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 (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 the DNA and RNA level with profound functional consequences. Examples include diversity generated by RNA splicing and V(D)J recombination. Today, these and other events are detected with fragmented bioinformatic tools that require predefining a form of transcript diversification; moreover, they rely on alignment to a necessarily incomplete reference genome, filtering out unaligned sequences which can be among the most interesting. Each of these steps introduces blindspots for discovery. Here, we develop NOMAD+, a new analytic method that performs unified, reference-free statistical inference directly on raw sequencing reads, extending the core NOMAD algorithm to include a micro-assembly and interpretation framework. NOMAD+ discovers broad and new examples of transcript diversification in single cells, bypassing genome alignment and without requiring cell type metadata and impossible with current algorithms. In 10,326 primary human single cells in 19 tissues profiled with SmartSeq2, NOMAD+ discovers a set of splicing and histone regulators with highly conserved intronic regions that are themselves targets of complex splicing regulation and unreported transcript diversity in the heat shock protein *HSP90AA1*. NOMAD+ simultaneously discovers diversification in centromeric RNA expression, V(D)J recombination, RNA editing, and repeat expansions missed by or impossible to measure with existing bioinformatic methods. NOMAD+ is a unified, highly efficient algorithm enabling unbiased discovery of an unprecedented breadth of RNA regulation and diversification in single cells through a new paradigm to analyze the transcriptome.

Introduction

In eukaryotes, each gene and non-coding RNA locus can produce diverse isoforms with sometimes opposite functions, including by well-studied mechanisms such as alternative splicing, RNA editing, and alternative 5' and 3' UTR use. Genetic changes in single cells - including insertion of mobile elements, repeat expansions, or segmental duplications - can further expand this diversity. In the adaptive immune system, V(D)J recombination determines the specificity and success of defense against pathogens; the genome has the potential to create more than 10^{13} genetic variants (Schroeder 2006). Together, transcript diversification can have significant functional consequences, including causal links to disease 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 that transcript diversity is regulated in single cells remains a significant open question in genome science. Bioinformatics today is fragmented: specialized approaches are required to identify editing, splicing, or V(D)J recombination, separately. Current computational

approaches to detect transcript diversity in single cells are also heavily reliant on references, beginning with alignment of reads to a reference genome before attempting to quantify diversity, 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). Some domain-specific approaches have been recently developed for RNA regulation analysis in single-cell RNA-seq (scRNA-seq), such as for splicing (Olivieri, Dehghannasiri, and Salzman 2022; Buen Abad Najar et al. 2022) or V(D)J recombination reconstruction (Lindeman et al. 2018), but other tasks, such as identifying somatically acquired repeats or RNA editing are not even attempted. Further, it remains an open question whether specialized algorithms are sensitive to all of the events they desire to detect. For example, reference-first algorithms 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, genomic inference today is conditional on the outputs of partially heuristic alignment algorithms, which often discard reads that do not map to the reference genome. Reference-first approaches also emphasize genomic health disparities because reference genomes are highly biased towards representing sequences of European ancestries (Sherman et al. 2019). Further, the statistical tests downstream of alignment or pseudo-alignment are typically themselves parametric or require randomized resampling, which can result in inaccurate or inefficient p-values (Figure 1A). Together, there is a strong argument to bypass reference alignment to find regulated sequence diversity through a fundamental unified framework.

We recently introduced NOMAD (Chaung et al. 2022), which shows that myriad biological processes that diversify transcripts can be detected with a unified reference-free algorithm, performing inference directly on raw, unaligned sequencing reads. This includes but is not limited to RNA splicing, mutations, RNA editing, and V(D)J recombination. Additionally, NOMAD can detect 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.

NOMAD's core includes a novel statistical test to detect sample-specific sequence variation that fills a gap in existing methods. Classical and parametric tests struggle to prioritize biologically important variation because they are overpowered in the context of noise generated from biochemical sampling, and such approaches may report inaccurate p-values. NOMAD's test provides finite-sample valid p-value bounds and, unlike Pearson's chi-squared test, controls false positive calls under commonly used modeling regimes such as negative binomial for scRNA-seq (Supplement, (Buen Abad Najar, Yosef, and Lareau 2020; Tavor Z. Baharav, David Tse, Julia Salzman, n.d.). NOMAD's test performs inference in scRNA-seq independent of any cell metadata (e.g., cell type), which can be difficult to generate and remains imprecise (Zeng 2022) and can miss important variation within cell types, such as B cell receptor variation (Watson and Breden 2012).

In this manuscript, we build on this core to introduce NOMAD+, which includes new approaches to analyze NOMAD's output, including a new, simple reference-free statistical approach to de novo assembly as well as a framework to interpret its results (Figures 1B,C). We use NOMAD+ to discover extensive RNA transcript diversification in 10,326 human single cells profiled using SmartSeq2 from 136 cell types and 12 donors from the Tabula Sapiens project (Tabula Sapiens Consortium* et al. 2022). NOMAD+ reveals new insights into the biology of single-cell regulation of transcript diversification – including features of RNA splicing, editing, and non-coding RNA expression missed by specialized, domain-specific bioinformatic pipelines. NOMAD+ detects sequences that have no known mapping to T2T, the latest human genome assembly (Altemose et al. 2022). Novel findings include (i) regulated expression of repetitive loci: RNU6 variants and of higher order repeats in centromeres including significant variation missed by mapping to the T2T reference genome (Hoyt et al. 2022); (ii) complex splicing programs including un-annotated variants in genes such as *CD47*, a major cancer immunotherapy target; (iii) pan-tissue regulation of splicing in splicing factors, histone regulation, and in the heat shock protein *HSP90AA1*; (iv) de novo rediscovery of immunoglobulin loci as the

most transcriptionally diverse human loci with improved sensitivity; and (v) single cells with transcribed repeat expansion and high levels of RNA editing. NOMAD+ makes discoveries that are impossible with existing algorithms, without cell metadata or reference genomes, avoiding biases towards alignments to genomes best curated for European ancestries. The results herein suggest NOMAD+ is a single algorithm that could replace the myriad custom bioinformatic approaches to detect 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

NOMAD is a highly efficient algorithm that operates directly on raw sequencing data to identify differentially diversified sequences (Figure 1A), characteristic of some of the most important transcript regulation such as splicing, RNA editing, V(D)J recombination. NOMAD achieves this by parsing reads into k-mer sequences, called anchors, that are followed by diverse sequences, called targets, a fixed distance downstream of them (Figure 1A, Supplement). NOMAD calls anchors that have sample-specific target expression, reflecting inter-cell variation in expression that can be signature of alternative RNA splicing, RNA editing, and among many other examples (Supplement). Anchors are too short (selected as 27-mers by default) for a biological interpretation of underlying mechanism generating sequence diversity. To fill this gap, we have expanded NOMAD to include a new algorithm that functions directly on raw sequences to provide sensitive, seed-based branched local de novo assembly (Figure 1B, Methods), which we call NOMAD+.

