exhibiting the lowest compactor diversity. (B) Fraction of reads assigned to each anchor category in each tissue, with base pair change anchors having the highest fraction of reads.

Suppl. Figure 6. Single-cell-regulated expression of the anchor containing pericentromeric repeat CCATT in tongue replicates (donors 4 and 7). Tongue basal cells consistently exhibit the highest expression in both replicates.

Suppl. Figure 7. The barplot showing the total read count across the entire dataset for each category of noncoding RNAs.

Suppl. Figure 8. Comparison of the SpliZ and SPLASH+ for detecting genes with significant alternative splicing in three replicates of muscle tissue (donors 1, 2, and 4), indicating higher concordance in SPLASH+ calls.

Suppl. Figure 9. Number of SRA studies reported by Pebblescout for *CD47* junctions, providing further evidence for the expression of the two novel splice predictions made by SPLASH+ for *CD47*.

Suppl. Figure 10. The two overrepresented cell types in bronchiole epithelium (ciliated and club cells) exhibit the highest expression fraction of the isoform with the microexon (dark blue isoform in Figure 3E) in lung replicates from donors 1 and 2, which is consistent with findings in the Visium data. Ciliated cells show 93 reads in 5 cells and 38 reads in 2 cells for donors 1 and 2, respectively, while club cells exhibit 22 reads in 2 cells for donor 2.

Suppl. Figure 11. Single-cell dependent alternative splicing of two hnRNPs: (A) *HNRNPDL* and (B) *HNRNPC* illustrated for the anchor expressed in the most tissues for each of these genes.

Suppl. Figure 12. (A) Potential Alu-insertion isoform in 3 tissues across two donors for gene *PSPH*. (B) Model illustrating an Alu element disrupting normal splicing, inserting into the middle of an exon and inhibiting the formation of typical isoforms, resulting in a new isoform, which consists of a portion of the original exon with the Alu element. The relative abundance of each isoform is detected from the compactor distribution, with the Alu-inserted isoform represented by compactors mapping uniquely to the human genome, with some portion that maps to transposable elements. A different compactor in the same tissue of a different donor exhibits an unannotated exon skipping event, which could point to an undetected Alu-insertion driving alternative splicing. (C) Compactor count per cell type for the *PSPH* anchor with Alu-insertion in muscle and mammary cells from donor 4 and small intestine cells from donor 2. (D) A potential Alu-insertion event in *AFMID* in the lung of donor 2.

Suppl. Figure 13. Query of the editing events detected in *ARPC2* using Pebblescout. (A) Number of studies reported by Pebblescout for the 25-mers with changes in the 6 RNA editing locations, as shown in Figure 6C. (B) Pebblescout reported compactors with editing in significantly more studies than decoy events or kmers constructed with base pair changes not predicted by SPLASH+.

Tables:

Table 1: Compactor benchmarking. This table contains the list of compactors that have match to the fusions in simulated benchmarking datasets.

Table 2: Noncoding RNA summary. Number of called noncoding RNAs separated by each category for each donor and tissue.

Table 3: Centromeric anchors summary. This table contains the sequence, read count, T2T CenSat annotation, and donor id, for each compactor of the centromeric anchors.

Table 4: Alternative splicing calls. For each detected splicing anchor across the entire dataset, this table provides gene name, donor and tissue information, effect size, number of reads for the anchor, number of cells, the coordinates for the associated splice junctions, and whether the alternative splicing is annotated according to human T2T annotation.

Table 5: SPLASH vs SpliZ Comparison. This table contains genes with significant alternative splicing detected by SPLASH+ and SpliZ in lung, blood, and muscle tissues.

Table 6: Compactor diversity summary. The number of distinct compactor sequences for each anchor in each donor.

Table 7: SPLASH vs BraceR comparison. This table contains the SPLASH IG_compactor sequences along with their IgBLAST annotations and BraCeR contig sequences from the same cell for spleen and lymph node cells from donors 2 and 7.

METHODS

Code availability

The code used in this work, including compactor generation and biological interpretation, is available at: https://github.com/salzman-lab/SPLASH-plus/.

Data availability

The fastq files for the Tabula Sapiens data were downloaded from https://tabula-sapiens-portal.ds.czbiohub.org/. Tabula Sapiens gene count table were obtained from the Tabula Sapiens AWS bucket. The Visium lung data was downloaded from SRA with accession ID SRR14851100. Filtered BraCeR contigs for V(D)J analysis were downloaded from: s3://czb-tabula-sapiens/Pilot*/immune-repertoire-analysis/bracer. The five fusion simulated datasets for compactor benchmarking were downloaded from: https://data.broadinstitute.org/Trinity/CTAT_FUSIONTRANS_BENCHMARKING/on_simulated_data/sim_50/reads. The TPM values for simulated fusions were downloaded from: https://data.broadinstitute.org/Trinity/CTAT_FUSIONTRANS_BENCHMARKING/on_simulated_data/sim_50/metadata/sim_50.fusion_TPM_values.dat.

SPLASH overview

SPLASH is a reference-free, annotation-free method that can be directly applied to raw sequencing reads and provide a unified statistical approach for the (co-)detection of various transcript diversification mechanisms (Chaung et al. 2023). Not requiring computational alignment of the reads to a reference genome, a feature commonplace in conventional RNA expression analysis methods, SPLASH can bypass inherent biases and blindspots in aligners. SPLASH takes sequencing files (FASTQ or FASTA) for input samples, and parses sequencing reads to find specific k-mers (substrings of length k) called *anchors* that are followed by a set diverse k-mers called *targets* when all reads from all samples are considered. In our analysis, we set the lengths for both anchors and targets to 27. SPLASH then builds a contingency table for each extracted anchor containing the read counts per sample for each target of the anchor. Using these contingency tables, SPLASH then performs a statistical test for each anchor and computes a closed-form valid p-value under the null hypothesis that the target frequencies for each anchor is independent of sample identity (Chaung et al. 2023), i.e., a significant p-value implies that the anchor has sample-dependent target distribution. The test statistic is constructed through random partitioning of the samples, and using random hash functions to map each target

to a random value in [0,1]. As the test statistic is computed for different random choices of sample partitioning and random hashes per split, the p-value is Bonferroni-corrected to account for multiple testing across random sample partitioning and number of generated random hashes per split.

SPLASH runs

Overall19 tissues and 12 donors from the Tabula Sapiens dataset (Tabula Sapiens Consortium* et al. 2022) that have been profiled by SmartSeq2 were used for our analysis (Suppl. Figure 2). SPLASH was run in unsupervised mode on each donor and tissue separately. For donor tissues with >400 cells, we randomly selected 400 annotated cells to ensure implicit cell count normalization (and approximately read depth). Eight donor-tissues had fewer than 400 cells: donor 2 trachea (119 cells), donor 5 eye (134 cells), donor 1 blood (138 cells), donor 4 tongue (209 cells), donor 12 heart (277 cells), donor 3 eye (291 cells), donor 6 trachea (358 cells), donor 2 kidney (370 cells). In total, we ran SPLASH on 13,500 SmartSeq2 cells from 136 cell types. SPLASH was run with default parameters except for the number of random partitions of input cells (set to 300) and number of random hashes per partition (set to 10). Anchors with >50 reads in >10 cells that have SPLASH's p-value < 0.05 and SPLASH's effect_size > 0.2 were called as significant anchors.

Compactor generation

To generate compactors for each anchor called as significant by SPLASH, first the input FASTQ files are searched for each anchor and the reads with exact match to the anchor are collected. The portion of the read that is upstream (left) of the anchor is clipped, i.e., all the actual read sequences used for assembly start with the anchor sequence. Starting from the first position after the anchor (since there is no variation within the anchor, we start from this position), the frequency of each nucleotide is computed across the collected reads for the anchor and a new branch corresponding to a nucleotide whose frequency exceeds a certain threshold is created. To create a new branch, we used the criterion that the nucleotide is in >10% of collected reads (if its count is >20) or in >80% of collected reads (if its count is >5). These thresholds are chosen based on the typical sequencing error rate and that we desire that the creation of a new branch due to sequencing error be highly unlikely, which is supported by our statistical analysis in the following section. Once a branch for a nucleotide is created at a given position, the subset of reads representing that nucleotide at that position is propagated to that branch (Figure 1B). Corresponding to each position of a branch, the *compactor* sequence is defined as the anchor sequence extended by the nucleotides added at each position from the start of extension up to that position of the branch (Figure 1B). If only one nucleotide satisfies the frequency criteria, no extra branch is created, all reads are propagated to the subsequent iteration in the current branch, and the nucleotide with the highest frequency is appended to the compactor sequence. This process is performed recursively until a user-specified number of recursions is reached (thus fixing compactor length for each set) or until the set of reads decreases to a user-specified read count threshold. When all compactors are generated for an anchor, each compactor is reported with the count and set of reads representing it exactly.

Compactor sequencing error model

Consider the compactor generation process, where we aim to decide on the nucleotide at a given position with N reads. Let $X_i \in \{A, T, C, G\}, 1 \leq i \leq N$ denote the nucleotide for read i at the given position. Under the null hypothesis, all variation in X_i is due to sequencing error. Assuming that sequencing errors occur independently and uniformly across the sequenced regions at rate ϵ, considering without loss of generality that A is the ground truth nucleotide, we model the probability of observing each nucleotide in each read as:

$$X_i = \begin{cases} A & \text{probability } 1-\epsilon, \\ T & \text{probability } \epsilon/3, \\ C & \text{probability } \epsilon/3, \\ G & \text{probability } \epsilon/3. \end{cases}$$

For each noisy nucleotide that is due to sequencing error (nucleotides other than A), the number of reads can be modeled as a binomial random variable with success probability $\frac{\epsilon}{3}$ and N trials denoted by $\mathrm{Bin}\left(N, \frac{\epsilon}{3}\right)$. The probability of observing each noisy nucleotide $\{T, C, G\}$ at least d times can be computed as the upper tail of $\mathrm{Bin}\left(N, \frac{\epsilon}{3}\right)$, which can be upper bounded using Chernoff bound (Arratia and Gordon 1989) as follows:

$$\mathbb{P}\left(\mathrm{Bin}(N, \epsilon/3) \geq d\right) \leq \exp\left(-N\mathrm{D}_{\mathrm{KL}}\left(\frac{d}{N}||\frac{\epsilon}{3}\right)\right),$$

where $\mathrm{D}_{\mathrm{KL}}(p||q)$ denotes the Kullback–Leibler divergence between two independent bernoulli random variables with heads probabilities p and q:

$$D_{\mathrm{KL}}\left(p \mid\mid q\right) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right).$$

We can now compute a closed-form null probability for creating a new branch due to sequencing error at a given position with N reads and sequencing error rate ϵ when the criterion for generating a new branch for a nucleotide is to observe the nucleotide at that position in at least 10% of reads (i.e., $d = \frac{N}{10}$). We use a conservative estimate of 1% for sequencing error rate. NovaSeq 6000 sequencing platform, the machine used to generate the Tabula Sapiens dataset, has a median error rate of 0.11% (Stoler and Nekrutenko 2021) and almost all studied datasets generated by NovaSeq 6000 had an error rate below 1% (Figure 2A in (Stoler and Nekrutenko 2021)). Under such sequencing error rate assumption, the null probability of creating a new branch for $N = 50, 1000, 1000$ is 0.012, 1.59×10^{-4}, and 1.23×10^{-38}, respectively, suggesting that it is highly unlikely to generate a new branch due to sequencing error when we consider a threshold of 10% for nucleotide frequency.

Compactors benchmarking

We considered five fusion simulated datasets (sim_adipose, sim_brain, sim_colon, sim_heart, and sim_testis), each having 500 simulated fusions. To obtain the anchors to be used as seeds for compactor generation, we downlaoded fusion sequences from: (https://data.broadinstitute.org/Trinity/CTAT_FUSIONTRANS_BENCHMARKING/on_simulated_data/simulated_fusion_transcript_sequences/). We then used the provided breakpoint positions in the downloaded fusion to obtain the seed anchor for each fusion as the sequence from `breakpoint_pos-26` to `breakpoint_pos`. We then employed these anchors and ran compactors on the fastq files (we used R1 fastq files for compactor generation) to obtain the compactor sequences. After generating compactors, the confirmation of fusion evidence in a compactor was based on finding a match for the junctional 20-mer of the fusion in the compactor. We should note that we discarded 185 fusions from our benchmarking analysis that did not have reads in R1 FASTQ files for their corresponding seed anchors.

Reference-free classification of compactors into biologically-relevant categories

To increase interpretability of the called anchors and to facilitate targeted downstream analysis for specific applications, we designed a biological interepretaion step in SPLASH+ to assign a biologically-relevant RNA diversifying event to each anchor based on the features directly derived from its compactors. We consider six categories for anchors (Suppl. Figure 2): splicing, internal splicing, base pair change, 3'UTR, centromere, and repeat. If an anchor is not assigned to any of these categories, it will be categorized as an unclassified anchor. We used a hybrid approach to assign classes to anchors, with some classes assigned independently of reference genome alignment (e.g., internal splicing and base pair change) and others assigned based on alignment (e.g., splicing, 3'UTR, centromere, and repeat). As each anchor might be qualified for more than one class, we prioritize classes in the following order: splicing, internal splicing, base pair change, 3'UTR,

centromere, repeat. To classify anchors, we consider only the top two most abundant compactors for each anchor (i.e., those with the highest fraction of anchor reads). If one compactor is longer than the other, we consider its substring equal in length to the shorter compactor. We then compute two different string metrics: Hamming distance and Levenstein distance. We should note that both Levenstein and Hamming distance for the two compactors are computed after removing the anchor sequence. Anchors with the same Hamming and Levenstein distances are classified as base pair change as this criterion suggests that only substitutions (i.e., nucleotide changes) account for the difference between the two compactors. We also utilize the sequence of Levenstein operations (comprising insertions, deletions, and substitutions) and classify an anchor as "internal splicing" if the Levenstein distance is less than the sum of of the longest stretch of deletions (*Run_D*) and insertions (*Run_I*) plus one (Suppl. Figure 2).

Reference-based classification

Other anchor categories (such as splicing) are assigned bsed on the alignment of compactors to the reference genome. To identify anchors potentially explained by alternative splicing, we align the two compactors for each anchor to the T2T human reference genome using STAR v2.7.5.a (Dobin et al. 2013) with the following parameters:

```
--twopassMode Basic --alignIntronMax 1000000 --chimJunctionOverhangMin 10
--chimSegmentReadGapMax 0 --chimOutJunctionFormat 1 --chimSegmentMin 12
--chimScoreJunctionNonGTAG -4 --chimNonchimScoreDropMin 10 --outSAMtype SAM
--chimOutType SeparateSAMold --outSAMunmapped None --clip3pAdapterSeq AAAAAAAAA
--outSAMattributes NH HI AS nM NM
```

We then extract information about the mapping flag, chromosome, coordinate, CIGAR string, and number of mismatches from the STAR BAM file (1st, 2nd, 3rd, 4th, 6th, and 16th columns). If at least one of the two compactors involves a split alignment, and the Hamming distance and Levenstein distance are not equal (as this would indicate a base pair change), we classify the anchor as "splicing". This classification is based on the presence of a splice junction in at least one compactor and that the difference between the compactor sequences cannot be explained by simple substitutions. Note that both compactors should overlap with the same gene to be considered as splicing anchor.

Anchors not classified as splicing, base pair change, or internal splicing are considered for further classification. We intersect compactor mapping positions with annotated 3'UTRs, centromere satellite elements (from T2T CenSat database), and repetitive elements (from RepeatMasker database) using BEDTools v2.25.0 (Quinlan 2014) intersect command. Anchors are then categorized to one of these categories based on this intersection (Suppl. Figure 2).

Comparison to SpliZ for alternative splicing

SpliZ is a statistical method for detecting genes with cell type-specific alternative splicing in scRNA-Seq (Olivieri, Dehghannasiri, and Salzman 2022). It assigns a single score to each pair of cell and gene and is reference-dependent in the sense that it needs the split reads mapping to the splice junctions of the gene. To compare SPLASH+ with SpliZ (as a state-of-the-art reference-dependent method for alternative splicing), we ran SpliZ on the same Tabula Sapiens dataset. To obtain high confidence splice junctions needed for SpliZ analysis, we first aligned reads to human reference genome using STAR and then ran SICILIAN (Dehghannasiri, Olivieri, and Salzman 2020), a statistical wrapper for detecting high-confidence splice junctions from spliced aligners, on the STAR BAM files. We applied SpliZ to the reads aligned to the detected splice junctions. SpliZ was ran on each donor separately and to avoid calling genes with tissue-specific splicing rather than cell type-specific splicing, its statistical test was performed separately across cell types within each tissue from that donor. Both SpliZ and SICILIAN were run with default parameters. After running SpliZ and obtaining the list of genes with cell-type-specific splicing for each donor and tissue, we compared the list of significant genes found by SPLASH+ and SpliZ. Since SPLASH calls are the anchor level, not gene-level, for a

consistent comparison with SpliZ (which reports at the gene-level), we adopt the convention that a gene is identified as significant by SPLASH+ if it contains at least one “splicing” anchor.

Querying SRA human RNA-Seq database using Pebblescout for novel splicing and RNA editing

Pebblescout (https://pebblescout.ncbi.nlm.nih.gov/) is a kmer-based query tool designed to take input sequences and report studies (RNA-Seq sequencing samples) within its indexed database that share matches with the provided input. Pebblescout randomly samples a 25-mer from every 42-mer in the query sequence, guaranteeing a match between the query and the study if the match is at least 42bp long. Given that Pebblescout reports matches based on any 25-mer in the input sequence, for each queried *CD47* and *HSP90AA1* isoform a 42-mer (the minimum acceptable length by Pebblescout for a query) is constructed by taking 21 bps from both the 5’ and 3’ sides of the junctions. For queries regarding *ARPC2* (Figure 6C), we extended each 20-mer by 22 bps. We further aligned the reported 25-mer matches by Pebblescout to the reference genome using Bowtie2 v2.2.1 to confirm their unique mapping to the *CD47*, *HSP90AA1*, and *ARPC2* loci.

To create a control set of unannotated junctions for novel *CD47* and *HSP90AA1* junctions, we first extracted exonic 21 bps leading to each annotated 5’ and 3’ exon boundary for each gene. We then concatenated the 21 bps from each 5’ splice site with those from its downstream 3’ splice sites to form 42-mers. These junctions that are not part of the annotated gene structure were then used as a control set for evaluating the prevalence of detected unannotated junctions by SPLASH+.

Analysis of Alu Element Insertion

To identify compactors with Alu insertions, we required that the anchor uniquely map to the human genome and that the compactor have >25 bps with a soft-clipped alignment to transposable elements, none of which must map to Illumina adaptors. To identify compactors with a novel exon skipping event in *PSPH*, compactors were filtered for anchors with the highest priority annotation to the human genome, compactors that align uniquely to the human genome, and compactors whose genes align to *PSPH*.

V(D)J rearrangement analysis and comparison to BraCeR

We defined immunoglobulin compactors (being referred to as “IG-compactors”) as the compactors mapped to an immunoglobulin gene *IGH*, *IGK*, or *IGV* by STAR. We further defined immunoglobulin-anchors (“IG-anchors”) as anchors with >20% of reads associated with IG-compactors. The unaligned compactors of an IG-anchor are still considered IG-compactors through “annotation-by-association”. To compare the SPLASH+ with BraCeR (Lindeman et al. 2018), we considered a stringent criterion for IG-compactors where we first annotated IG-compactors with AssignGenes.py (ChangeO v1.3.0) (Gupta et al. 2015) and IgBLAST v1.21.0 (Ye et al. 2013) and considered only those IG-compactors that had both variable (V) and joining (J) immunoglobulin gene segments identified thorugh IgBLAST. The IG-compactors are annotated as in-frame by IgBLAST if the last triplet of the annotated V gene is in-frame with the first triplet of the annotated J gene. Also, to test whether SPLASH+ provides support for the detection of the same B cell receptors (BCRs) as BraCeR, for each cell we computed the minimum Hamming distance between the compactors of IG-anchors and the filtered BraCeR contigs.

Repeat expansion analysis

For each cell, we obtain total read count for each repeat length by summing the reads across all compactors associated with that repeat length. To characterize the repeat length for the two dominant alleles per cell, we refer to the repeat length with the highest read count as the “primary repeat length” and the repeat length with the second-highest read count as the “secondary repeat length”.

Analyzing the abundance of the third most abundant target under a two-true-target assumption

We statistically quantify how unlikely it is that we observe a high count for the third most common target when there are only two ground truth targets and all others are due to sequencing error. This could arise when the anchor corresponds to alternative splicing involving two isoforms (each target represents an isoform) or corresponds to an SNP (one target represents the transcript without SNP and the other target represents the transcript with SNP). We assume that the anchors are observed without sequencing error. We also make the same assumption of uniformity and independence for sequencing error rate ϵ as discussed in the section on analyzing sequencing errors in compactor construction. Under these assumptions, we can provide p-values on the abundance of the third most frequent target, showing that it is very unlikely to observe a high count for the third most abundant target when only two ground truth targets exist, with all others attributed to sequencing errors.

Let M be the total number of counts, and L be the target length. The two ground truth targets are denoted by y_1 and y_2. For each possible target (a sequence of length L from A/G/C/T) $y \in \{A, C, T, G\}^L$, we define $\ell(y)$ to be the minimum of the hamming distance between y and y_1, and y and y_2, i.e., $\ell(y) = \min(d_{\mathrm{H}}(y, y_1), d_{\mathrm{H}}(y, y_2))$, where d_{H} denotes the hamming distance between two strings. Note that $l(y)$ is always between 1 and L. We can now compute the probability of observing a noisy target p_y due to sequencing error using its minimum hamming distance $l(y)$ relative to the two true targets. For this target y, assuming that targets are generated randomly from y_1, y_2 and that $\mathbb{P}(\text{unerrored target } y_1) = q$:

$$\begin{aligned}
\mathbb{P}(\text{observe y}) &= \mathbb{P}(\text{observe y}|\text{unerrored target is } y_1)\mathbb{P}(\text{unerrored target is } y_1) \\
&\qquad + \mathbb{P}(\text{observe y}|\text{unerrored target is } y_2)\mathbb{P}(\text{unerrored target is } y_2) \\
&= q\left((1-\epsilon)^{k-d_{\mathrm{H}}(y,y_1)}(\epsilon/3)^{d_{\mathrm{H}}(y,y_1)}\right) + (1-q)\left((1-\epsilon)^{k-d_{\mathrm{H}}(y,y_2)}(\epsilon/3)^{d_{\mathrm{H}}(y,y_2)}\right) \\
&\leq (1-\epsilon)^{k-\ell}(\epsilon/3)^{\ell}.
\end{aligned}$$

We used the law of total probability for the first equality, and the last upper bound is because both hamming distances $d_H(\cdot)$ are greater than ℓ.

We can now compute probability of observing more than k counts for the third most abundant target. Let $X_1, X_2, \ldots, X_{4^L}$ to be the counts of the 4^L possible targets. We consider X_1 and X_2 to be the counts for the top two true targets y_1 and y_2, respectively. Under the null, $X_3, \ldots, X_k$ are due to sequencing error, and we assume these are arbitrarily ordered (e.g., lexicographically), with the target sequence corresponding to X_i being y_i. Therefore, the counts for the third target is obtained as $\max(X_3, \ldots, X_{4^L})$. We can now compute the probability of at least K counts for the third target as follows:

$$
\begin{aligned}
\mathbb{P}\left(\max(X_3,\ldots,X_{4^L}) \geq K\right) &= \mathbb{P}\left(\bigcup_{i=3}^{4^L}\{X_i \geq K\}\right) \\
&= \sum_{i=3}^{4^L}\mathbb{P}\left(X_i \geq K\right) \\
&= \sum_{i=3}^{4^L}\mathbb{P}\left(\mathrm{Bin}(M, p_{\ell(y_i)}) \geq K\right) \\
&= 2\sum_{\ell=1}^{L}\binom{L}{\ell}3^{\ell}\mathbb{P}\left(\mathrm{Bin}(M,(1-\epsilon)^{L-\ell}(\epsilon/3)^{\ell}) \geq K\right) \\
&\leq 2\sum_{\ell=1}^{L}\binom{L}{\ell}3^{\ell}\exp\left(-MD_{\mathrm{KL}}\left(\frac{K}{M}\,\middle\|\,(1-\epsilon)^{L-\ell}(\epsilon/3)^{\ell}\right)\right)
\end{aligned}
$$

The first equality implies that in order for the maximum count of noisy targets be greater than K, at least one noisy target must have counts greater than K. The third equality states that X_i can be modeled as a Binomial random variable $\mathrm{Bin}(M, p_{\ell(y_i)})$ using the total count and probability of observing y_i which is a function of its minimum hamming distance to the top two targets. The fourth equality is because there are $2\binom{L}{\ell}3^{\ell}$ targets with minimum hamming distance ℓ. For y_1, there are exactly $\binom{L}{\ell}3^{\ell}$ targets of hamming distance exactly ℓ from y_1 (choose ℓ out of the L possible locations, and which of the 3 nucleotides to observe at each of the ℓ locations), and therefore there are at most $2\binom{L}{\ell}3^{\ell}$ targets with minimum hamming distance ℓ between y_1, y_2. The final inequality utilizes a Chernoff bound (Arratia and Gordon 1989) for the binomial distribution. This bound we obtained for the count of the third target K decays very quickly as a function of K. For example, for $M = 50$ and $\epsilon = 0.01$, $K = 4$ yields $.01$, $K = 5$ yields $1E-5$, and $K = 7$ yields $1E-10$. We have provided a rigorous statistical framework to show that it is very unlikely to observe many counts for noisy targets that are due to sequencing error.

Bibliography

Abdel-Fatah, Tarek M. A., Robert C. Rees, A. Graham Pockley, Paul Moseley, Graham R. Ball, Stephen Y. T. Chan, Ian O. Ellis, and Amanda K. Miles. 2017. “The Localization of Pre mRNA Splicing Factor PRPF38B Is a Novel Prognostic Biomarker That May Predict Survival Benefit of Trastuzumab in Patients with Breast Cancer Overexpressing HER2.” *Oncotarget* 8 (68): 112245–57.

Altemose, Nicolas, Glennis A. Logsdon, Andrey V. Bzikadze, Pragya Sidhwani, Sasha A. Langley, Gina V. Caldas, Savannah J. Hoyt, et al. 2022. “Complete Genomic and Epigenetic Maps of Human Centromeres.” *Science* 376 (6588): eabl4178.

Al-Zeer, Munir A., Mariola Dutkiewicz, Annekathrin von Hacht, Denise Kreuzmann, Viola Röhrs, and Jens Kurreck. 2019. “Alternatively Spliced Variants of the 5’-UTR of the ARPC2 mRNA Regulate Translation by an Internal Ribosome Entry Site (IRES) Harboring a Guanine-Quadruplex Motif.” *RNA Biology*. https://doi.org/10.1080/15476286.2019.1652524.

Arratia, R., and L. Gordon. 1989. “Tutorial on Large Deviations for the Binomial Distribution.” *Bulletin of Mathematical Biology* 51 (1): 125–31.

Arzalluz-Luque, Ángeles, and Ana Conesa. 2018. “Single-Cell RNAseq for the Study of Isoforms-How Is That Possible?” *Genome Biology* 19 (1): 110.

Baharav, Tavor Z., David Tse, and Julia Salzman. 2023. “OASIS: An Interpretable, Finite-Sample Valid Alternative to Pearson’s X2 for Scientific Discovery.” *bioRxiv*. https://doi.org/10.1101/2023.03.16.533008.

Barrett, T. B., P. Sampson, G. K. Owens, S. M. Schwartz, and E. P. Benditt. 1983. “Polyploid Nuclei in Human Artery Wall Smooth Muscle Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 80 (3): 882–85.

Biterge, Burcu, Florian Richter, Gerhard Mittler, and Robert Schneider. 2014. “Methylation of Histone H4 at Aspartate 24 by Protein L-Isoaspartate O-Methyltransferase (PCMT1) Links Histone Modifications with Protein Homeostasis.” *Scientific Reports* 4 (October): 6674.

Bonnal, Sophie C., Irene López-Oreja, and Juan Valcárcel. 2020. “Roles and Mechanisms of Alternative Splicing in Cancer — Implications for Care.” *Nature Reviews Clinical Oncology*. https://doi.org/10.1038/s41571-020-0350-x.

Buen Abad Najar, Carlos F., Prakruthi Burra, Nir Yosef, and Liana F. Lareau. 2022. “Identifying Cell State-Associated Alternative Splicing Events and Their Coregulation.” *Genome Research*, July. https://doi.org/10.1101/gr.276109.121.

Buen Abad Najar, Carlos F., Nir Yosef, and Liana F. Lareau. 2020. “Coverage-Dependent Bias Creates the Appearance of Binary Splicing in Single Cells.” *eLife* 9 (June). https://doi.org/10.7554/eLife.54603.

Buratti, Emanuele, and Francisco E. Baralle. 2001. “Characterization and Functional Implications of the RNA Binding Properties of Nuclear Factor TDP-43, a Novel Splicing Regulator ofCFTR Exon 9.” *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.m104236200.

Chaung, Kaitlin, Tavor Z. Baharav, George Henderson, Ivan N. Zheludev, Peter L. Wang, and Julia Salzman. 2023. “SPLASH: A Statistical, Reference-Free Genomic Algorithm Unifies Biological Discovery.” *Cell* 186 (25): 5440–56.e26.

Chen, Yuhao, and Xiaowei Wang. 2020. “miRDB: An Online Database for Prediction of Functional microRNA Targets.” *Nucleic Acids Research* 48 (D1): D127–31.

Cmero, Marek, Breon Schmidt, Ian J. Majewski, Paul G. Ekert, Alicia Oshlack, and Nadia M. Davidson. 2021. “MINTIE: Identifying Novel Structural and Splice Variants in Transcriptomes Using RNA-Seq Data.” *Genome Biology* 22 (1): 296.

Cohen-Fultheim, Roni, and Erez Y. Levanon. 2021. “Detection of A-to-I Hyper-Edited RNA Sequences.” *Methods in Molecular Biology* 2181: 213–27.

Cully, Megan, Jessica Shiu, Roland P. Piekorz, William J. Muller, Susan J. Done, and Tak W. Mak. 2005. “Transforming Acidic Coiled Coil 1 Promotes Transformation and Mammary Tumorigenesis.” *Cancer Research* 65 (22): 10363–70.

Dehghannasiri, Roozbeh, Julia Eve Olivieri, and Julia Salzman. 2020. “Specific Splice Junction Detection in Single Cells with SICILIAN.” *bioRxiv*. https://doi.org/10.1101/2020.04.14.041905.

Ding, Fangyuan, Christina J. Su, Kehuan Kuo Edmonds, Guohao Liang, and Michael B. Elowitz. 2022. “Dynamics and Functional Roles of Splicing Factor Autoregulation.” *Cell Reports* 39 (12): 110985.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21.

Eisenberg, Eli, and Erez Y. Levanon. 2018. “A-to-I RNA Editing — Immune Protector and Transcriptome Diversifier.” *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-018-0006-1.

Estany, Joan, Marc Tor, Daniel Villalba, Lluís Bosch, David Gallardo, Neus Jiménez, Laura Altet, et al. 2007. “Association of CA Repeat Polymorphism at Intron 1 of Insulin-like Growth Factor (IGF-I) Gene with Circulating IGF-I Concentration, Growth, and Fatness in Swine.” *Physiological Genomics* 31 (2): 236–43.

Ezkurdia, Iakes, Jose Manuel Rodriguez, Enrique Carrillo-de Santa Pau, Jesús Vázquez, Alfonso Valencia, and Michael L. Tress. 2015. “Most Highly Expressed Protein-Coding Genes Have a Single Dominant Isoform.” *Journal of Proteome Research*. https://doi.org/10.1021/pr501286b.

Freedman, Adam H., Michele Clamp, and Timothy B. Sackton. 2021. “Error, Noise and Bias in de Novo Transcriptome Assemblies.” *Molecular Ecology Resources* 21 (1): 18–29.

Gordon, Sydney R., Roy L. Maute, Ben W. Dulken, Gregor Hutter, Benson M. George, Melissa N. McCracken, Rohit Gupta, et al. 2017. “PD-1 Expression by Tumour-Associated Macrophages Inhibits Phagocytosis and Tumour Immunity.” *Nature* 545 (7655): 495–99.

Gray, Phillip J., Thomas Prince, Jinrong Cheng, Mary Ann Stevenson, and Stuart K. Calderwood. 2008. “Targeting the Oncogene and Kinome Chaperone CDC37.” *Nature Reviews. Cancer* 8 (7): 491–95.

Gupta, Namita T., Jason A. Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Gur Yaari, and Steven H. Kleinstein. 2015. “Change-O: A Toolkit for Analyzing Large-Scale B Cell Immunoglobulin Repertoire Sequencing Data.” *Bioinformatics* 31 (20): 3356–58.

Haas, Brian J., Alexander Dobin, Bo Li, Nicolas Stransky, Nathalie Pochet, and Aviv Regev. 2019. “Accuracy Assessment of Fusion Transcript Detection via Read-Mapping and de Novo Fusion Transcript Assembly-Based Methods.” *Genome Biology* 20 (1): 213.

Hoter, Abdullah, Marwan E. El-Sabban, and Hassan Y. Naim. 2018. “The HSP90 Family: Structure, Regulation,

Function, and Implications in Health and Disease." *International Journal of Molecular Sciences* 19 (9). https://doi.org/10.3390/ijms19092560.
Iwai, Naomi, Zhijian Zhou, Dennis R. Roop, and Richard R. Behringer. 2008. "Horizontal Basal Cells Are Multipotent Progenitors in Normal and Injured Adult Olfactory Epithelium." *Stem Cells* 26 (5): 1298–1306.
Jaberi, Elham, Emilie Tresse, Kirsten Grønbæk, Joachim Weischenfeldt, and Shohreh Issazadeh-Navikas. 2020. "Identification of Unique and Shared Mitochondrial DNA Mutations in Neurodegeneration and Cancer by Single-Cell Mitochondrial DNA Structural Variation Sequencing (MitoSV-Seq)." *EBioMedicine* 57 (July): 102868.
Kalvari, Ioanna, Eric P. Nawrocki, Joanna Argasinska, Natalia Quinones-Olvera, Robert D. Finn, Alex Bateman, and Anton I. Petrov. 2018. "Non-Coding RNA Analysis Using the Rfam Database." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 62 (1): e51.
Kim, Eunhee, Janine O. Ilagan, Yang Liang, Gerrit M. Daubner, Stanley C-W Lee, Aravind Ramakrishnan, Yue Li, et al. 2015. "SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition." *Cancer Cell* 27 (5): 617–30.
Kino, Tomoshige, Darrell E. Hurt, Takamasa Ichijo, Nancy Nader, and George P. Chrousos. 2010. "Noncoding RNA Gas5 Is a Growth Arrest– and Starvation-Associated Repressor of the Glucocorticoid Receptor." *Science Signaling*. https://doi.org/10.1126/scisignal.2000568.
Kojima, Yoko, Jens-Peter Volkmer, Kelly McKenna, Mete Civelek, Aldons Jake Lusis, Clint L. Miller, Daniel Direnzo, et al. 2016. "CD47-Blocking Antibodies Restore Phagocytosis and Prevent Atherosclerosis." *Nature* 536 (7614): 86–90.
Königs, Vanessa, Camila de Oliveira Freitas Machado, Benjamin Arnold, Nicole Blümel, Anfisa Solovyeva, Sinah Löbbert, Michal Schafranek, et al. 2020. "SRSF7 Maintains Its Homeostasis through the Expression of Split-ORFs and Nuclear Body Assembly." *Nature Structural & Molecular Biology* 27 (3): 260–73.
Kung, Che-Pei, Leonard B. Maggi Jr, and Jason D. Weber. 2018. "The Role of RNA Editing in Cancer Development and Metabolic Disorders." *Frontiers in Endocrinology* 9 (December): 762.
Lareau, Liana F., Maki Inada, Richard E. Green, Jordan C. Wengrod, and Steven E. Brenner. 2007. "Unproductive Splicing of SR Genes Associated with Highly Conserved and Ultraconserved DNA Elements." *Nature* 446 (7138): 926–29.
Lindeman, Ida, Guy Emerton, Lira Mamanova, Omri Snir, Krzysztof Polanski, Shuo-Wang Qiao, Ludvig M. Sollid, Sarah A. Teichmann, and Michael J. T. Stubbington. 2018. "BraCeR: B-Cell-Receptor Reconstruction and Clonality Inference from Single-Cell RNA-Seq." *Nature Methods* 15 (8): 563–65.
Lu, Junjie, and David M. Gilbert. 2007. "Proliferation-Dependent and Cell Cycle Regulated Transcription of Mouse Pericentric Heterochromatin." *The Journal of Cell Biology* 179 (3): 411–21.
Mabin, Justin W., Peter W. Lewis, David A. Brow, and Heidi Dvinge. 2021. "Human Spliceosomal snRNA Sequence Variants Generate Variant Spliceosomes." *RNA* 27 (10): 1186–1203.
Mahmoudabadi, Gita, Tabula Sapiens Consortium, and Stephen R. Quake. 2023. "Single Cell Transcriptomics Reveals the Hidden Microbiomes of Human Tissues." *bioRxiv*. https://doi.org/10.1101/2022.10.11.511790.
Ma, Yiyi, Eric B. Dammer, Daniel Felsky, Duc M. Duong, Hans-Ulrich Klein, Charles C. White, Maotian Zhou, et al. 2021. "Atlas of RNA Editing Events Affecting Protein Expression in Aged and Alzheimer's Disease Human Brain Tissue." *Nature Communications*. https://doi.org/10.1038/s41467-021-27204-9.
Mistry, Jaina, Robert D. Finn, Sean R. Eddy, Alex Bateman, and Marco Punta. 2013. "Challenges in Homology Search: HMMER3 and Convergent Evolution of Coiled-Coil Regions." *Nucleic Acids Research* 41 (12): e121.
Morimoto, Hirofumi, Yasuhiro Hida, Nako Maishi, Hiroshi Nishihara, Yutaka Hatanaka, Cong Li, Yoshihiro Matsuno, Toru Nakamura, Satoshi Hirano, and Kyoko Hida. 2021. "Biglycan, Tumor Endothelial Cell Secreting Proteoglycan, as Possible Biomarker for Lung Cancer." *Thoracic Cancer* 12 (9): 1347–57.
Mourtada-Maarabouni, M., M. R. Pickard, V. L. Hedge, F. Farzaneh, and G. T. Williams. 2009. "GAS5, a Non-Protein-Coding RNA, Controls Apoptosis and Is Downregulated in Breast Cancer." *Oncogene* 28 (2): 195–208.
Ni, Julie Z., Leslie Grate, John Paul Donohue, Christine Preston, Naomi Nobida, Georgeann O'Brien, Lily Shiue, Tyson A. Clark, John E. Blume, and Manuel Ares Jr. 2007. "Ultraconserved Elements Are Associated with Homeostatic Control of Splicing Regulators by Alternative Splicing and Nonsense-Mediated Decay." *Genes & Development* 21 (6): 708–18.
O'Donnell-Luria, Anne H., Lynn S. Pais, Víctor Faundes, Jordan C. Wood, Abigail Sveden, Victor Luria, Rami Abou Jamra, et al. 2019. "Heterozygous Variants in KMT2E Cause a Spectrum of Neurodevelopmental Disorders and Epilepsy." *American Journal of Human Genetics* 104 (6): 1210–22.