Each significant anchor sequence is used as a seed in the assembly step. First, fastq files are parsed for each called anchor and the reads containing each anchor are collected. For each position downstream of the anchor with multiple observed nucleotides, the local assembly potentially branches into multiple sequences using a simple statistical criterion, depending on the number of nucleotides with frequencies exceeding a fixed threshold. This rule is applied recursively, resulting in extended anchors called "compactors" whose length is at most the input read length and required to have raw read support (Figure 1B, Methods). After this step, each called anchor is associated with a set of compactor sequences along with the set of reads assigned to each compactor during its branching process (Methods). Compactors can be thought of as a simple seed based micro assembly.

Compactors denoise input reads and enable discrimination of splice isoforms, editing events, V(D)J recombinants and sequences outside of the human reference. Unlike any other de novo transcript assembly in use for scRNA-seq, to our knowledge, compactors can be statistically analyzed to quantify the probability that reads supporting an artifactual compactor will be observed (Methods). Compactors also enable huge reductions in computational burden for the number of sequences analyzed in any downstream analysis such as alignment to the genome. In this study, compactors reduced the number 120 fold: from 183,471,175 raw reads to 1,515,555 compactors.

NOMAD+ also includes a classification procedure operating on compactors to assign NOMAD's calls to biologically-meaningful categories such as splicing, among many other categories (Methods, Figure 1C, Supplement), improving interpretability of the NOMAD+ calls and facilitating either targeted or integrative downstream analysis on the anchors within and across multiple categories. Each anchor is classified based on its two most abundant compactors into 6 different categories: splicing, internal splicing, base pair change, 3'UTR, centromere, and repeat (Methods, Suppl. Figure 3). Our classification is based on both computing edit distance between the two compactors for each anchor, and also mapping these compactors to the T2T genome using a spliced aligner (we used STAR (Dobin et al. 2013)). An anchor is classified as splicing if STAR provides a spliced mapping for at least one of the compactors (Methods). When both compactors lack splice junctions, the anchor is classified as internal splicing or base pair change according to the string edit distance between the compactors. The mapping positions for the remaining unclassified anchors are further intersected with annotation databases for 3' UTRs, centromeric repeats, and repeats to classify to one of these categories accordingly. Additionally, soft-clipped portions of the compactor are realigned to the reference genome to allow

for putative transposable elements, circular RNA, and other intra-, inter- or extra-genic transcription (Methods). Through our classification, each compactor becomes part of a family of compactors defined by their shared anchor. Therefore, for each anchor, compactors that fail to map can still be annotated by the annotation of the most abundant compactor for that anchor, which we call *annotation by association* described below. Together, compactors enable interpretation of transcript diversity that can bypass the biases introduced by reference-first approaches, increase the interpretability of calls, and allow a direct comparison of NOMAD+ to existing algorithms: by using references only for interpretability rather than for statistical inference, NOMAD+ obtains statistically valid and unbiased inference, as well as interpretable results

NOMAD+ detects transcript diversity in repetitive RNA loci including RNU-6 and centromeres

We ran NOMAD+ without cell type metadata on 10,326 cells profiled with SmartSeq2 from 19 tissues and 12 donors (29 donor-tissue pairs) and 346 cell types across from the Tabula Sapiens Dataset (Tabula Sapiens Consortium* et al. 2022). 10 tissues, including Blood, Muscle, and Lung had at least two donors, allowing us to analyze reproducibility. NOMAD+ including compactor analysis and downstream steps were run on each donor-tissue pair separately (Methods). Over all tissues, an average of 82.93%, 46.10%, and 28.36% of anchors map to the human genome, the Rfam database of RNA families (Kalvari et al. 2021), and the Dfam database of transposable element DNA sequence alignments (Hubley et al. 2016), respectively. The 3'UTR and splicing classes had the highest and lowest average compactor number per anchor, respectively (Supplement). This finding matches the expectation that most genes have only a handful of highly expressed splicing variants per tissue (Ezkurdia et al. 2015), whereas 3' UTR variation can be driven by diversity of polyadenylation or primer binding sites which can be extensive.

NOMAD+ classified 5.75% (20,891) anchors as centromeric (19,989,187 reads) (Supplement). The centromere was assembled for the first time in the T2T genome (Altemose et al. 2022), but much population-level or single-cell variation could be missing in T2T, which is based on a single cell line. Supporting the limitations of genome alignment, 86% (46,348) of centromeric anchors' compactors and 14% (2,800,038) of reads supporting them failed to align to the T2T genome assembly. 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 been studied in primary cells conditions at single-cell resolution (Altemose et al. 2022). NOMAD+ detected 4,815 anchors containing two consecutive repeats CCATT or its reverse complement which define HSat2. The highest number of distinct compactors for an anchor containing CCATT repeats was 190, which was for an anchor found in 6 donors (donors 1, 2, 4, 7, 8, and 12). Compactor sequence diversity is extensive as illustrated in the multiway alignment (Figure 2A). 81 compactors (48% of reads for the anchor) did not map to T2T, and 53 compactors failed alignment with BLAST and BLAT. We also observed substantial expression variation in multiple cell types including those with proliferative potential (skeletal muscle satellite, mesenchymal, and basal cells of the tongue (Figure 2B). Five T cell types also showed diverse and abundant expression variants. Tongue basal cells (donor 4), thought to be the source of stem cell progenitors (Sullivan, Borecki, and Oleskevich 2010), express 23 of the 26 compactors from this anchor that were found in this donor and tissue, and have cell-specifically expressed variants. Similar diversity is observed in basal cells of donor 7 (Supplement). The enrichment in proliferative cell populations suggest the hypothesis that expression levels of pericentromeric repeats and replication are linked.

We also investigated if NOMAD+ detected expression of Live higher-order repeats (HORs), where the histone variant Centromere Protein A (CENP-A) was found to be bound and is thought to have low transcriptional activity (Hoyt et al. 2022). For anchors classified as centromeric, an anchor's compactors lacking T2T_CenSat annotation are assigned the category defined as the T2T_CenSat annotation of the anchor's most abundant compactor. The category with most assigned reads was HOR_1_5(S1C1/5/19H1L): 2,461 reads supporting 22 distinct compactors, an annotation reflecting a live centromere (Figure 2C). We compared the expression of compactors from the above anchors (pericentromeric and HOR_1_5 repeats)

normalized by the number of cells in which it was detected: 2.1 and 5.6 counts per cell respectively, which is in contrast to previous findings based on single cell line and custom biochemical and bioinformatics pipelines that HORs have very low expression (Hoyt et al. 2022). Also, while (peri)centromeric repeats are expected to be depleted in polyA primed SS2 libraries, normalized expression was within an order of magnitude to moderately expressed transcripts such as *MYL6* (~38 counts/cell) and *GAS5* (~12 counts per cell). This indicates that in contrast to findings of low transcription in the CHM13 cell line, some primary cells have appreciable expression of HOR_1_5 (Hoyt et al. 2022). NOMAD+ analysis of Smart-seq2 enables detection of diverse centromeric RNA expression, difficult or impossible with methodology.

We also used NOMAD+ to identify variant expression in non-coding RNA loci, which are difficult to map due to high copy number, via querying compactors for similarity to the Rfam database using Infernal cmscan (Mistry et al. 2013) , bypassing genome alignment (Methods). The most highly expressed families were ribosomal RNA in Eukaryotes (6.9M reads. 65.99% of reads assigned to Rfam-annotated compactors, Prokaryotes (2.1M reads; 20.64%), Archaea (871K reads; 8.28%), and Microsporidia (432K reads; 4.11%). 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, n.d.) (Supplement). The most abundant non-ribosomal noncoding RNA was U6-snRNA (28K reads; 0.26%), a small nuclear RNA and one of the core components of the spliceosome (Supplement). RNU6 has recently been shown to have high cytoplasmic representation, suggesting this abundance may be expected due to potential polyadenylation (Mabin et al. 2021). More than 75% of compactors assigned to RNU6 by HMMER (RNU6-annotated) failed to map by STAR, presumably due to multimapping to the 1000+ annotated RNU6 loci (Figure 2D). We used the Rfam mapping to define RNU6 compactors by association: the unaligned compactors that share an anchor with at least one compactor matching Rfam and annotated as RNU6 are annotated as *RNU6* compactors by association. We further compared the 71 such compactors to each annotated *RNU6* variant by Hamming distance (Methods, Figure 2E). RNU6 compactors with higher expression had lower minimum Hamming distances to annotated RNU6 variants (data not shown). Hamming distances between U6-directly-annotated and U6-by-association compactors to their best RNU6 reference are comparable, suggesting the U6-by-association compactors are false negatives by Rfam annotation (Figure 2E). Eight distinct annotated *RNU6* or RNU6-pseudogene variants had compactors with unique matches to them with hamming distance 0 that include non-uniform single-cell expression in donor 2 skin (Figure 2F). RNU6-8 perfect-mapping compactors are exclusively expressed in muscle and salivary glands while RNU6-6P compactors are exclusively expressed in skin (Supplement).

More than 30% of RNU6 compactors had substantial (20-40 bps) sequence matching genomic context past the 3' end of the gene (while only 3% mapped upstream of the gene). For example, RNU6-8 and RNU6-6P had perfect alignments that spanned 45 bps downstream of the annotated end and 7 bps upstream of the annotated start, respectively (with 659 and 315 supporting reads, respectively; Figure 2G). Neither RNU6-8 nor RNU6-6P are intergenic. Together, this evidence strongly supports expression of multiple and non-canonical 3' end variants including expression of pseudogenes. To our knowledge, this is the first finding of U6-snRNA variants with extended 3' ends and with multiple variants expressed in primary tissue (Mabin et al. 2021).

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

More than 95% of human genes are known to be alternatively spliced (Pan et al. 2008), but the number of dominant expressed isoforms is mainly based on bulk tissue-level analyses and continues to be debated (Ezkurdia et al. 2015; Arzalluz-Luque and Conesa 2018). This debate is based on alignment-first and reference-based approaches, many focused on specific tasks such as detecting linear alternative splicing,

performing metadata (e.g., cell type) guided testing, and approximate statistical inference due to problems associated with mapping to multiple isoforms (Zheng, Ma, and Kingsford 2022).

We used NOMAD+ to test whether a statistics-first approach could lend clarity to this debate. NOMAD+ reported 20,385 anchor calls classified as splicing across all donors and tissues, including 11,995 and 3,700 unique anchor sequences and genes, respectively. First, we evaluated robustness to biochemical sampling obstacles including dropouts and PCR bias present in SS2 libraries that can be approximated with the negative binomial probability distribution(Jiang et al., n.d.). Under sampling counts from negative binomial distribution identical for each sample which formally violates the null hypothesis of targets are drawn from a distribution independent of cell identity but does not represent real biological signal, Pearson's chi-squared, a widely-used classic statistical test of target-sample independence, calls a high fraction of false positives (FDR >80% at a nominal FDR of 5%), while NOMAD+ retains a maximum FDR of 5% (Methods, Supplement). We also evaluated NOMAD's robustness by recovery of annotated splicing events and reproducibility of calls in the same tissue but between different donors.

73.2% of anchors classified as splicing had compactors mapping to annotated alternative splicing junctions (Figure 3A, Methods). A minority (7.5% or 1,537 anchors) had compactor sequences mapping to an unannotated junction, including 1,387 anchor calls with >10% reads mapping to the unannotated junction. 706 of these anchors found in more than one donor-tissue pair. NOMAD+ improved the reproducibility and precision of the current best performing algorithm for detection of single cell regulated alternative splicing, SpliZ (Olivieri, Dehghannasiri, and Salzman 2022), and predicted complex splicing patterns missed by this method. This improvement by NOMAD+ is noteworthy: it did not use any cell type metadata and was run on subsamples of these tissues, not matched for cell type composition, and thus tissue-donor samples are not formally biological replicates. Focusing on the lung, blood, and muscle - where >1 donor was available, NOMAD+ achieved higher concordance for the same tissue between different donors (Figure 3B, Supplement). In blood and lung, NOMAD+ showed significantly higher reproducibility between donors compared to SpliZ: 77/501 (15.4%) compared to 9/272 (3.3%) for blood; and 202/1250 (16.2%) compared to 75/1289 (5.8%) for lung (Figure 3B). For muscle, both NOMAD+ and SpliZ called almost the same number of unique genes across 3 donors; however, 111 NOMAD+ calls were shared across all three donors, versus only 17 were shared by SpliZ (Supplement). The most highly expressed anchor found in all three muscle replicates and missed by the SpliZ 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 (Figure 3C).

Recent reference-based metadata-guided studies on human cells and experimental validations have found that *MYL6* and genes in the TPM family undergo highly cell type-specific alternative splicing (Olivieri et al. 2021). All these genes were also found by NOMAD+ (Supplement) highlighting its power for detecting cell type-specific patterns even without cell metadata. NOMAD+ re-identified regulated splicing patterns and extended findings to reveal combinatorial expression of isoforms. 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 NOMAD+ 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, NOMAD+ detected 10 distinct spliced isoforms, including 5 novel isoforms (Figure 3D). One anchor 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 novel isoforms detected is denoted 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, a second novel prediction.

NOMAD+ revealed new insights into splicing of *RPS24*, a highly conserved essential component of the ribosome; annotations show > 5 annotated isoforms that include ultraconserved intronic sequence and microexons (Olivieri et al. 2021). NOMAD+ detected 4 isoforms in lung cells from donor 2, respecting our previous findings of compartment specificity for this gene. However, they are also extended, with NOMAD+ identifying a novel isoform containing only a microexon which both STAR and current annotation miss (Figure 3E). Together, this analysis shows that without using any cell metadata or reference before significance calls are made, NOMAD+ has greater sensitivity and reproducibility than SpliZ.

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

2,118 of the genes with splicing anchor called by NOMAD+ are found in more than one tissue, including 10 genes found in 18/19 tissues (Figure 4A). We performed GO enrichment analysis on genes found to have splicing anchors in at least 15 tissues (61 genes). 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.