Ola, Roxana, Sylvie Lefebvre, Karl-Heinz Braunewell, Kirsi Sainio, and Hannu Sariola. 2012. “The Expression of Visinin-like 1 during Mouse Embryonic Development.” *Gene Expression Patterns: GEP* 12 (1-2): 53–62.

Olivieri, Julia Eve, Roozbeh Dehghannasiri, and Julia Salzman. 2022. “The SpliZ Generalizes ‘Percent Spliced in’ to Reveal Regulated Splicing at Single-Cell Resolution.” *Nature Methods* 19 (3): 307–10.

Olivieri, Julia Eve, Roozbeh Dehghannasiri, Peter Wang, Sori Jang, Antoine de Morree, Serena Y. Tan, Jingsi Ming, et al. 2021. “RNA Splicing Programs Define Tissue Compartments and Cell Types at Single Cell Resolution.” *bioRxiv*. https://doi.org/10.1101/2021.05.01.442281.

Pan, Qun, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. 2008. “Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing.” *Nature Genetics* 40 (12): 1413–15.

Payer, Lindsay M., Jared P. Steranka, Maria S. Kryatova, Giacomo Grillo, Mathieu Lupien, Pedro P. Rocha, and Kathleen H. Burns. 2021. “Insertion Variants Alter Gene Transcript Levels.” *Genome Research* 31 (12): 2236–48.

Perissi, Valentina, Kristen Jepsen, Christopher K. Glass, and Michael G. Rosenfeld. 2010. “Deconstructing Repression: Evolving Models of Co-Repressor Action.” *Nature Reviews. Genetics* 11 (2): 109–23.

Picardi, Ernesto, Anna Maria D’Erchia, Claudio Lo Giudice, and Graziano Pesole. 2017. “REDIportal: A Comprehensive Database of A-to-I RNA Editing Events in Humans.” *Nucleic Acids Research* 45 (D1): D750–57.

Probst, Aline V., Ikuhiro Okamoto, Miguel Casanova, Fatima El Marjou, Patricia Le Baccon, and Geneviève Almouzni. 2010. “A Strand-Specific Burst in Transcription of Pericentric Satellites Is Required for Chromocenter Formation and Early Mouse Development.” *Developmental Cell* 19 (4): 625–38.

Quinlan, Aaron R. 2014. “BEDTools: The Swiss-Army Tool for Genome Feature Analysis.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 47 (September): 11.12.1–34.

Raihan, Obayed, Afrina Brishti, Qin Li, Qilun Zhang, Dingfeng Li, Xiaohui Li, Qingyang Zhang, et al. 2019. “SRSF11 Loss Leads to Aging-Associated Cognitive Decline by Modulating LRP8 and ApoE.” *Cell Reports*. https://doi.org/10.1016/j.celrep.2019.06.002.

Riley, Donald E., and John N. Krieger. 2009. “UTR Dinucleotide Simple Sequence Repeat Evolution Exhibits Recurring Patterns Including Regulatory Sequence Motif Replacements.” *Gene* 429 (1-2): 80–86.

Schmucker, D., J. C. Clemens, H. Shu, C. A. Worby, J. Xiao, M. Muda, J. E. Dixon, and S. L. Zipursky. 2000. “Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity.” *Cell* 101 (6): 671–84.

Schroeder, Harry W., Jr. 2006. “Similarity and Divergence in the Development and Expression of the Mouse and Human Antibody Repertoires.” *Developmental and Comparative Immunology* 30 (1-2): 119–35.

Sherman, Rachel M., Juliet Forman, Valentin Antonescu, Daniela Puiu, Michelle Daya, Nicholas Rafaels, Meher Preethi Boorgula, et al. 2019. “Assembly of a Pan-Genome from Deep Sequencing of 910 Humans of African Descent.” *Nature Genetics* 51 (1): 30–35.

Shinde, D. 2003. “Taq DNA Polymerase Slippage Mutation Rates Measured by PCR and Quasi-Likelihood Analysis: (CA/GT)n and (A/T)n Microsatellites.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkg178.

Shkreta, Lulzim, Aurélie Delannoy, Anna Salvetti, and Benoit Chabot. 2021. “SRSF10: An Atypical Splicing Regulator with Critical Roles in Stress Response, Organ Development, and Viral Replication.” *RNA* 27 (11): 1302–17.

Sproviero, William, Aleksey Shatunov, Daniel Stahl, Maryam Shoai, Wouter van Rheenen, Ashley R. Jones, Safa Al-Sarraj, et al. 2017. “ATXN2 Trinucleotide Repeat Length Correlates with Risk of ALS.” *Neurobiology of Aging* 51 (March): 178.e1–178.e9.

Stoler, Nicholas, and Anton Nekrutenko. 2021. “Sequencing Error Profiles of Illumina Sequencing Instruments.” *NAR Genomics and Bioinformatics* 3 (1): lqab019.

Swanson, Lucas, Gordon Robertson, Karen L. Mungall, Yaron S. Butterfield, Readman Chiu, Richard D. Corbett, T. Roderick Docking, et al. 2013. “Barnacle: Detecting and Characterizing Tandem Duplications and Fusions in Transcriptome Assemblies.” *BMC Genomics* 14 (August): 550.

Tabula Sapiens Consortium*, Robert C. Jones, Jim Karkanias, Mark A. Krasnow, Angela Oliveira Pisco, Stephen R. Quake, Julia Salzman, et al. 2022. “The Tabula Sapiens: A Multiple-Organ, Single-Cell Transcriptomic Atlas of Humans.” *Science* 376 (6594): eabl4896.

Takahashi, Nobuhiro, Noboru Sasagawa, Koichi Suzuki, and Shoichi Ishiura. 2000. “The CUG-Binding Protein Binds Specifically to UG Dinucleotide Repeats in a Yeast Three-Hybrid System.” *Biochemical and Biophysical Research Communications*. https://doi.org/10.1006/bbrc.2000.3694.

Travaglini, Kyle J., Ahmad N. Nabhan, Lolita Penland, Rahul Sinha, Astrid Gillich, Rene V. Sit, Stephen Chang, et al. 2020. “A Molecular Cell Atlas of the Human Lung from Single-Cell RNA Sequencing.” *Nature* 587 (7835): 619–25.

Wang, Rui, Jingyun Li, Xin Zhou, Yunuo Mao, Wendong Wang, Shuai Gao, Wei Wang, et al. 2022. “Single-Cell Genomic and Transcriptomic Landscapes of Primary and Metastatic Colorectal Cancer Tumors.” *Genome Medicine* 14 (1): 93.

Watson, C. T., and F. Breden. 2012. “The Immunoglobulin Heavy Chain Locus: Genetic Variation, Missing Data, and Implications for Human Disease.” *Genes and Immunity* 13 (5): 363–73.

Werner, Stephan, Lukas Schmidt, Virginie Marchand, Thomas Kemmer, Christoph Falschlunger, Maksim V. Sednev, Guillaume Bec, et al. 2020. “Machine Learning of Reverse Transcription Signatures of Variegated Polymerases Allows Mapping and Discrimination of Methylated Purines in Limited Transcriptomes.” *Nucleic Acids Research* 48 (7): 3734–46.

Westoby, Jennifer, Pavel Artemov, Martin Hemberg, and Anne Ferguson-Smith. 2020. “Obstacles to Detecting Isoforms Using Full-Length scRNA-Seq Data.” *Genome Biology* 21 (1): 74.

Xu, Caipeng, Xiaohua Chen, Xuetian Zhang, Dapeng Zhao, Zhihui Dou, Xiaodong Xie, Hongyan Li, et al. 2021. “RNA-Binding Protein 39: A Promising Therapeutic Target for Cancer.” *Cell Death Discovery* 7 (1): 214.

Yagi, Takeshi. 2008. “Clustered Protocadherin Family.” *Development, Growth & Differentiation* 50 Suppl 1 (June): S131–40.

Ye, Jian, Ning Ma, Thomas L. Madden, and James M. Ostell. 2013. “IgBLAST: An Immunoglobulin Variable Domain Sequence Analysis Tool.” *Nucleic Acids Research* 41 (Web Server issue): W34–40.

Ying, Zhe, Hyae Ran Byun, Qingying Meng, Emily Noble, Guanglin Zhang, Xia Yang, and Fernando Gomez-Pinilla. 2018. “Biglycan Gene Connects Metabolic Dysfunction with Brain Disorder.” *Biochimica et Biophysica Acta, Molecular Basis of Disease* 1864 (12): 3679–87.

Yum, Kevin, Eric T. Wang, and Auinash Kalsotra. 2017. “Myotonic Dystrophy: Disease Repeat Range, Penetrance, Age of Onset, and Relationship between Repeat Size and Phenotypes.” *Current Opinion in Genetics & Development* 44 (June): 30–37.

Zeng, Hongkui. 2022. “What Is a Cell Type and How to Define It?” *Cell* 185 (15): 2739–55.

Zheng, Hongyu, Cong Ma, and Carl Kingsford. 2022. “Deriving Ranges of Optimal Estimated Transcript Expression due to Nonidentifiability.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 29 (2): 121–39.

Zuehlke, Abbey D., Kristin Beebe, Len Neckers, and Thomas Prince. 2015. “Regulation and Function of the Human HSP90AA1 Gene.” *Gene* 570 (1): 8–16.

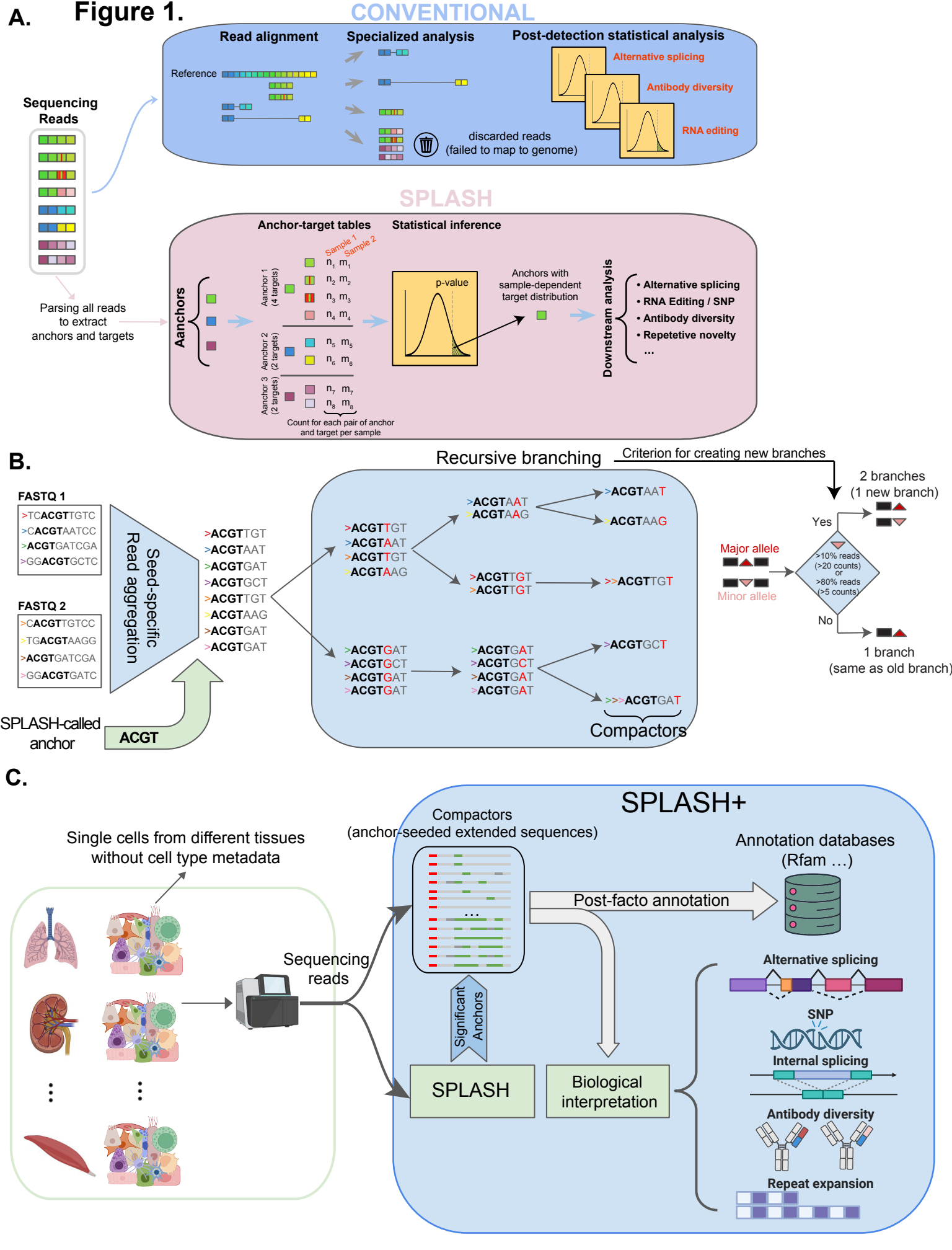
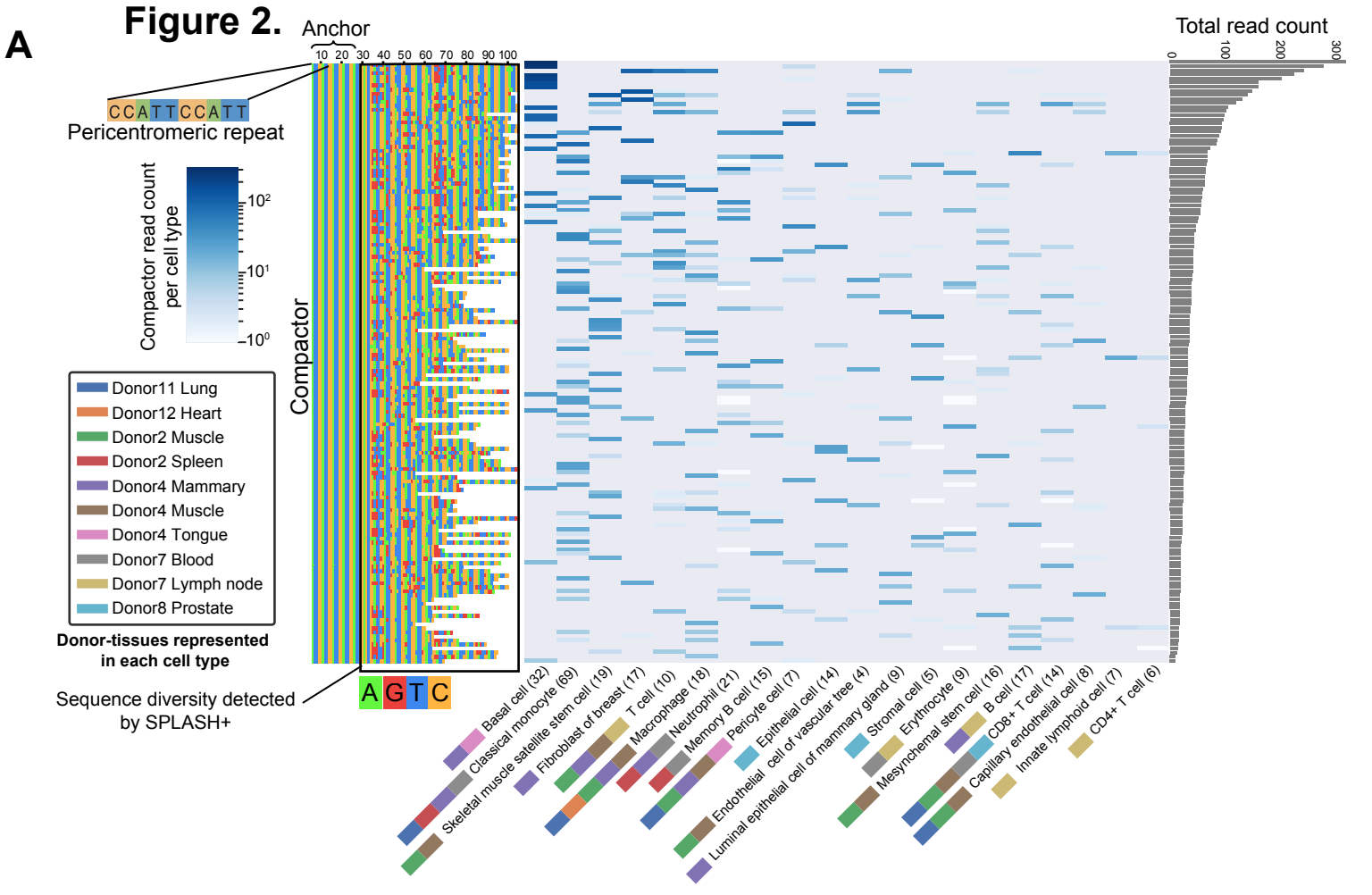
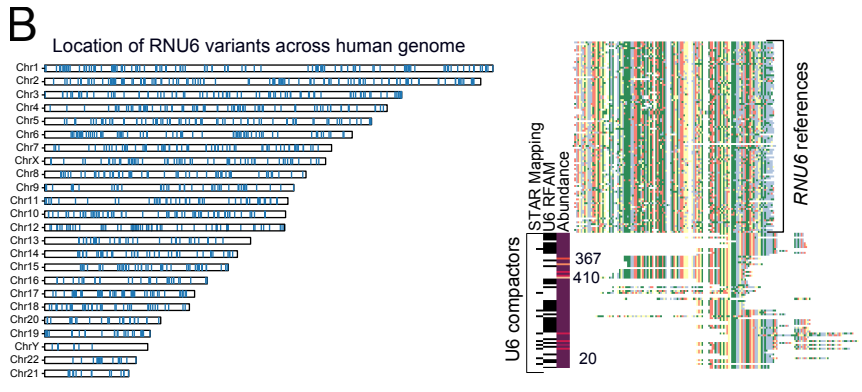
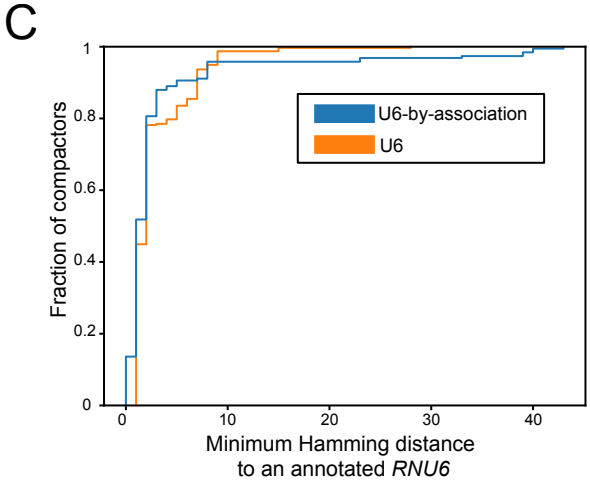
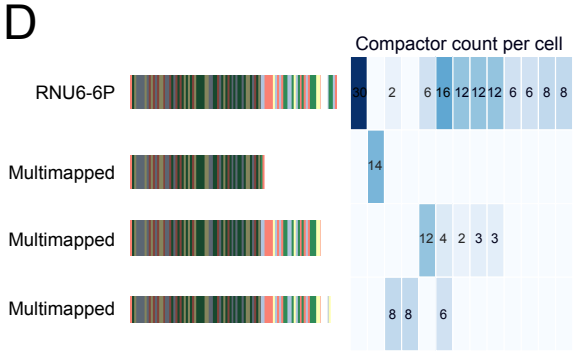

A.
Figure 1.
CONVENTIONAL
Read alignment
Specialized analysis
Post-detection statistical analysis
Reference
Alternative splicing
Antibody diversity
RNA editing
discarded reads
(failed to map to genome)
Sequencing Reads
Parsing all reads to extract anchors and targets
SPLASH
Anchor-target tables
Statistical inference
Sample 1
Sample 2
Aanchors
Aanchor 1 (4 targets)
Aanchor 2 (2 targets)
Aanchor 3 (2 targets)
Count for each pair of anchor and target per sample
p-value
Anchors with sample-dependent target distribution
Downstream analysis
• Alternative splicing
• RNA Editing / SNP
• Antibody diversity
• Repetetive novelty
...
B.
FASTQ 1
FASTQ 2
Seed-specific Read aggregation
SPLASH-called anchor
ACGT
Recursive branching
Criterion for creating new branches
2 branches (1 new branch)
Yes
Major allele
Minor allele
>10% reads (>20 counts) or >80% reads (>5 counts)
No
1 branch (same as old branch)
Compactors
C.
Single cells from different tissues without cell type metadata
Sequencing reads
SPLASH+
Compactors (anchor-seeded extended sequences)
Annotation databases (Rfam ...)
Post-facto annotation
Significant Anchors
SPLASH
Biological interpretation
Alternative splicing
SNP
Internal splicing
Antibody diversity
Repeat expansion

Figure 2.


A
Anchor
10 20 30 40 50 60 70 80 90 100
CCATTCCATT
Pericentromeric repeat
Compactor read count per cell type
10^2
10^1
10^0
Compactor
Total read count
0 100 200 300
Donor11 Lung
Donor12 Heart
Donor2 Muscle
Donor2 Spleen
Donor4 Mammary
Donor4 Muscle
Donor4 Tongue
Donor7 Blood
Donor7 Lymph node
Donor8 Prostate
Donor-tissues represented in each cell type
Sequence diversity detected by SPLASH+
A G T C
Basal cell (32)
Classical monocyte (69)
Skeletal muscle satellite stem cell (19)
Fibroblast of breast (17)
T cell (10)
Macrophage (18)
Neutrophil (21)
Memory B cell (15)
Pericyte cell (7)
Epithelial cell (14)
Endothelial cell of vascular tree (4)
Luminal epithelial cell of mammary gland (9)
Stromal cell (5)
Erythrocyte (9)
Mesynchemal stem cell (16)
B cell (17)
CD8+ T cell (14)
Capillary endothelial cell (8)
Innate lymphoid cell (7)
CD4+ T cell (6)


B
Location of RNU6 variants across human genome
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
ChrX
Chr8
Chr9
Chr11
Chr10
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr20
Chr19
ChrY
Chr22
Chr21
STAR Mapping
U6 RFAM
Abundance
U6 compactors
367
410
20
RNU6 references


C
Fraction of compactors
0
0.2
0.4
0.6
0.8
1
U6-by-association
U6
0
10
20
30
40
Minimum Hamming distance to an annotated RNU6


D
Compactor count per cell
RNU6-6P
30 2 6 16 12 12 12 6 6 8 8
Multimapped
14
Multimapped
12 4 2 3 3
Multimapped
8 8 6

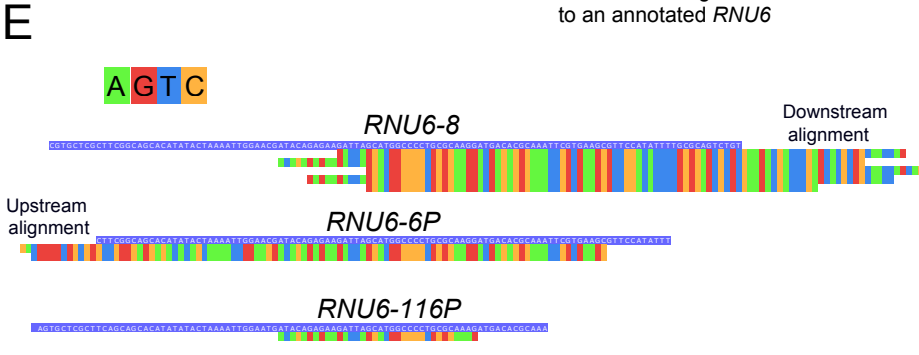

E
A G T C
RNU6-8
Downstream alignment
Upstream alignment
RNU6-6P
RNU6-116P

Figure 3.

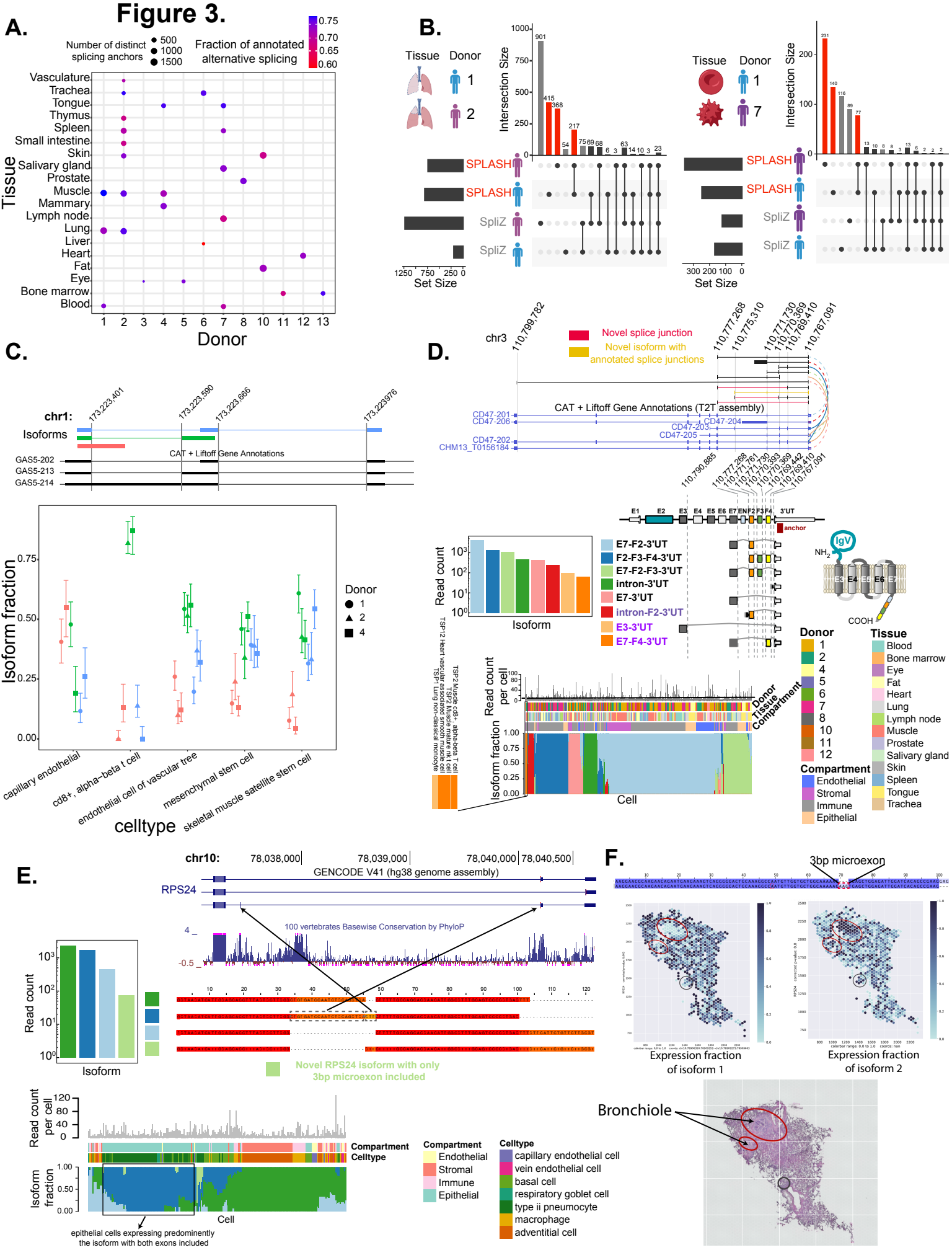

A.
Number of distinct splicing anchors
500
1000
1500
Fraction of annotated alternative splicing
0.75
0.70
0.65
0.60
Tissue
Vasculature
Trachea
Tongue
Thymus
Spleen
Small intestine
Skin
Salivary gland
Prostate
Muscle
Mammary
Lymph node
Lung
Liver
Heart
Fat
Eye
Bone marrow
Blood
Donor
B.
Tissue
Donor
Intersection Size
SPLASH
SpliZ
Set Size
C.
chr1:
173,223,401
173,223,590
173,223,666
173,223976
Isoforms
CAT + Liftoff Gene Annotations
GAS5-202
GAS5-213
GAS5-214
Isoform fraction
Donor
capillary endothelial
cd8+, alpha−beta t cell
endothelial cell of vascular tree
mesenchymal stem cell
skeletal muscle satellite stem cell
celltype
D.
chr3
Novel splice junction
Novel isoform with annotated splice junctions
CAT + Liftoff Gene Annotations (T2T assembly)
CD47-201
CD47-206
CD47-204
CD47-203
CD47-205
CD47-202
CHM13_T0156184
E1
E2
E3
E4
E5
E6
E7
F2
F3
F4
3'UT
anchor
Read count
Isoform
E7-F2-3'UT
F2-F3-F4-3'UT
E7-F2-F3-3'UT
intron-3'UT
E7-3'UT
intron-F2-3'UT
E3-3'UT
E7-F4-3'UT
IgV
NH2
COOH
Read count per cell
Isoform fraction
Donor
Tissue
Compartment
Cell
Donor
Tissue
Compartment
Endothelial
Stromal
Immune
Epithelial
Blood
Bone marrow
Eye
Fat
Heart
Lung
Lymph node
Muscle
Prostate
Salivary gland
Skin
Spleen
Tongue
Trachea
E.
chr10:
78,038,000
78,039,000
78,040,000
78,040,500
GENCODE V41 (hg38 genome assembly)
RPS24
100 vertebrates Basewise Conservation by PhyloP
Read count
Isoform
Novel RPS24 isoform with only 3bp microexon included
Read count per cell
Isoform fraction
Compartment
Celltype
Cell
epithelial cells expressing predominently the isoform with both exons included
Compartment
Endothelial
Stromal
Immune
Epithelial
Celltype
capillary endothelial cell
vein endothelial cell
basal cell
respiratory goblet cell
type ii pneumocyte
macrophage
adventitial cell
F.
3bp microexon
Expression fraction of isoform 1
Expression fraction of isoform 2
Bronchiole

Figure 4.

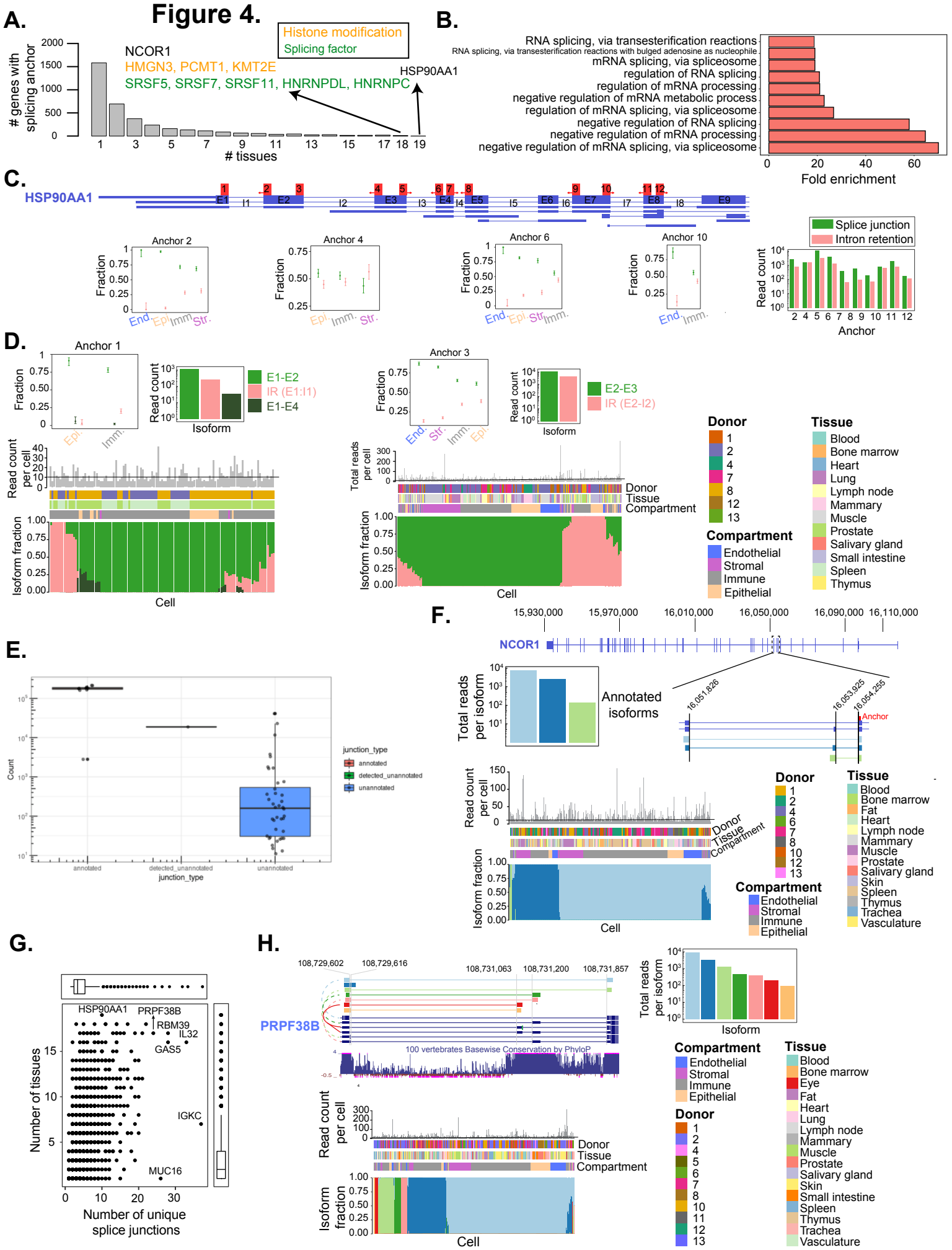

A.
Histone modification
Splicing factor
NCOR1
HMGN3, PCMT1, KMT2E
SRSF5, SRSF7, SRSF11, HNRNPDL, HNRNPC
HSP90AA1
genes with splicing anchor
tissues
B.
RNA splicing, via transesterification reactions
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
mRNA splicing, via spliceosome
regulation of RNA splicing
regulation of mRNA processing
negative regulation of mRNA metabolic process
regulation of mRNA splicing, via spliceosome
negative regulation of RNA splicing
negative regulation of mRNA processing
negative regulation of mRNA splicing, via spliceosome
Fold enrichment
C.
HSP90AA1
Anchor 2
Anchor 4
Anchor 6
Anchor 10
Fraction
Splice junction
Intron retention
Read count
Anchor
D.
Anchor 1
Anchor 3
Read count
Isoform
E1-E2
IR (E1:I1)
E1-E4
E2-E3
IR (E2-I2)
Read count per cell
Isoform fraction
Total reads per cell
Donor
Tissue
Compartment
Cell
Donor
Tissue
Compartment
Endothelial
Stromal
Immune
Epithelial
E.
Count
junction_type
annotated
detected_unannotated
unannotated
F.
NCOR1
Total reads per isoform
Annotated isoforms
Anchor
Read count per cell
Isoform fraction
Donor
Tissue
Compartment
Cell
Donor
Tissue
Compartment
Endothelial
Stromal
Immune
Epithelial
G.
HSP90AA1
PRPF38B
RBM39
IL32
GAS5
IGKC
MUC16
Number of tissues
Number of unique splice junctions
H.
PRPF38B
100 vertebrates Basewise Conservation by PhyloP
Total reads per isoform
Isoform
Read count per cell
Isoform fraction
Donor
Tissue
Compartment
Cell
Compartment
Endothelial
Stromal
Immune
Epithelial
Donor
Tissue
Blood
Bone marrow
Eye
Fat
Heart
Lung
Lymph node
Mammary
Muscle
Prostate
Salivary gland
Skin
Small intestine
Spleen
Thymus
Trachea
Vasculature