We call the 10 genes found in 18/19 tissues as *core genes*. NOMAD+ reveals diverse splicing regulation of the 10 core genes: compactors detect more than 71 isoform variants across these 10 genes, including 4 unannotated isoforms, one each in *HNRNPC*, *KMT2E*, *SRSF7*, and *SRSF11*. While each of the core genes are appreciated to have significant regulatory roles, the extent and complexity of their splicing regulation, as revealed by NOMAD+, 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 7, and a novel splicing between the first and fourth exons (Figure 4C). Detected intron retentions are highly compartment-specific, with higher expression fraction for immune and stromal cells compared to epithelial and endothelial cells (Figure 4C): for 9 anchors immune cells have 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. 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* (Figure 4D), 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) as well as intron retention (Supplement). Beyond annotated variants, NOMAD+ identified a novel microdeletion of 6 bases in its 9th exon of *SRSF5* donor 8 prostate, a case of automatic statistical discovery from NOMAD's compactors missed by traditional splicing analysis. 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). NOMAD+ detects a poison exon and intron retention event in *NCOR1* (Figure 4E) 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 missing (Ding et al. 2022; Lareau et al. 2007).

We also used NOMAD+ 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 measured genes with the highest number of variants: across all anchors, donors and tissues *IL32*, *GAS5*, and *RBM39* had the most unique splice junctions (33, 28, and 28 junctions, respectively); 49 genes, including *PRPF38B*, *TACC1*, *CCDC66*, and *TAX1BP3*, had at least 15 distinct splice junctions (Methods, Figure 4F). 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). *PRPF38B* is a splicing factor with prognostic biomarker potential in breast cancer (Abdel-Fatah et al. 2017) with 17 distinct detected splice junctions (across all of its anchors) in 17 tissues. One of its anchors shows compactors with complex alternative splicing involving both 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 4G). NOMAD+ reveals complexity of splicing regulated at the single cell level missed by current methods, and supports the idea that many human genes have cell-specific splicing patterns, rather than exclusively favoring a dominant form.

NOMAD+ detects Alu element insertion polymorphisms *de novo*

We used NOMAD+ to nominate 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). NOMAD+ detects an Alu insertion in two donors for *PSPH*, a Phosphoserine phosphatase in the small intestine of donor 2, and mammary and muscle of donor 4 (Figures 4H). 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 (Figure 4H). In donor 2 muscle, other compactors that map to *PSPH* support a novel unannotated exon skipping event (Figure 4I), also found in 8 other donor-tissues (Methods). In the muscle of donor 4, inclusion of the Alu-inserted varies by cell type (Figure 4I): for example, the Alu-insertion isoform comprises 100% (14/14) of reads in immune macrophage cells and 81.25% (78/96) in stromal skeletal muscle satellite stem cells. NOMAD+ detected other Alu insertions, including in muscle and lung of donor 1 in *ANAPC16*. The majority of compactors (51-65%) in each tissue exhibit a split mapping to the gene *ANAPC16* and Alu elements. Other events were represented in one donor-tissue pair: in the lung of donor 2, >15% of reads originating in the gene AFMID show exonization of an Alu element after an exon (Supplement).

NOMAD+ 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. NOMAD+ unifies this discovery by a hypothesis testing framework: under the null, all cells in any donor have only two

alleles of any fixed splice variant. Under this null, at most two compactors sharing a splice junction should be observed (Methods). We used this framework to validate NOMAD+ calls and prioritize its predictions of transcriptional novelty. Positive controls expected to violate the null include mitochondria where genomes are polyploid (Barrett et al. 1983), as well as the the rearrangement of immunoglobulin loci in B cells and T cell receptor loci which undergo V(D)J recombination, among other examples. Other events also expected to violate the null include post-transcriptionally generated variation, such as RNA editing or repeat expansions.

To investigate global trends, for each anchor, we found distinct compactors across all donor-tissues. Genes annotated as immunoglobulin had the highest number of compactors (Figure 5A). An anchor mapping to immunoglobulin kappa chain (*IGKC*) has the greatest number of compactors–140– across all donors and tissues of any gene (Table 1). Interestingly, these anchors were observed in all immune tissues in our dataset (Figure 5A). Other immunoglobulin genes were also enriched for high numbers of compactors and differentiated from other genes on these purely numerical criteria. Centromeres are thought to have high sequence diversity within and across individuals, thus are expected to have high compactor number across tissues: centromeric anchors defined as containing CCATTCCATT or their reverse complements (Figure 5A), have highest numbers of compactors across donors and tissues. Mitochondrial genes also had high diversity, highest in an anchor with compactors mapping to *MT-ND5,* with 24 compactors in donor 1 lung, the greatest number of compactors generated in a single donor-tissue (Table 3) by any anchor with compactors mapping to an annotated gene. 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, NOMAD+ enables a statistics-first micro-assembly to detect variants in the B cell receptor (BCR) locus avoiding genome alignment. First, we evaluated whether NOMAD+ discovered BCR rearrangements, which we call an "IG-compactor" defined as a compactor mapping to gene with an immunoglobulin annotation (Methods) and matched those found by the state-of-the-art custom pipeline BraCeR (Lindeman et al. 2018)(Supplement). In donor 2 and donor 7 spleen, NOMAD+ detected an IG-compactor in every cell which BraCeR reported a BCR contig, as well as additional cells which BraCeR missed. For example, in spleen plasma B cells from donor 2, not only did NOMAD+ detect IG-compactors in all the 7 cells with BraCeR calls, but it also found IG-compactors in 12 additional cells. Similarly, in donor 7 spleen BraCeR and NOMAD+ found evidence of BCR rearrangement in the same 47 plasma B cells, but NOMAD+ found IG-compactors in 2 additional plasma B cells which BraCeR did not (Figure 5B, Methods). There are instances where NOMAD+ is less sensitive than BraCeR, such as in spleen memory B cells from donor 7, where BraCeR and NOMAD+ both found BCR evidence in the same 68 cells, but NOMAD+ misses 10 cells which BraCeR calls due to NOMAD+'s requirement that IG-compactors be supported by at least 5 reads.

Out of the cells analyzed by both algorithms, NOMAD+ called IG-compactors in 142 cells from donor 2 spleen and 142 and 123 in donor 7 spleen and lymph node. We tested if NOMAD+'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 NOMAD+ IG-compactors for each cell. A high fraction of cells have perfect matches to BraCeR's calls in the same cell: 58.1%, 65.8%, and 64.1% for donor 2 spleen, donor 7 spleen, and donor 7 lymph node respectively, with increasing concordance as for more relaxed minimum Hamming distance criterion for calling a match between IG-compactors with BracerR calls (Figure 5C).

We then investigated NOMAD+ calls missed by BraCeR, further restricting such candidate compactors to have a minimum Hamming distance of greater than 30 bps to all BraCeR contigs, as well as requiring 3 or more compactors per anchor, with either 20 soft-clipped bases, split-mapping, or >4 mismatches to the genome to support their being called due to rearrangement and or hypermutation. NOMAD+ detected 416 anchors with IG-compactors with the above stringent qualities. Another anchor contained 8 compactors, which each aligned perfectly to different IGHV loci, likely representing distinct V segment inclusion. The alignment of one of these compactors is shown as well as the sequence similarity between all eight compactors of the

anchor (Figure 5D). In summary, NOMAD+ automatically detects V(D)J rearrangement agreeing with, but extending that detected by BraCeR in expected B cell subtypes, with implications for downstream biological inference and opportunities to explore other sequences nominated by NOMAD+ 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 impacted by V(D)J. This list includes anchors with compactors showing abundant editing in *ANAPC16* (Figure 6A), a regulator of anaphase, anaphase promoting complex subunit 16 and *AGO2* (Figure 6B), the argonaut protein involved in miRNA targeting, which are targets of canonical Adenosine-to-inosine (A-to-I) RNA editing, the most prevalent form of RNA modification in mammals carried out by Adenosine deaminase acting on RNA (ADAR).

In *AGO2*, an enzyme critical for RNA interference, NOMAD+ generated 18 compactors across donor 2 skin, lung, and donor 4 muscle from the same 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. These compactors support canonical intronic hyperediting in an Alu element defined by A-to-I editing at 5 most commonly edited positions across donor-tissues, in compactors to which 47%, 40%, 36%, 36%, 22%, and 17% of reads were assigned across the three donor-tissues investigated (reported coordinates in hg38 to match REDIPortal chr8 140612531, 140612514, 140612536, 140612540, 140612521, 140612522). REDIPortal only reports 5/6 of these edits, and it reports fewer than 14 of the 9,642 studies in REDIPortal corroborating each event. The A-to-I edit that was missed by REDIPortal occurs at hg38 chr8 140612536 and is supported by 42% of reads assigned to a compactor containing this edit in donor 2 skin, 31% in donor 2 lung and 31% in donor 4 muscle. 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 and position of editing in *AGO2* is highly cell type-specific; 14 of the 20 donor-tissue specific cell types express the 4 unedited alleles at prevalence >25%. However, T cells, goblet and dendritic cells, type II pneumocytes, skeletal muscle satellite stem cells, and CD4-positive alpha beta T cells have no detectable expression of the reference sequence in any of the three donors. More than 75% of assigned reads in donor 4 skeletal muscle satellite stem cells and 100% of assigned reads in donor 2 lung macrophages support a variant with 4 edits. (Supplement). In donor 4 muscle, a compactor composing >90% of reads from CD4-positive alpha beta T cells contained a cluster of 3 A–to-I edits, together with a fourth; this compactor was also observed in mesenchymal stem cells, suggesting a common theme for hyperediting in some stem cells and T cell subsets (Supplement). The extent of editing in these loci and extremely low support from a comprehensive, ultra-deep reference database suggests reads at this locus would be unmapped or mismapped with conventional pipelines (Eisenberg and Levanon 2018).

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 with the largest number of generated compactors in a single donor; this anchor generated 16 compactors in donor 2 muscle (Table 2). Compactors for this anchor represent prevalent base pair changes with respect to the reference (Figure 6C). This apparent editing spectrum lacks a known mechanistic explanation. 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 NOMAD+ was underpowered to detect it. We generated compactors using the above fixed anchor for all cells in the study: 15% of all reads in this dataset containing this anchor have discrepant bases in the 17nt window (chr2: 218217466–218217483), and 11% have base pair changes in the

start codon, much higher than by expected by chance under sequencing error for this study (median error rate .01% for the Illumina NovaSeq 6000) (Stoler and Nekrutenko 2021). High editing rates are observed in diverse tissues including in bone marrow which had a consistent editing rate in two different donors (36% in both donors 11 and 13) (Figure 6C). The reproducibility across donors, tissue specificity, stereotyped positions, and level of diversity are strong evidence against these base-pair changes being an artifact or arising in DNA. In summary, NOMAD+'s automatic statistical inference identifies extensive and novel editing de novo in single cells and in cell types that include high levels of canonical editing in stem and T cell populations; to our knowledge, these events have not been and cannot be detected current custom workflows (Cohen-Fultheim and Levanon 2021).

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

Other anchors with highest levels of diversity show evidence of repeat polymorphism, for example in 3' UTR of *BGN*, compactors show multiple AG dinucleotide repeat lengths (Figure 6D). *BGN* codes for biglycan, which has suggested 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 can be generated during PCR, a process called slippage. Thus, we investigated if polymerase slippage could explain the repeat polymorphisms nominated by NOMAD+. The profile of Taq PCR slippage has been studied, showing a non-negligible probability a repeat is contracted during each PCR cycle. We used this model as it is the only one available, though Kapa HiFi is used in this study (Shinde 2003). Under this error model, allelic variants with dinucleotide repeat regions should be co-detected with variants having a continuum of contracted repeats.

We tested if observed repeat variants in BGN are consistent with the error model. The reference allele reported in the T2T assembly contains 15 AG repeats. In donor 2 lung, the dominant repeat numbers were 15 and 10, (20% and 53% of reads, resp.), inconsistent with being generated in vitro by PCR slippage. Instead, this supports a model that donor 2 has two repeat lengths in *BGN*. To further test if these variable repeat lengths could be explained by PCR artifacts, we generated compactors in tissues NOMAD+ did not call *BGN*: donor 2 muscle and matched tissues: lung and muscle in donor 1. Donor 1 showed no evidence of two repeat lengths: the reference allele comprised 78% of reads, and contractions of 1 or 2 AG dinucleotides accounted for the remaining reads. In contrast, in donor 2 lung and muscle, the reference allele represented 12% of reads (20% in lung and 4% in muscle)–but the dominant repeat length was a contraction of 5 AG repeats called in 60% of reads. >17% of reads had a further contraction to total 6. To analyze single cell variation, we computed the two most abundant repeat numbers per cell: cells in donor 2 express alleles with repeat length modes at 0 and -6 with respect to the reference and no intermediate repeats in the interval (-4,-3) in donor 2 lung. This, together with the donor-specific repeat polymorphism is strong evidence that NOMAD+ calls BGN because single cells express different allelic repeat lengths, perhaps due to allelic imbalance, versus the calls being due to PCR-artifact (Figure 6D).

Other repeat polymorphisms were found in NOMAD+ calls, including in *VSNL1*. Prior literature shows that repeat polymorphism in *VSNL1*, Visinin like 1 protein, a neuronal sensor calcium protein, 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); compactors representing 8%, 7%, and 10% of assigned reads had contractions of 0, 1 and 2 repeats respectively with no intermediate repeats. 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 (Fig 6E). This, alongside the diverse single-cell expression of non-allelic repeat variants, suggest donor-specific somatic diversification of repeat number, likely somatic variation within a donor, rather than PCR, generates these variants. Other NOMAD+ calls had compactors potentially representing repeat variants in transcribed RNA: In donor 1 Lung, *WARS1*,

Tryptophanyl-TRNA Synthetase 1, had 19 compactors. In donor 1 lung, 28.6% of reads from the anchor map to a circRNA with 13 number of repeats, -4 from the reference. This circRNA contains the 5' UTR and coding start and is an example of NOMAD's co-detection of circRNA and repeat variation with respect to the reference (Supplement).