Figure 5

(A)

Compactor category

- IGKC (# 4,272)
- IGH (# 2,294)
- IG (# 2,164)
- CCATTCCATT or AATGGAATGG (# 19,434)
- Mitochondrial gene (# 1,861)
- Other compactors (# 291,515)

donors where this anchor appears

Number of compactors per anchor

Donor 7 Spleen
Donor 2 Spleen
Donor 7 Lymph Node
Donor 1 Blood
Donor 13 Bone Marrow
Donor 2 Thymus
Donor 7 Salivary Gland

(B)

BraCeR		SPLASH+
missed	/	called
called	/	missed
called	/	called

Donor 2 Spleen

Memory B cell: 41 (93%), 76
Plasma cell: 12 (100%), 7

Read count per cell

Donor 7 Spleen

Memory B cell: 2 (100%), 1, 77
Naive B cell: 1, 14
Plasma cell: 2 (100%), 47

Read count per cell

(C)

Sample
TSP2 Spleen
TSP7 Lymphnode
TSP7 Spleen

Percent cells BraCeR match

Minimum Hamming Distance

(D)

Compactor 1
Compactor 2
Compactor 3
Compactor 4
Compactor 5
Compactor 6
Compactor 7
Compactor 8
Anchor

Chr14 100,791,000 100,791,100 100,791,150

Compactor 1
Compactor 3
Compactor 4
Compactor 5
Compactor 6
Compactor 7
Compactor 8
IGHV3-43-201

Chr14 100,425,450 100,425,550 100,425,650

Compactor 1
Compactor 2
Compactor 8
IGHV3-15-201

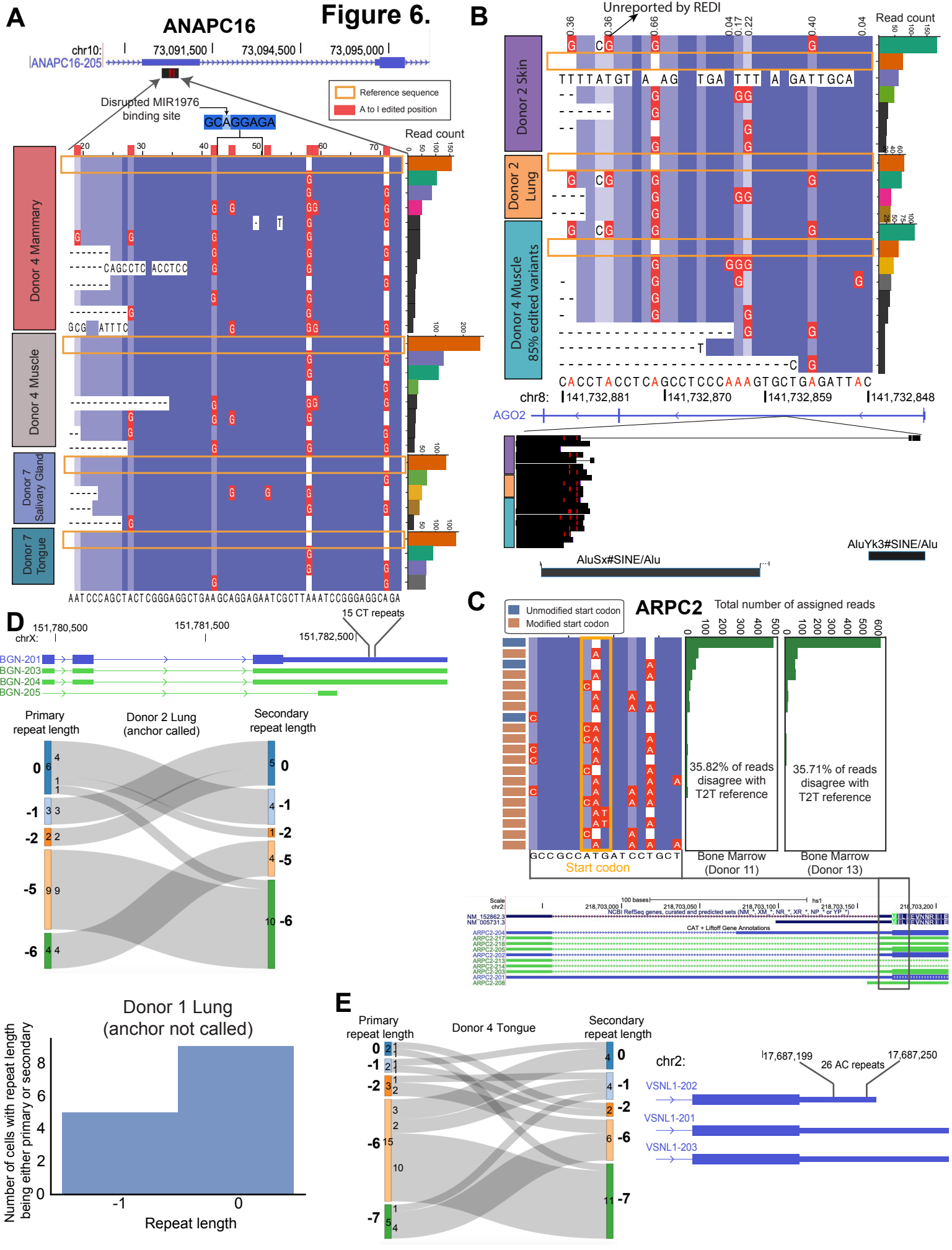
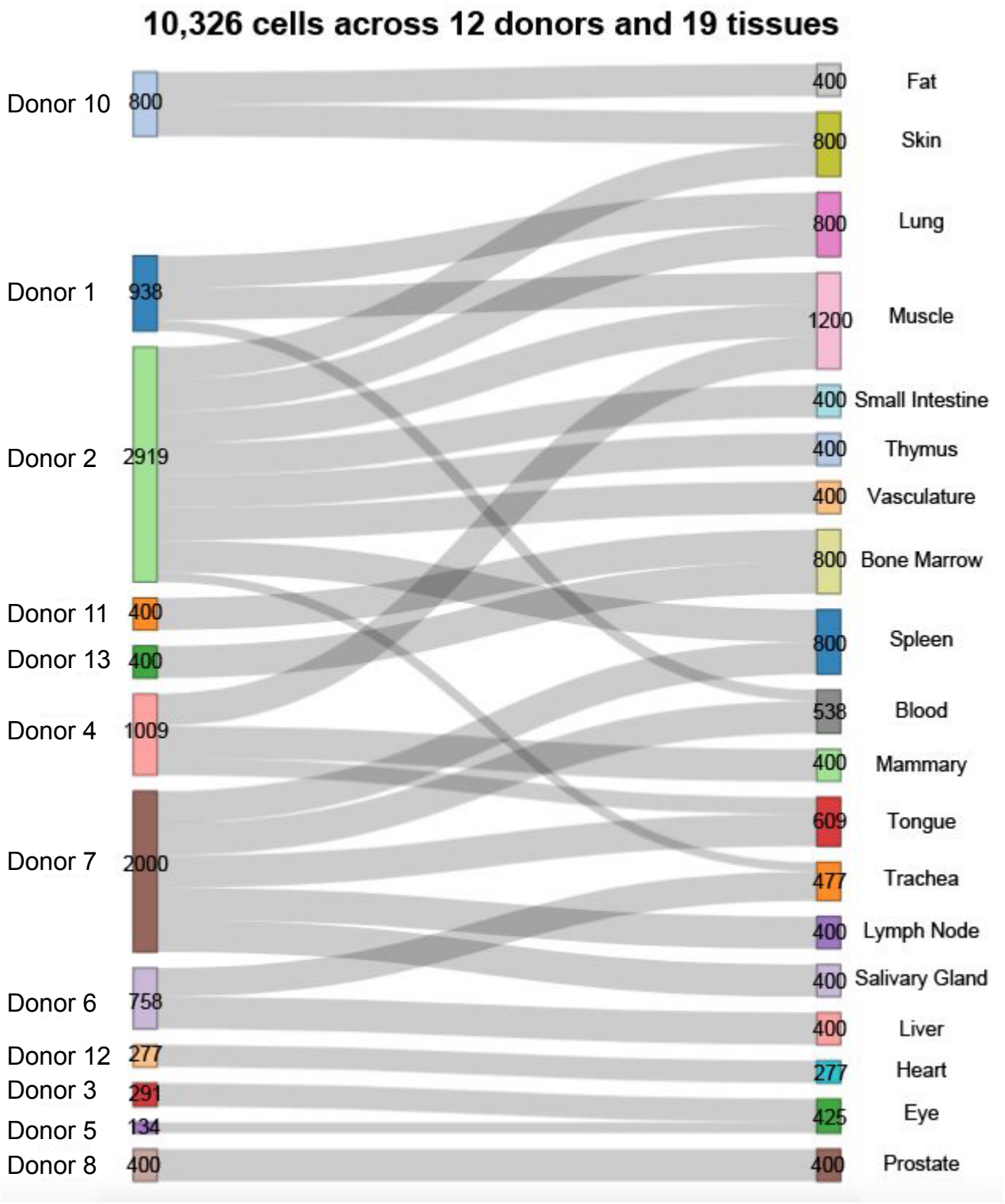


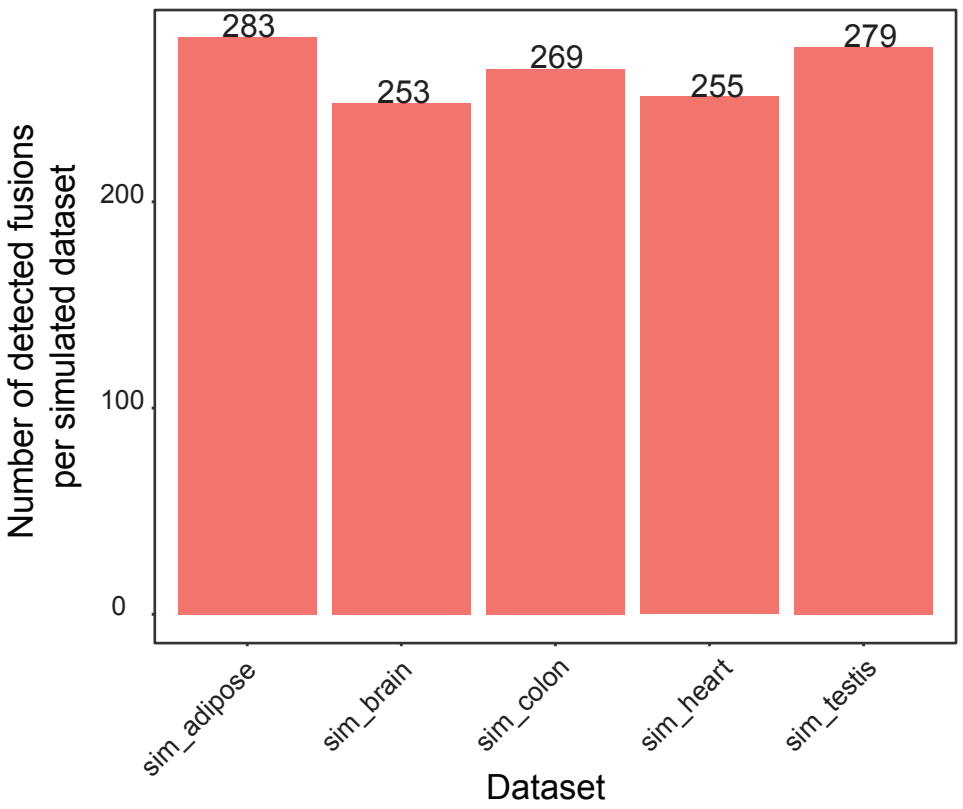
Figure 6.
A
ANAPC16
chr10:
73,091,500
73,094,500
73,095,000
ANAPC16-205
Disrupted MIR1976 binding site
GCAGGAGA
Reference sequence
A to I edited position
Read count
Donor 4 Mammary
Donor 4 Muscle
Donor 7 Salivary Gland
Donor 7 Tongue
AATCCCAGCTACTCGGGAGGCTGAAGCAGGAGAATCGCTTAAATCCGGGAGGCAGA
B
Unreported by REDI
Read count
Donor 2 Skin
Donor 2 Lung
Donor 4 Muscle 85% edited variants
CACCTACCTCAGCCTCCCAAAGTGCTGAGATTAC
chr8:
141,732,881
141,732,870
141,732,859
141,732,848
AGO2
AluSx#SINE/Alu
AluYk3#SINE/Alu
C
ARPC2
Total number of assigned reads
Unmodified start codon
Modified start codon
35.82% of reads disagree with T2T reference
35.71% of reads disagree with T2T reference
Start codon
Bone Marrow (Donor 11)
Bone Marrow (Donor 13)
D
chrX:
151,780,500
151,781,500
151,782,500
15 CT repeats
BGN-201
BGN-203
BGN-204
BGN-205
Primary repeat length
Donor 2 Lung (anchor called)
Secondary repeat length
Donor 1 Lung (anchor not called)
Number of cells with repeat length being either primary or secondary
Repeat length
E
Primary repeat length
Donor 4 Tongue
Secondary repeat length
chr2:
17,687,199
17,687,250
26 AC repeats
VSNL1-202
VSNL1-201
VSNL1-203

Suppl. Figure 1

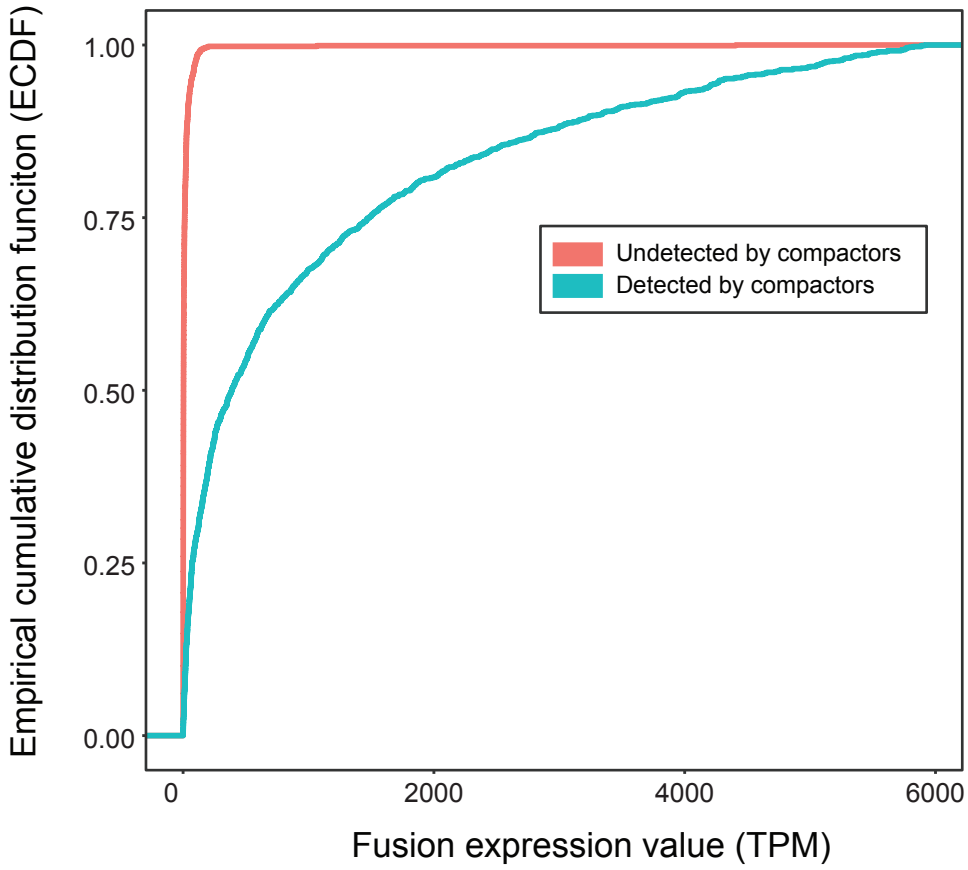


Suppl. Figure 2

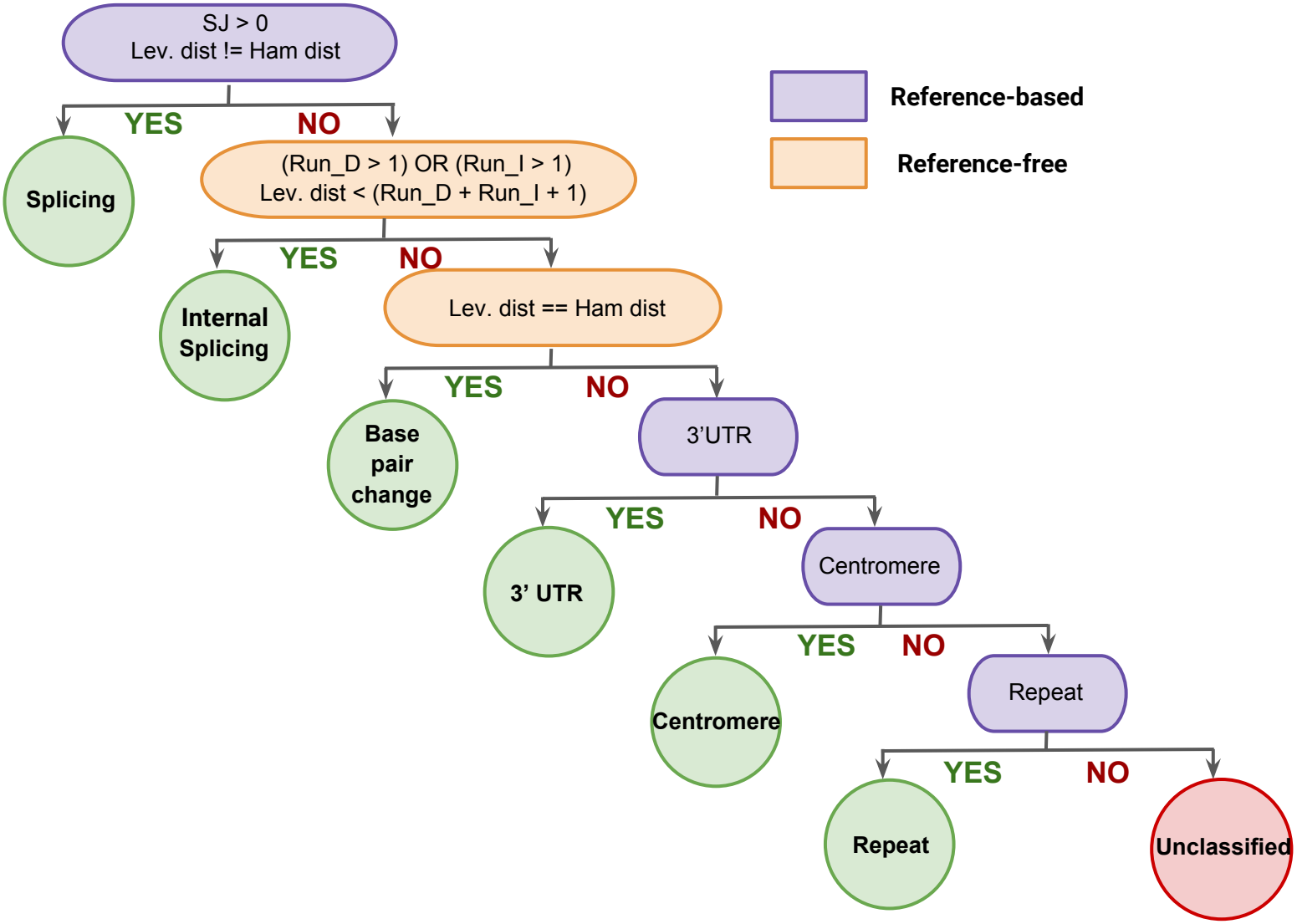
A



B

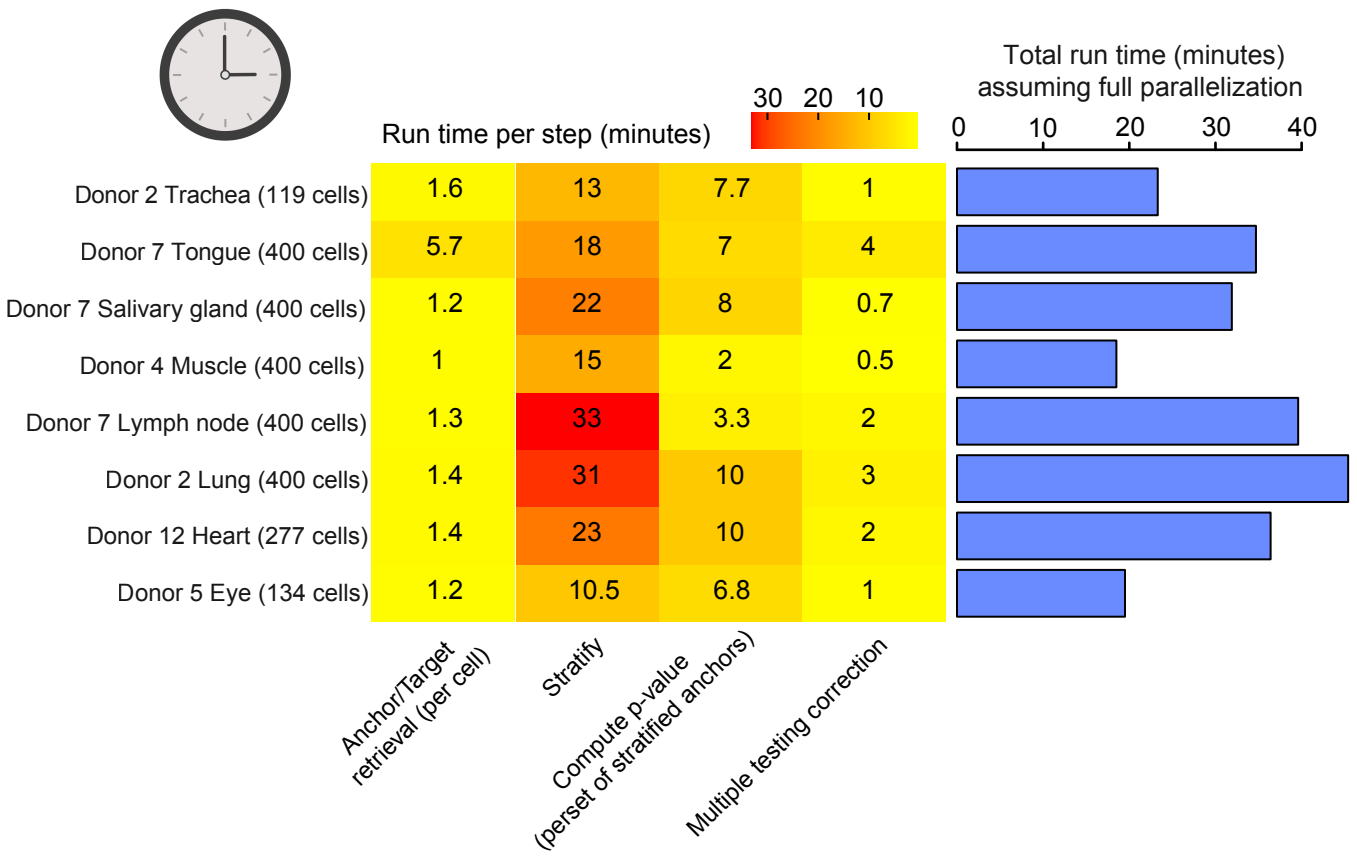


Suppl. Figure 3

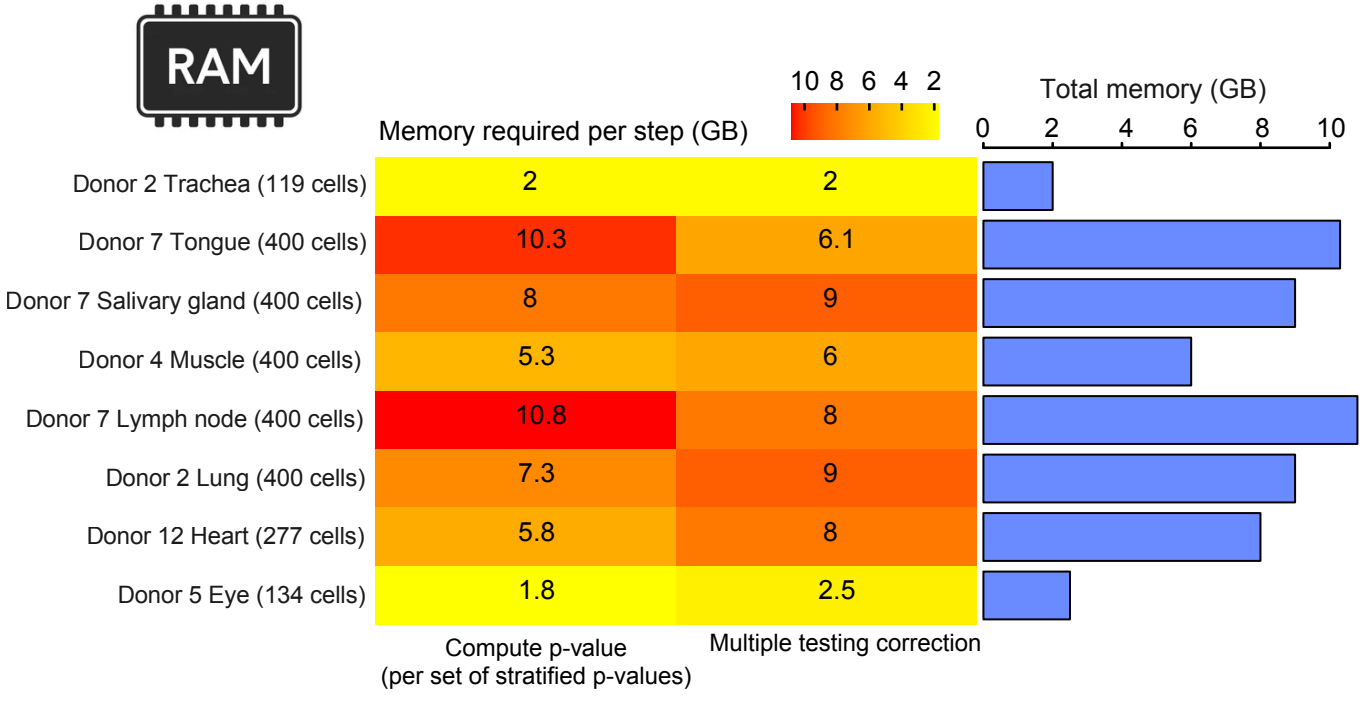


Suppl. Figure 4

A

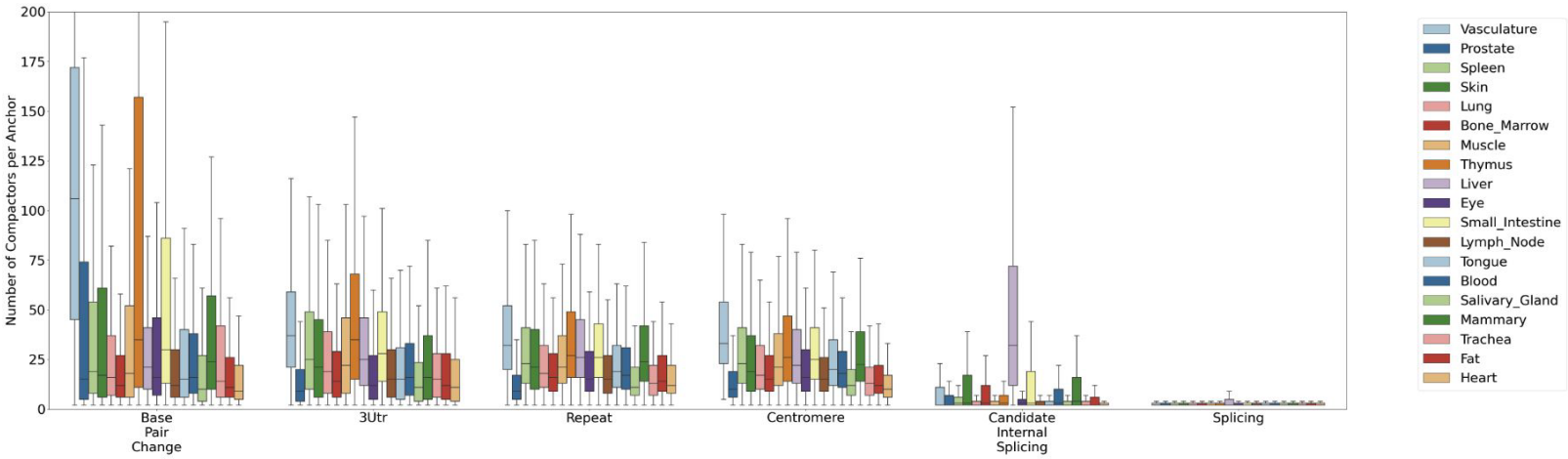


B

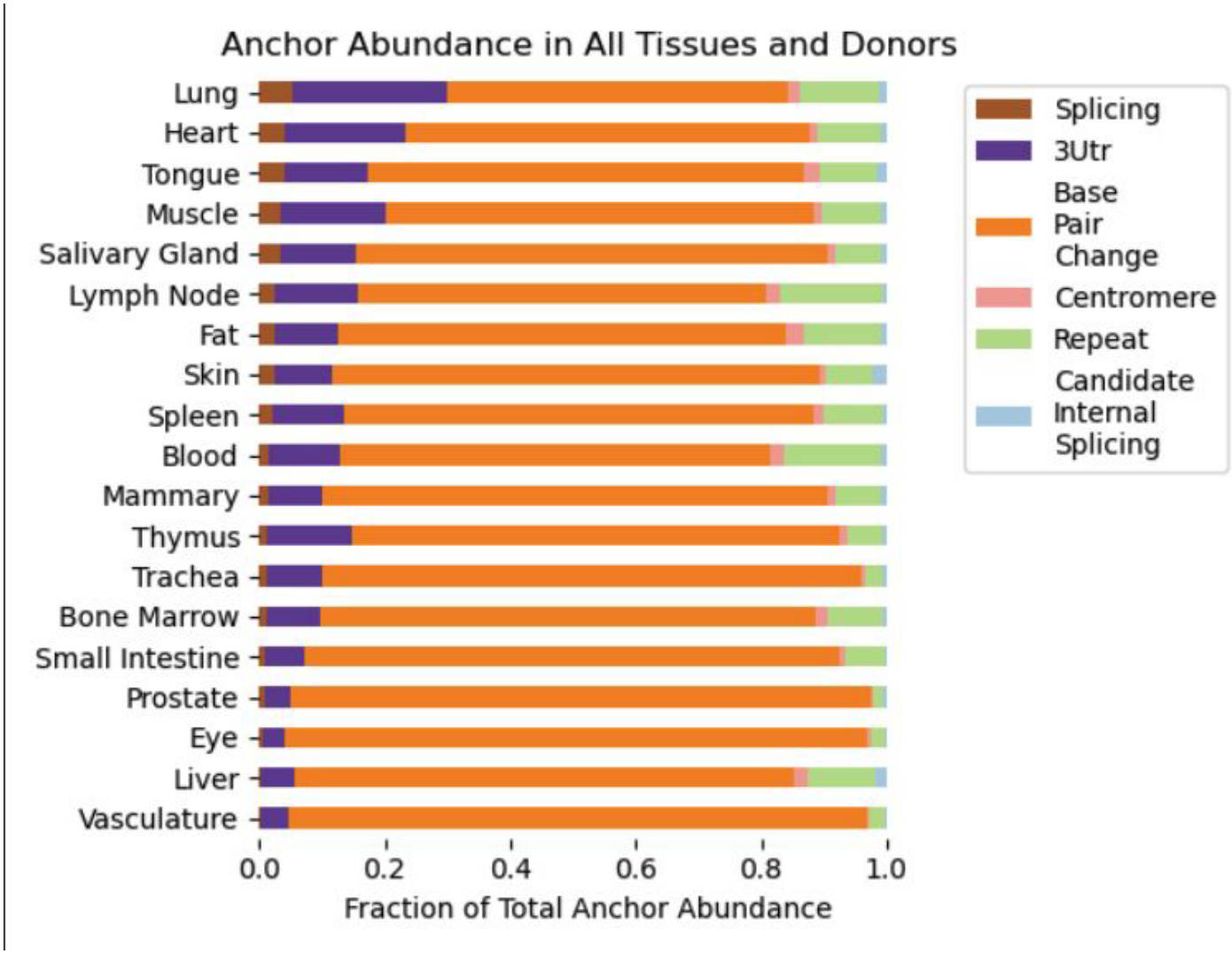


Suppl. Figure 5