Conclusion

Alignment or pseudo-alignment to human reference genome and transcriptome are thought to be a prerequisite for analysis of RNA-sequencing, and great efforts have been made to provide complete and curated reference genomes and transcript annotation. Similarly, cell type metadata, or its generation, are currently thought to be critical starting or ending points for the analysis of single cell sequencing experiments. In this work, we show that novel biology of transcriptome regulation is 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, NOMAD+'s reference-free approach predicts biology in single cells that has been missed by customized bioinformatics methods in multiple domains. NOMAD+ 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, NOMAD+ unifies and extends discovery. NOMAD+'s findings of complex splicing regulation in splicing factors provides direct evidence of and extends previous predictions of such regulation from EST databases and DNA sequence. We also uncovered novel cell-regulated splicing in myriad other genes, including noncoding RNAs such as the non-coding RNA *GAS5*, together suggesting new candidates for functional studies to prioritize. The extent of splicing diversity uncovered by NOMAD+ 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 shows NOMAD+'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 NOMAD+ 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 NOMAD+, 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 discussion of other dimensions of transcript diversity found by NOMAD+, 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 NOMAD+ 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, NOMAD+ establishes a unified statistics-first approach to sequence analysis, which reveals prevalent transcript diversity regulated in single cells and missed by current bioinformatics. The examples discussed here only scratch the surface of its complexity due to subsampling of human cells and tissues in this study. Analysis of larger single cell data sets as well as of DNA sequencing data may enable a new generation of genetic and transcriptomic analyses as predictors of cellular phenotype or disease. Indeed, NOMAD+ is general, applicable to any RNA-seq or DNA-seq study. Further developments and applications of NOMAD+ promise to enable massive

scale, statistically driven study of transcriptomes, including up and downstream regulation, previously impossible.

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 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 Federation 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.

Tables

Table 1: Anchors, tissue, junction calls, number of distinct compactors, splice junctions for analysis of splicing

All classified_compactors.tsv for the donor-tissues considered in this paper were concatenated. Sequences having homopolymers of length greater than five were excluded, and sequences having one or zero STAR T2T alignments and having one or zero gene annotations a were selected.

Tables 2,3: Donor or donor-tissue, anchor, junction, gene annotation, compactor counts
Anchors with the highest diversity of compactors with identical spliced alignments and gene annotations were investigated, restricting to compactors aligned to one or zero loci (gene name + junction). The number of compactors aggregated by each splice-junction, compactor_gene, donor tuple (Table 2) and for The number of compactors aggregated by each splice-junction, compactor_gene, donor, tissue tuple (Table 3).

Figures:

Figure 1. A general reference-independent alignment-free approach with diverse RNA-Seq analysis applications
(A) Current methods for RNA-seq analysis rely on alignment of the sequencing reads to a reference genome, introducing biases and blindspots and are tailored for specific applications such as quantifying RNA alternative splicing. NOMAD+ provides a unified statistical solution for detecting myriad mechanisms that diversify transcriptomes. It bypasses sequence alignment and rather directly analyzes kmer composition. The core of NOMAD+ is to test if constant kmers (anchors) have non-uniform expression of downstream sequences (targets) across cells. For each observed anchor, NOMAD first creates contingency tables of target counts per sample (here each single cell), and then computes a robust p-value using a closed-form expression that avoids resampling. The resulting anchors with significant p-values have sample-dependent distribution of target counts evidencing regulation at single cell level. NOMAD+ then performs a local assembly approach to reconstruct the compactor sequences, which are then classified and annotated through post-facto steps to categorize NOMAD+ calls for interpretation, such as splicing, V(D)J recombination. (B) Anchors called by NOMAD+ are extended through a local assembly approach to generate “compactors” that are used in the subsequent classification step for inference and downstream analysis. (C) Compactors for the called anchors by NOMAD+ are classified into multiple biologically relevant groups: splicing, base pair change, internal splicing, 3’UTR, centromere, and repeats. The softclipped bases for the compactors that have been aligned to the genome through softclipping are realigned to the genome using Bowtie aligner and are used to find

compactors with evidence for circular RNA, gene fusions, strandcrosses. Compactors that fail to map are annotated by the identity of other compactors in its family, called *annotation by association.*

Figure 2: NOMAD+ detects differential RNU6 spliceosomal RNA compactor usage per cell type in TSP4 Muscle and TSP2 Skin
(A) The sensitivity of NOMAD+ to variation in the centromeric repeat array is illustrated for the anchor ATTCCATTCCATTCCATTCCATTCCAC, having 5 contiguous repeats of the canonical pericentromeric repeat ATTCC. Of anchors containing the subsequence ATTCCATTCC, this anchor has the largest number of compactors across donor-tissues. Multiway alignment shows the 190 compactors detected for this anchor ordered concordantly. 53 of the 190 compactors did not BLAT or BLAST-align. (B) The heatmap shows the natural log of read counts assigned to each compactor per cell type collapsed across the donors and tissues in which the anchor was called; rows and columns are ordered by descending sum. (C) Dot plot showing the number of compactors versus the fraction of reads supporting compactors that did not STAR align for centromeric anchors subcategorized as a higher-order repeat (HOR) array. The dot size corresponds to the number of reads in each category. The STAR-unaligned compactors were subcategorized by the annotation of the most abundant compactor sharing their anchor. 15 HOR categories with the most assigned reads are shown. (D) *RNU6* has 1,281 gene and pseudo-gene loci scattered throughout the human genome. A randomly selected subset of 100 *RNU6* reference sequences and the 71 *RNU6* compactors were multiway aligned with ClustalOws showing high conservation. Differential usage of *RNU6* compactors per cell. (E) Direct and by-annotated *RNU6* compactors have comparable abundances and sequence similarity to *RNU6* reference genes. (F) *RNU6-6P* pseudo-gene mapping compactor, and the other *RNU6* mapping compactors from the same anchor which do not have perfect and unique mapping. Heatmap shows differential compactor usage per cell across donor 2 skin. (G) Perfect multiway alignments of compactors to *RNU6* loci show alignment both upstream and downstream of the annotated regions.