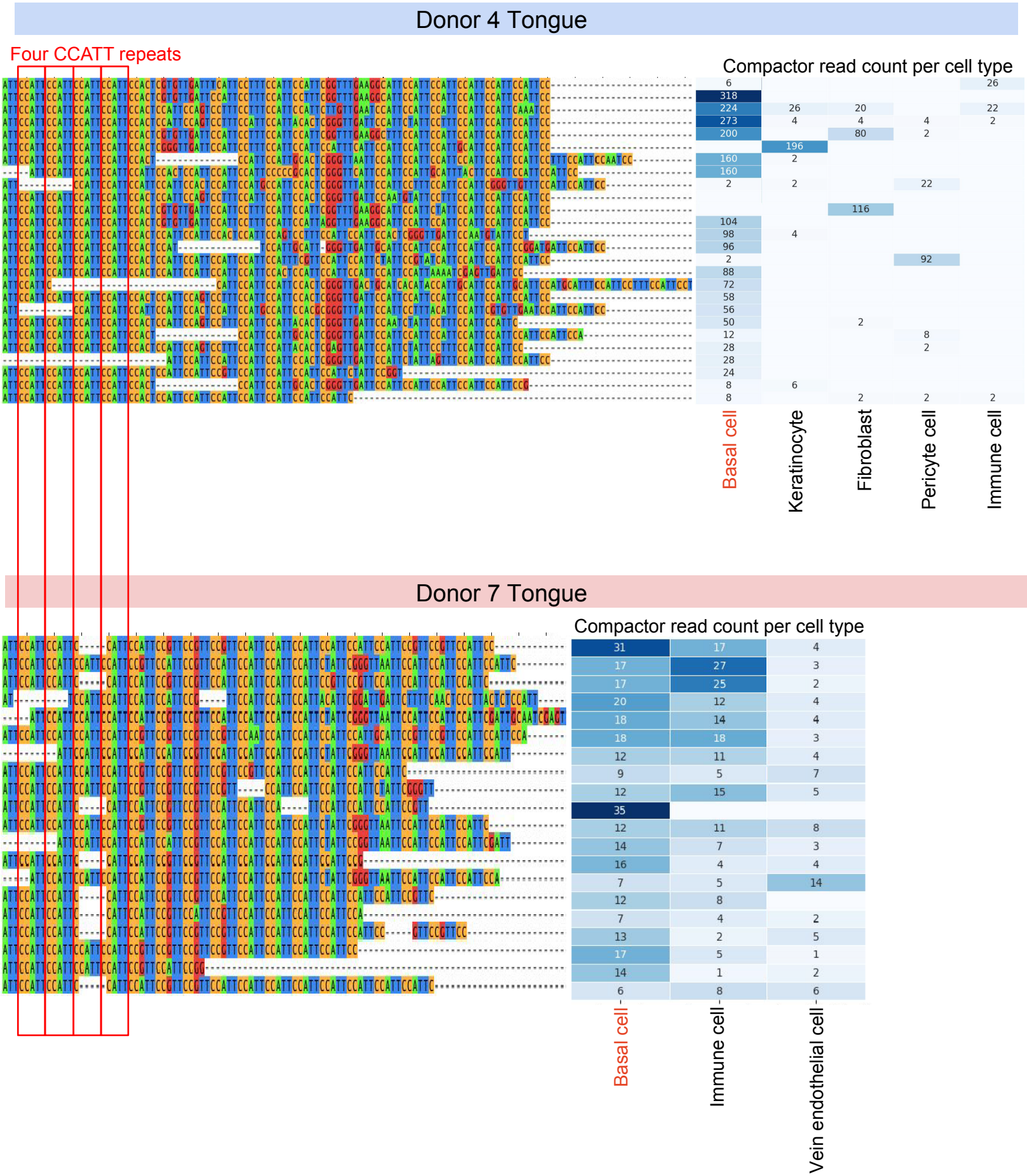
A



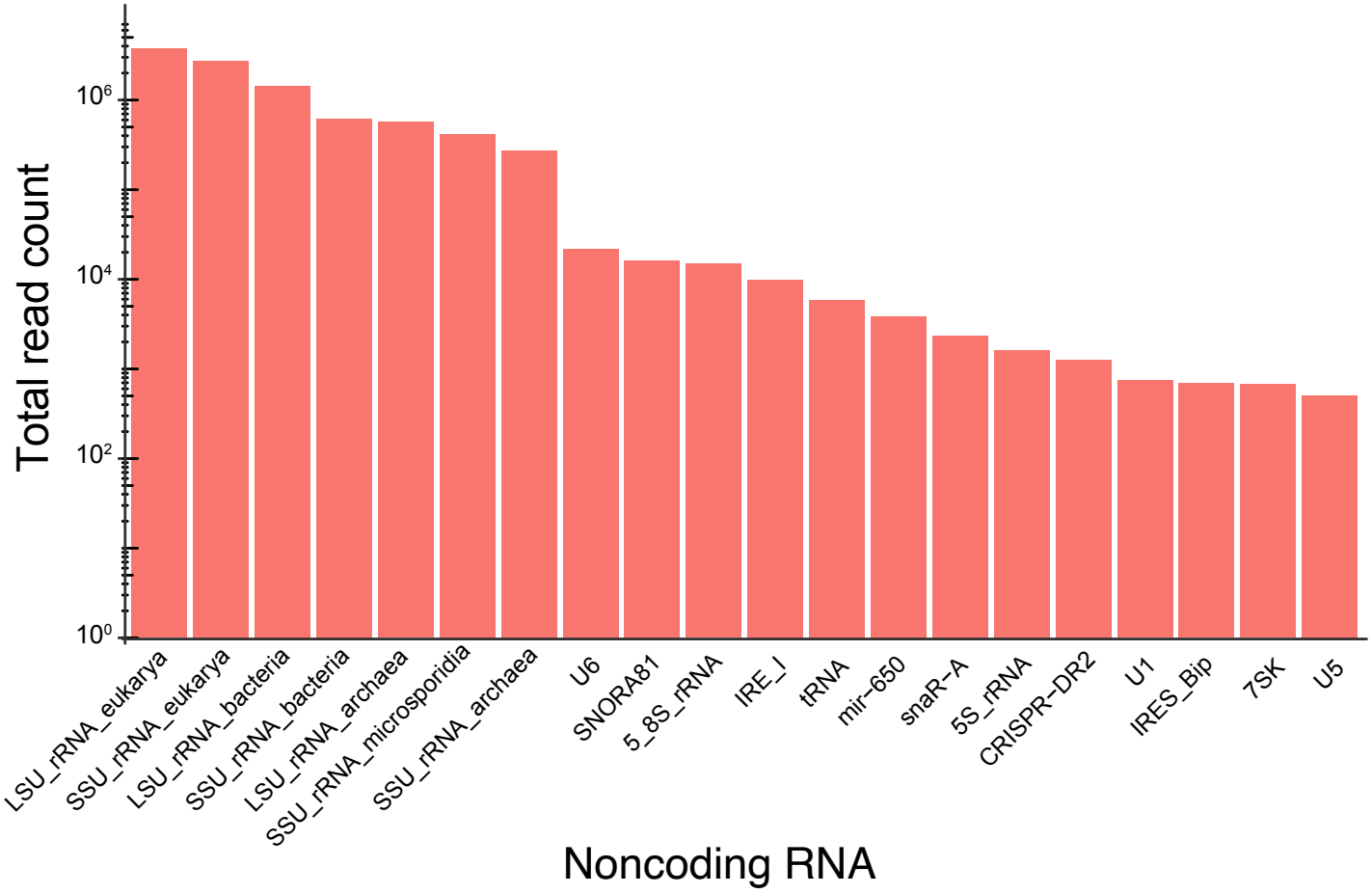
B



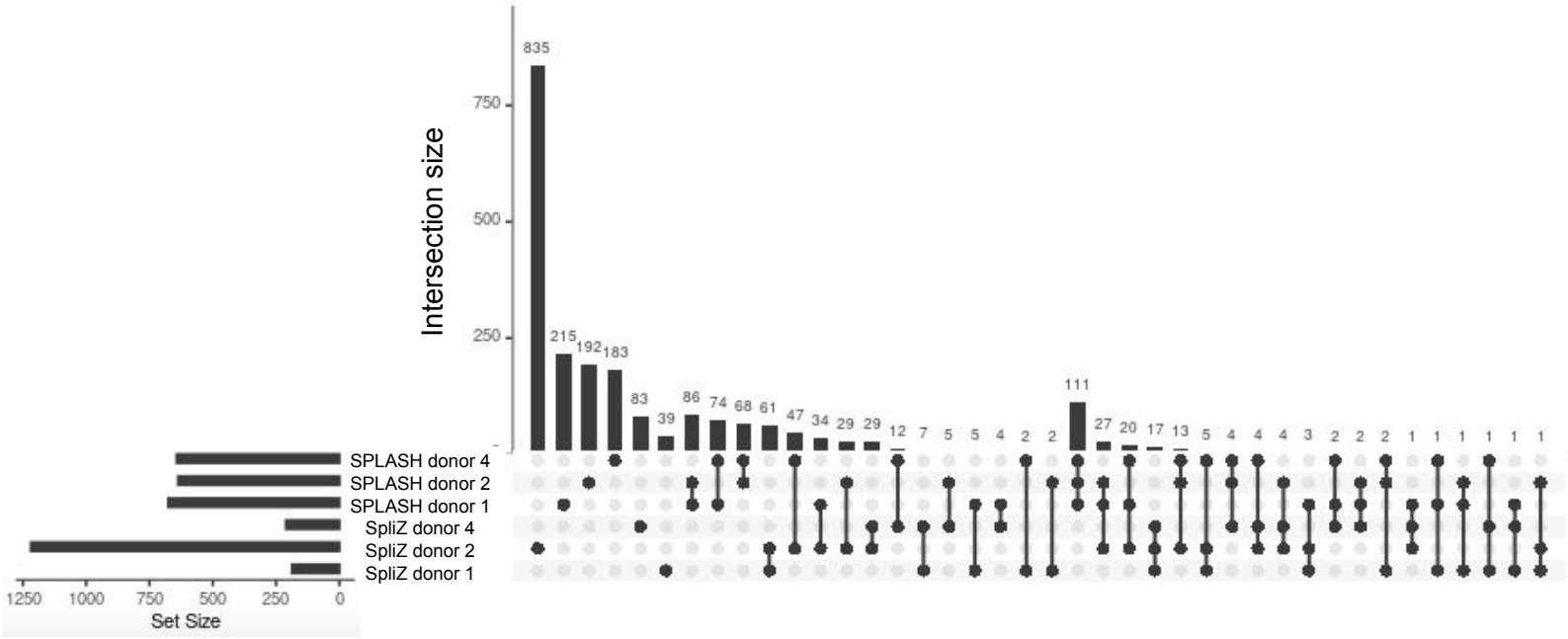
Suppl. Figure 6



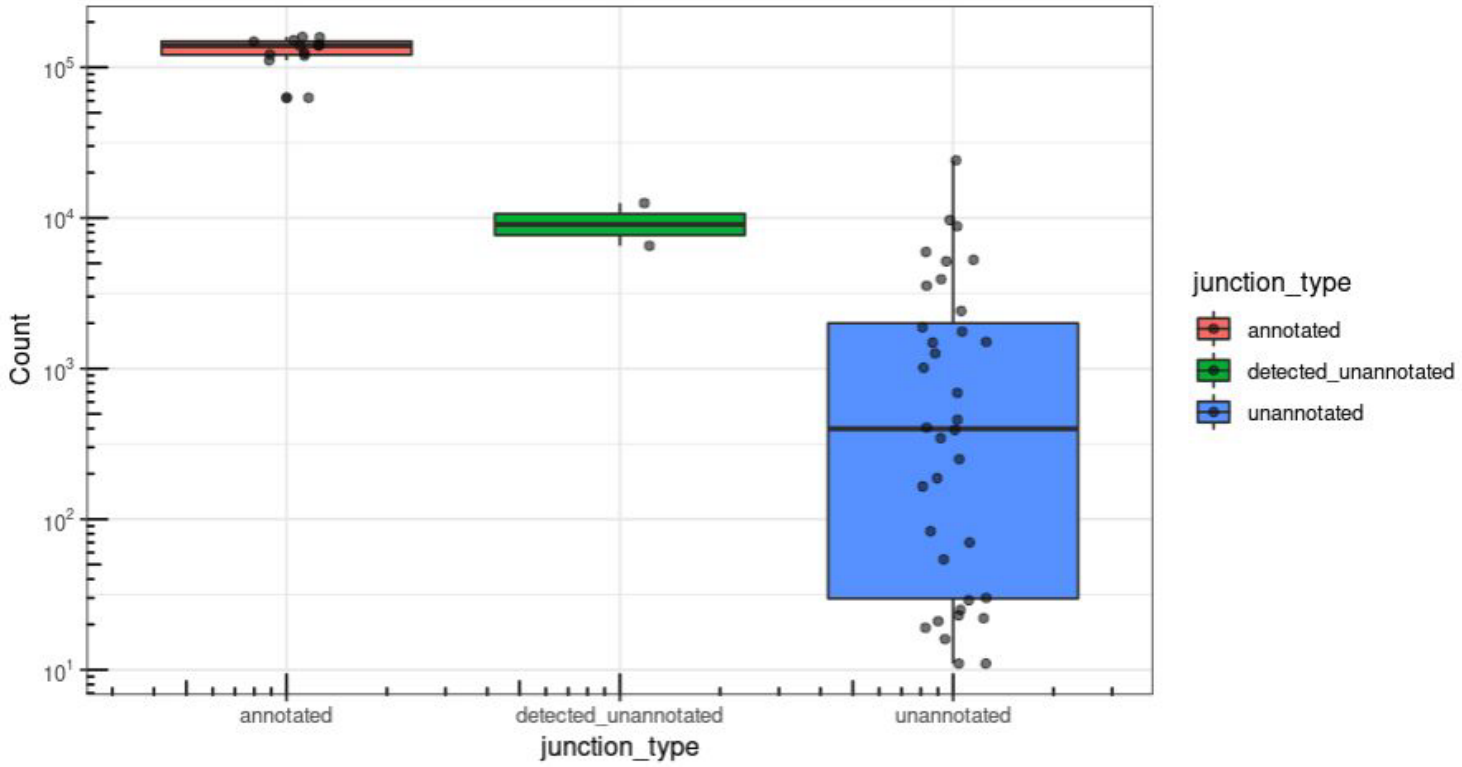
Suppl. Figure 7



Suppl. Figure 8

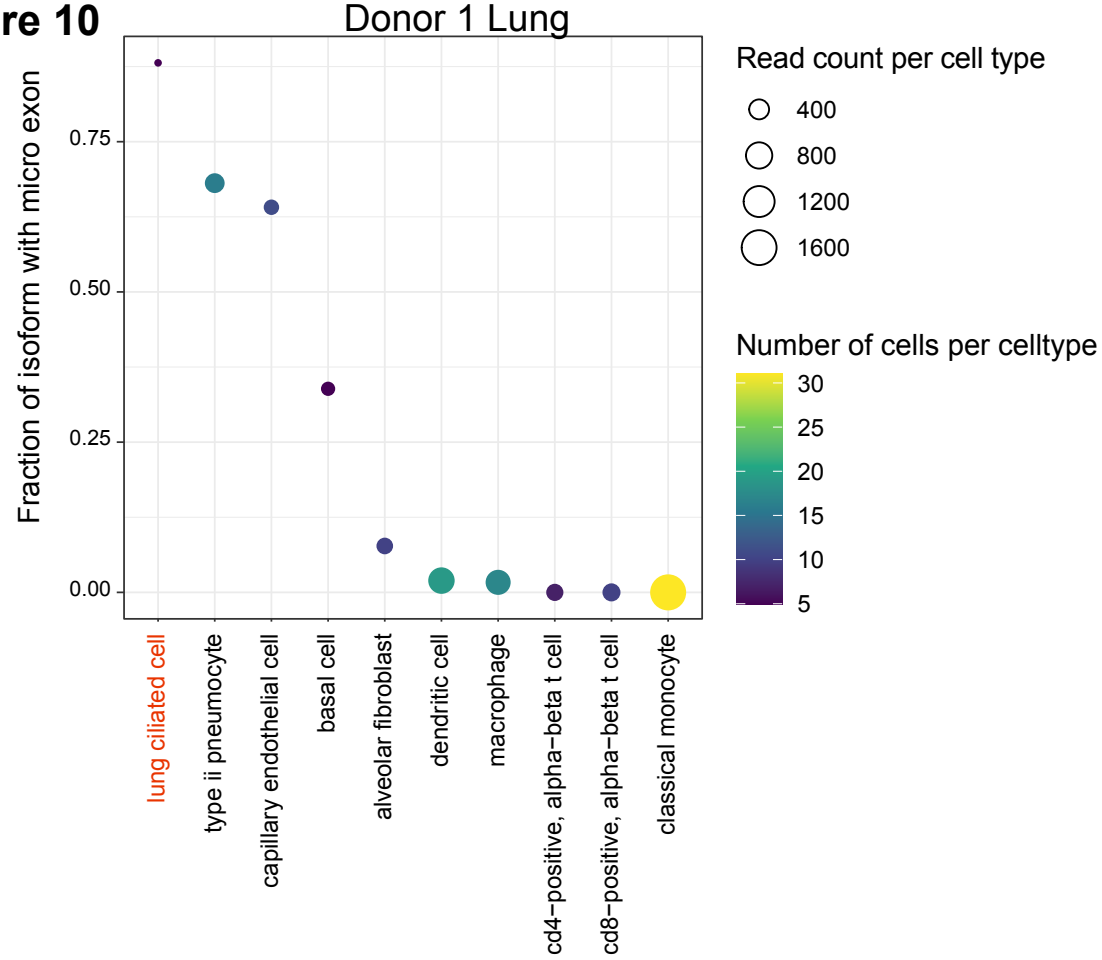


Suppl. Figure 9

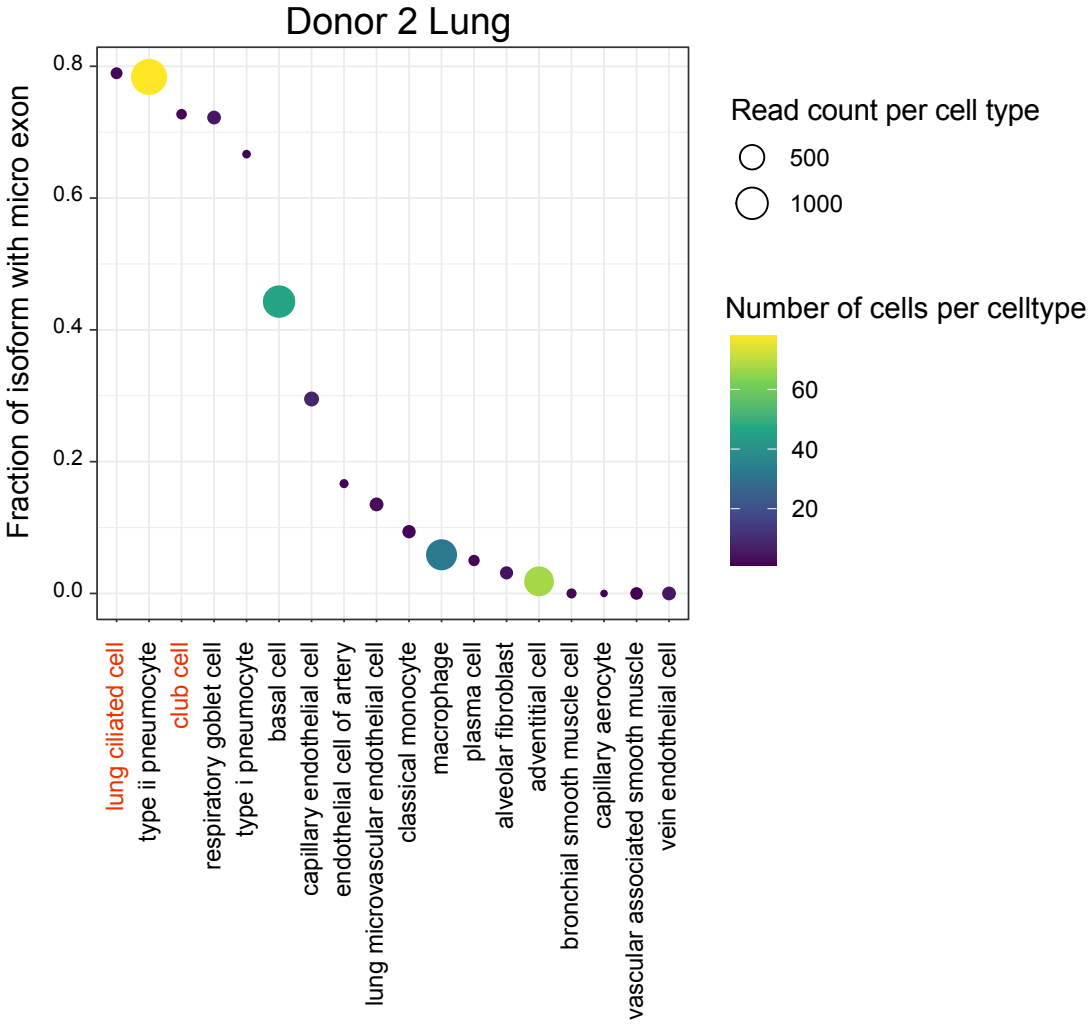


Suppl. Figure 10

A.