Figure 3: NOMAD+ enables de-novo analysis of alternative splicing in single cells. (A) Dot plot of the number of distinct anchors classified as splicing in each pair of donor and tissue. >55% of the splicing anchors found in each donor and tissue involve an annotated alternative splicing event. (B) Upset plots comparing the concordance of the significant splicing genes called by NOMAD+ (red) and SpliZ in two replicates of lung (left) and blood (right). NOMAD+ achieves higher concordance of the called genes in the same tissue from different donors compared to SpliZ despite not using cell identity, which have different distributions and or composition in the two donors. (C) The plot shows that the compartment-specific alternative splicing of *GAS5* is reproducible in muscle cells from 3 different donors. The error bar for each dot shows the 95% binomial confidence interval for each isoform fraction. In all 3 donors, CD8+, alpha-beta t cells possess higher inclusion rate for the isoform with shorter intron. (D) Heatmap showing the fraction of eight *CD47* isoforms detected for an anchor of this gene in each single cell across 10 donors and 14 tissues. Cells with >5 reads are included and horizontal side bars show the donor, tissue, and compartment of each cell. Cells are sorted based on hierarchical clustering applied directory to the heatmap. NOMAD+ detects extensive expression variation, including novel isoforms. Among other detected *CD47* isoforms is a novel isoform (shown in yellow) with annotated junctions (chr3:110767091:110771730--chr3:110771761:110775311). (E) Extensive alternative splicing of *RPS24* for an anchor with 4 different splicing isoforms. The alternative spliicng involves inclusion/exclusion of two cassette exons in ultraconserved regions, including a microexon of 3 bps. NOMAD+ detected a novel isoform which includes only the micro exon detected. The multiway alignment confirmed the inclusion of the microexon while both STAR and BLAT were unable to detect the microexon. The heatmap shows the fraction of these 4 different isoforms across lung cells from donor 2, which also shows the compartment-specific alternative splicing where the isoform with both exons included is predominantly expressed in epithelial cells.

Figure 4: Splicing factors and histone modification genes are enriched in genes with pan-tissue regulation of their alternative splicing. (A) Bar plot showing that 57% of the genes (2,118 genes) with significant splicing found in at least 2 tissues, including 10 genes found in at least 18 tissues. 8 of these genes are either splicing factors (shown in green) or histone modifications (shown in brown). (B) GO enrichment analysis of genes found in >15 tissues revealed enrichment for pathways related to mRNA processing and splicing regulation (Fisher test, FDR corrected p-value< 0.05). (C) Extensive alternative splicing of *HSP90AA1*,

the only gene found to have alternative splicing in all 19 tissues. We show the 12 anchors for this genes that involve distinct intron retention events in 7 different introns. For 6 of these anchors, we show the fraction of each splicing isoform in each compartment. Intron retentions are shown in pink and splice isoforms are shown in light green, except for anchor 1 whose second splice isoform is shown in dark green. Intron retentions are compartment-specific, with immune cells having the highest fraction for most of the anchors. The barplot on the right shows the total read count for the intron retention and splicing isoform for each anchor. For two of the anchors (anchors 1 and 3) that have the most compactors (3 for anchor 1) and were found in the most tissues (anchor3), we show the haetmaps for the fraction of each isoform in each single cell. (D) The plots showing the fraction of each isoform for (D) *HNRNPDL* and *NCOR1* (E) in each cell. (F) Plot shows the number of tissues and number of unique splice junctions for each gene with detected splicing anchor. Genes *IL32*, *RBM39*, and *IGKC* have the most unique splice junctions. (G) Alternative splicing of gene *PRPF38B* which involves an intron retention and 6 alternative isoforms. This gene is among the genes with the most diverse isoform structure with significant alternative splicing in 11 donors and 17 tissues. (H) Distribution of the compactor abundance for the anchor with Alu-insertion in *PSPH* in muscle cells from donor 4. (I) Model depicting an Alu element disrupting a normal splicing program. The Alu element inserts into the middle of an exon, inhibiting the formation of the typical isoforms and resulting in the formation of a new isoform, which consists of a portion of the original exon with the Alu element. The relative abundance of each isoform can be detected from the distribution of the compactors, in which the Alu-inserted isoform is represented by compactors that map uniquely to the human genome with some portion that maps to transposable elements. A potential Alu-insertion isoform in 3 tissues across two donors, in the gene *PSPH*. 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. A potential Alu-insertion event in *ANAPC16* in the muscle of donor 1, in which the compactor maps to the exon containing the UTR. A potential Alu-insertion event in AFMID in the lung of donor 2.

Figure 5: Global trends of compactor diversity reveal V(D)J recombination has highest levels of transcriptome diversity and BraCeR comparison.
(A) The set of all NOMAD+ called anchors are subset to those having one or zero STAR alignments. For each anchor, the number of donors in which this anchor was found and the anchor's total number of compactors are computed (log scale). Color corresponds to four categories: 'CCATTCCATT', which indicates that the anchor contains this centromeric repeat motif or its reverse complement; 'IGKC', 'IGH', and 'IG', indicating the compactor gene annotation contains one such substring (if *IGKC*, categorized as *IGKC* rather than IG) ', 'mitochondrial', and 'other', which contains all other anchors. The top marginal histogram shows the probability that each category falls into a ¼ x-unit range, for example [0, ¼), [¼ , 1) etc. The right marginal histogram shows the probability that each point occurs at each value of 'number of samples' (logarithmic scale). Multiple sequence alignment of the anchor with the most distinct compactors. (C) Comparison of NOMAD+ and BraCeR for donor 2 and donor 7 spleen among cells analyzed by both algorithms. Each dot represents a cell which is BraCeR+ if a BraCeR contig was called, and NOMAD+ if an IG-compactor was found with stringent filters (Methods) and expression over 5 counts. The value on the x-axis is the expression of the maximally expressed IG-compactor found in that cell. NOMAD+ largely agrees with BraCeR but both (i) missed calls BraCeR makes such as in donor 7 memory B cells (red), and (ii) finds IG events in B cells that BraCeR does not, such as donor 2 plasma and memory B cells (green). (D) Fraction of cells run through both BraCeR and NOMAD+ where at least one IG-compactor has Hamming distance less than threshold to at least one BraCeR contig in the same cell. (E) *IGHV3* alignment of multiple compactors from the same anchor, showing split mapping to *IGHV3-53-201* and *IGHV3-53-201*.

Figure 6: RNA Editing and Repeat polymorphism.
Multiple sequence alignment of compactors generated de novo by NOMAD+ for (A) *ANAPC16* and (B) *AGO2*.. A-to-I edits are colored red, and the compactors matching the reference allele are shown by orange boxes in each donor-tissue. Marginal histograms show the number of reads for each compactor sequence. The axis line indicates a count of 100 reads. A predicted miRNA binding site in *ANAPC16,* which is disrupted by observed edits in all four donor-tissues (Chen and Wang 2020). Also a circular RNA was identified in the third compactor from donor 2 skin in *AGO2*. BLAT displays the position of *ANAPC16* compactors on the 5' UTR and within an

Alu repeat region. (C) *ARPC2* sequence diversity having 20 distinct variants occurs in a region of 17 bps, i.e., positions 58-74 in the compactor sequences. Multiple sequence alignment of 20 17mers in *ARPC2* exon 2, where base-pair changes occur between the 5' UTR and translational start (T2T chr2: 218217466–218217483) with all kmer counts from donor bone marrow at rate orders of magnitude higher than expected under sequencing error. (D) and (E) Barplot and scatterplots of single-cell resolved repeat length in *BGN* and *VSNL1*. Blue and pink colors represent a donor-tissue called by NOMAD+ and a compactor generation analyzed as a control. Low single cell variation is expected. Each cell contributes one vote for each of its two most abundant variants to scatterplots univariate histograms, quantifying expression of single cells' two alleles across a donor-tissue. If a cell displays only one repeat variant, it votes twice for this repeat number. In the scatterplot, the variant with smaller repeat number presents its repeat number on the x-axis, and the variant with the larger repeat number presents on the y-axis. Dots are orange if a cell's most abundant allele's repeat number is greater than its second-most abundant allele's repeat number, and dots are blue if the most abundant allele's repeat number is smaller than that of its second-most abundant allele. Dot sizes correspond to the number of cells which possess this combination (x >y) of repeat variants.