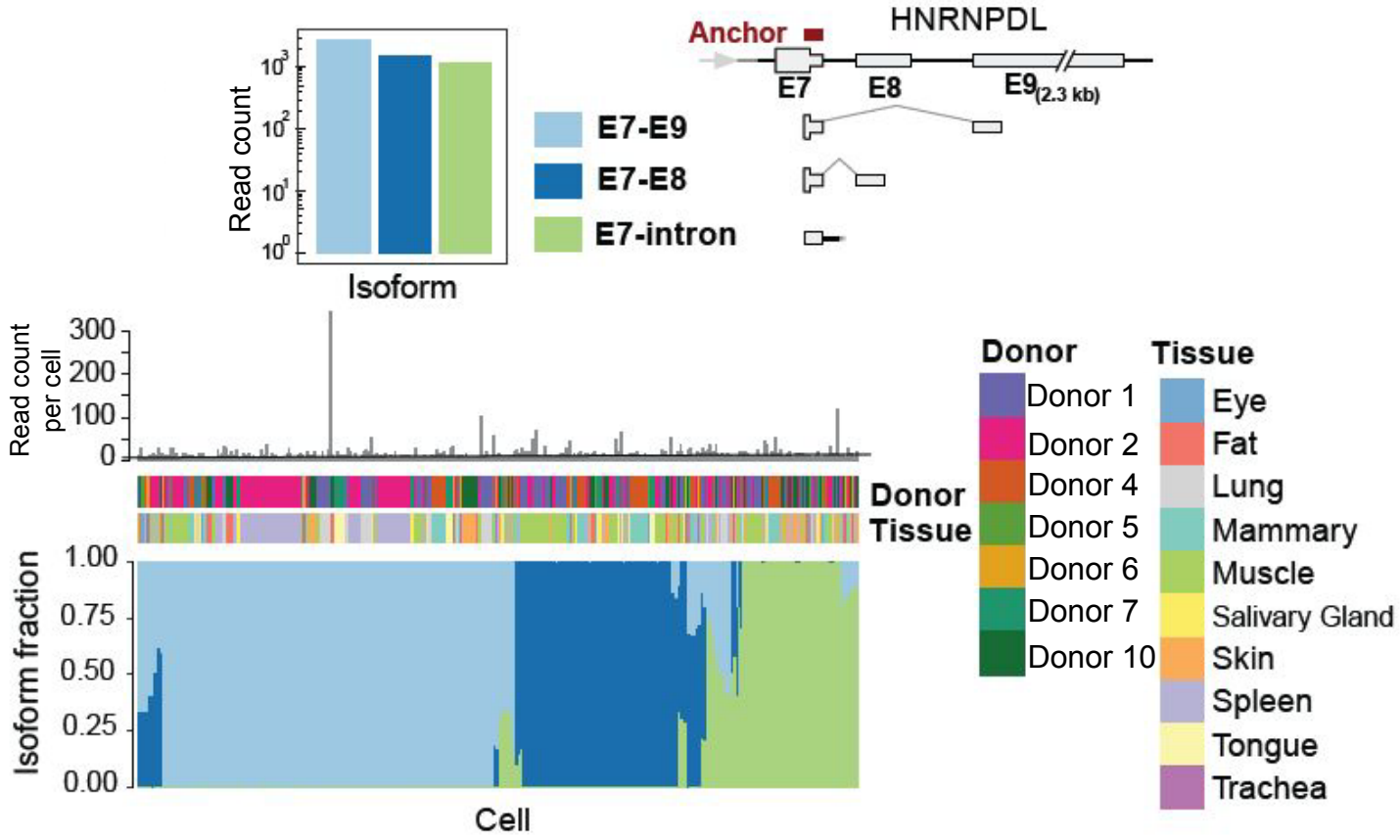


B.

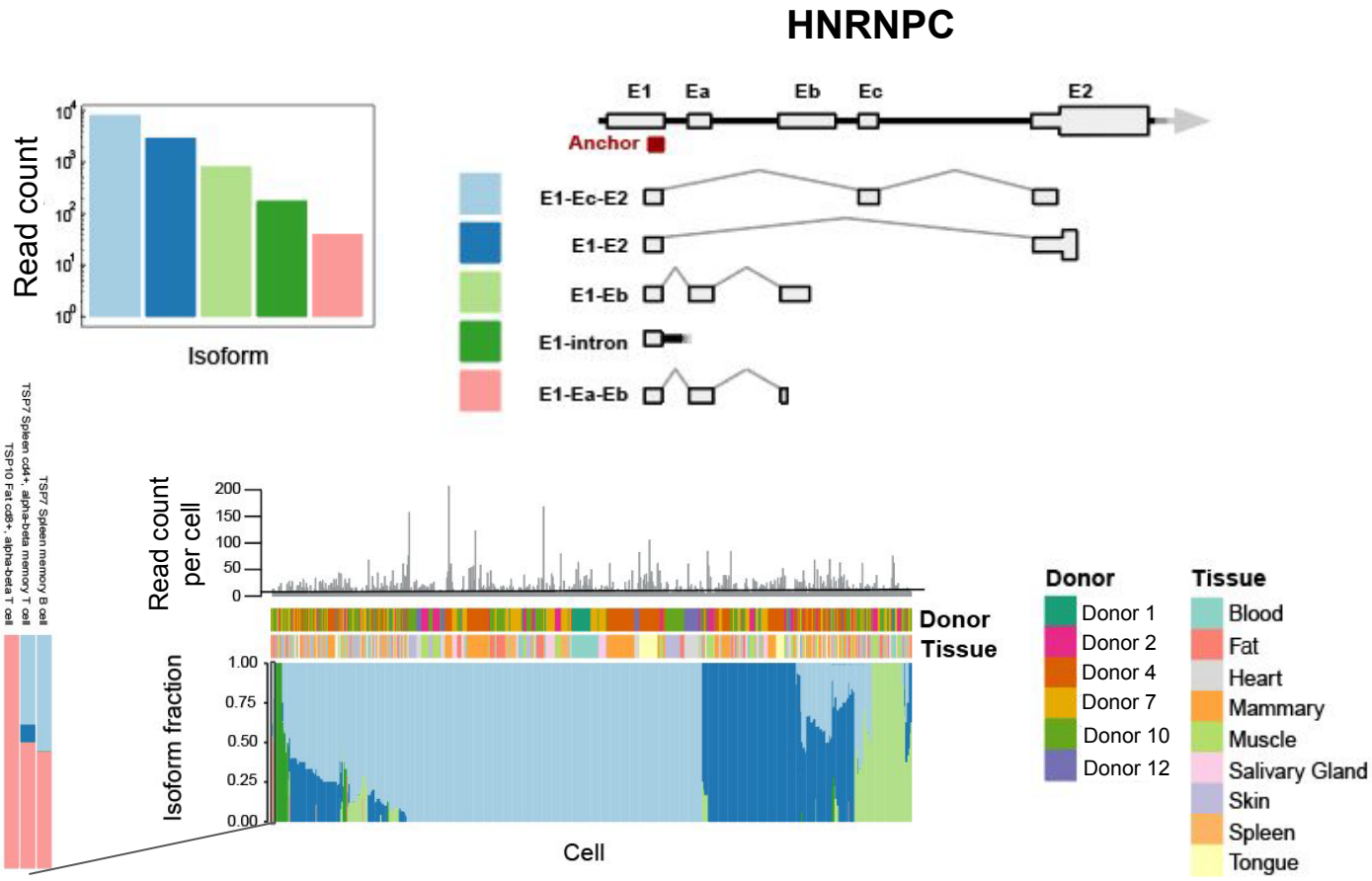


Suppl. Figure 11

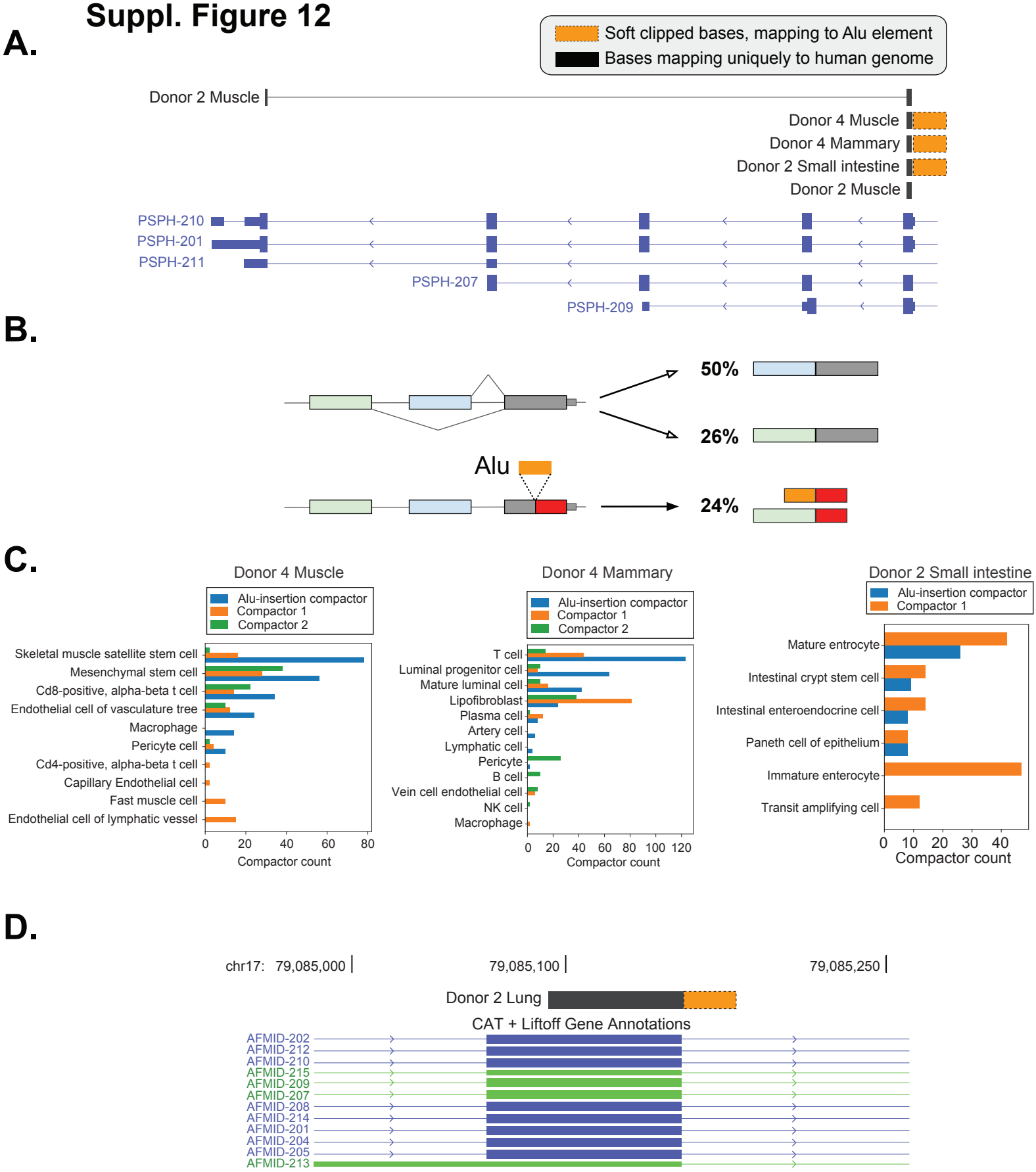
A



B

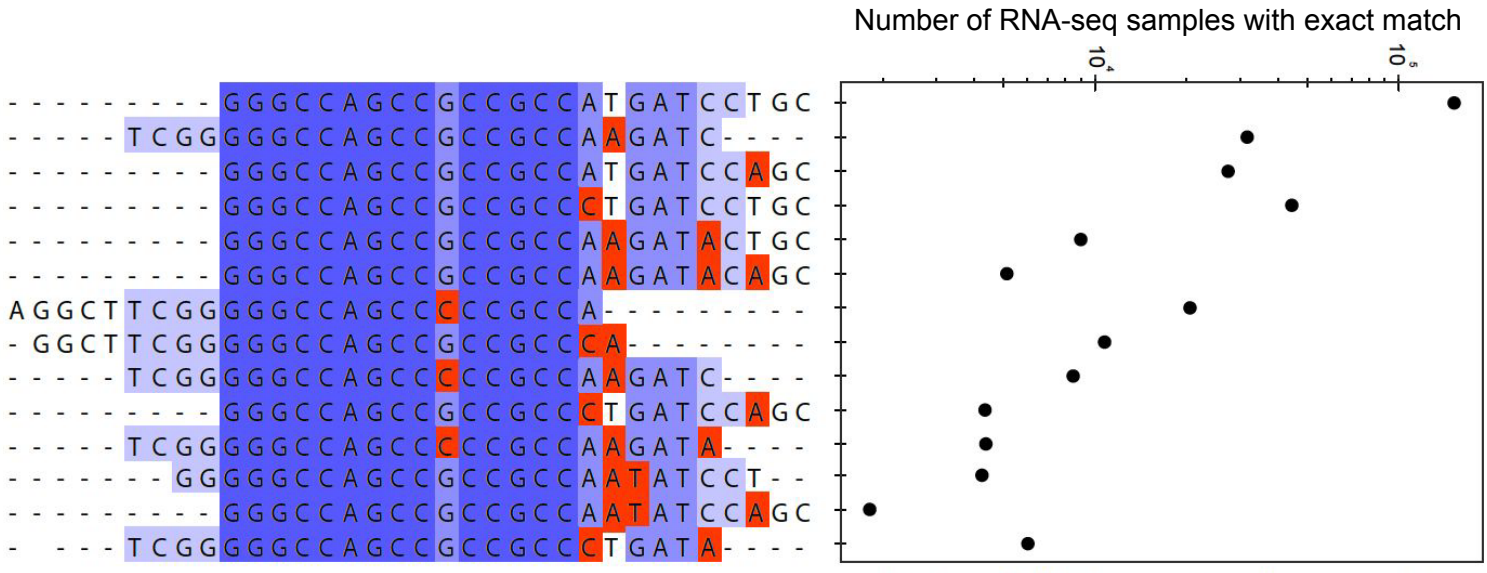
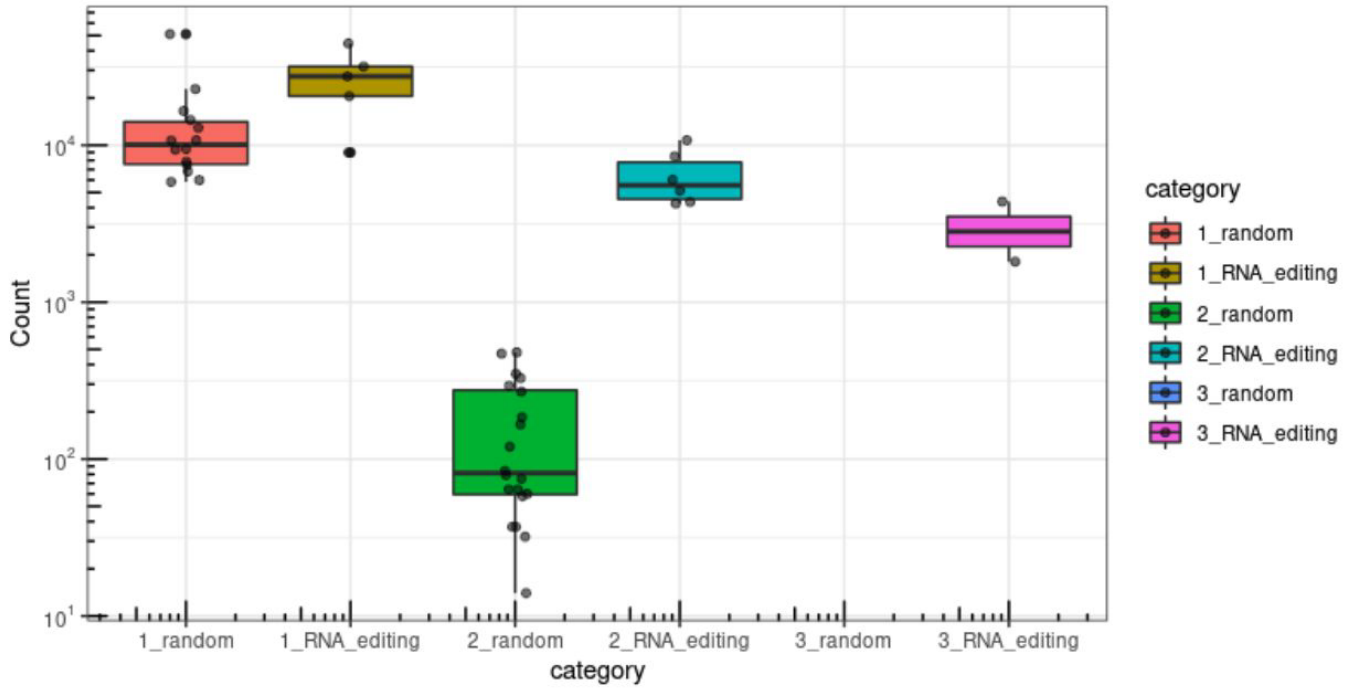


Suppl. Figure 12



Suppl. Figure 13

A



B