Code and data availability

The code used in this work is available at https://github.com/salzman-lab/nomad and at https://github.com/salzman-lab/compactor.

Methods

NOMAD overview

NOMAD 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 including (nut not limited to): alternative splicing, RNA editing, V(D)J recombination, and chimeric RNAs such as gene fusions, inversions, and circular RNAs. Not requiring computational alignment of the reads to a reference genome, a feature commonplace in conventional methods for RNA expression analysis, NOMAD can bypass inherent biases and blindspots in aligners leading to discoveries not possible by other methods. NOMAD searches for sequences of certain length (anchors) which are followed by diverse sequences (targets). Then for each extracted anchor, a contingency table containing the read counts for each target and each sample is generated. NOMAD then performs a statistical test with closed-form solution for valid p-value to find anchors with significant sample-dependent target count 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}. P-values for each anchors are corrected for multiple testing across generated random partitions (L_num_random_Cj) and number of generated random hashes for each partition (K_num_hashes). P-value for each anchor is corrected for multiple testing across the number of partitions and hashes.

NOMAD runs

NOMAD was run in unsupervised mode on the fastq files from each donor and tissue in Tabula Sapiens data set. 19 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). We randomly selected 400 annotated cells with cell type information from a tissue and donor, if it had more than 400.Eight donor-tissues had fewer than 400 cells: Trachea TSP2 (119 cells), Eye TSP5 (134 cells), Blood TSP1 (138 cells), Tongue TSP4 (209 cells), Heart TSP12 (277 cells), Eye TSP3 (291 cells), Trachea TSP6 (358 cells), Kidney TSP2 (370 cells). Because 400 cells were sampled for each donor and tissue (except 8 tissues with between 119 and 370 cells (Methods)), cell number normalization is implicit, and read depth is approximately so. In total, we ran NOMAD on 13,500 SmartSeq2 cells from 136 cell types. NOMAD was run with default parameters except for

the following parameters for number of random partitions for input cells and number of random hashes for each partition (Chaung et al. 2022): L_num_random_Cj = 300 and K_num_hashes = 10.

Anchors with |NOMAD effect_size| > 0.2, target Levenshtein distance >1, number of reads assigned to the anchor > 50, and observed in >10 samples (cells) selected for compactor generation and downstream analysis.

SpliZ Runs

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 compute. To compare NOMAD with SpliZ (as a state-of-the-art reference-dependent method), we ran SpliZ on the same Tabula Sapiens dataset. We first aligned reads to human hg38 reference genome using STAR and then ran the STAR BAM files through SICILIAN (Dehghannasiri, Olivieri, and Salzman 2020), which is a statistical wrapper for detecting high-confidence splice junctions from spliced aligners. We then applied SpliZ to the detected splice junctions. SpliZ was ran on data from each donor separately and to avoid calling genes that have tissue-specific splicing rather than cell type-specific splicing, its statistical test was performed separately across cell types within each tissue from that donor.

Comparison to STAR and SpliZ

After running SpliZ and obtaining the list of genes with cell type-specific splicing for each donor and tissue called by SpliZ, we compare the list of significant genes found by NOMAD and SpliZ for each donor and tissue separately. Since NOMAD calls are the anchor-level and not gene-level to have consistent comparison with SpliZ whose calls are at the gene-level, we use the convention that a gene is found to be significant by NOMAD if it calls at least one “splicing” anchor for that gene.

Compactor generation

Reads containing significant anchors are aggregated across FASTQs, and read segments upstream of the anchor sequence are discarded. For each anchor, reads are traversed left-to-right and nucleotides at each position are tested for support in the reads. If at least 20 reads present a particular nucleotide and the read number exceeds 10% of reads on the current branch (or 5 reads and 80%), then reads containing this nucleotide are branched to be traversed and tested independently, and this nucleotide is appended to their representative compactor sequence. This rule is applied recursively, resulting in subsets of all anchor-reads each represented by a distinct compactor sequence.

Compactor Pfam, Rfam, and BLAST alignment.

We submit all STAR-unaligned compactor sequences to hmmsearch for alignment to the Pfam and to cmscan search of the Rfam database. For BLAST, we produce the following two compactor subsets for each anchor:

We first collect the set of sequences which have been unaligned by STAR.

1. If the anchor has 100 or more compactors, we take the 10 most abundant compactors to be BLAST-aligned. If the anchor has fewer than 100 compactors, we take the 2 most abundant compactors to be BLAST-aligned.

We then take the union of these subsets and submit to BLAST with the following parameters:

```
-evalue 0.1     -dust no   -word_size 24   -reward 1  -penalty -3
-max_target_seqs 4
```

Reference-free classification of compactors into biologically-relevant categories

To increase interpretability of the inferred compactors and to be able to perform targeted downstream analysis for a specific RNA diversity event, we assign an RNA event to each anchor based on its compactors and the features directly derived from the compactors. We consider six categories for anchors: 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 unclassified. We take a hybrid approach for assigning classes to the anchors, where some classes are assigned independently from the alignment to a reference genome (e.g., internal splicing and base pair change) and some classes are assigned based on the reference genome (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 (i.e., those with the highest fraction of anchor reads) for each anchor. If one of the compactors is longer than the other one, we consider its substring that is of the same length as the other compactor. We then compute two different distance metrics: hamming distance and Levenstein distance. Both strings should be of the same length to be able to compute Levenstein distance. We should note that both Levenstein and hamming distance are computed when the anchor sequence is removed from the beginning of each compactor sequence. Anchors with the same hamming and Levenstein distances are classified as mutations as this criterion indicates that only substitutions (i.e., nucleotide changes) can explain the difference between the compactors.

If the Levenstein operations sequence for an anchor includes a number of consecutive insertions and deletions, the anchor is classified as "internal splicing" if: `Levenstein_distance <(run_length_D+run_length_I+1)`, where `run_length_D` and `run_length_I` are the longest stretch of deletions and insertions in the Levenstein operations sequence, respectively.

Reference-based classification

To find anchors that can be explained by an alternative splicing, we map the two compactors for each anchor to the reference genome using a spliced aligner (we used STAR, but any other splice aligner could be used). We aligned compactors to the reference human genome T2T using STAR in two-pass mode 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, mapping chromosome, mapping coordinate, CIGAR string, and number of mismatches from STAR BAM file (1st, 2nd, 3rd, 4th, 6th, and 16th columns). If at least one of the compactors for an anchor involves a split mapping and the hamming distance and Levenstein distance are not equal, we classify the anchor as "splicing", as at least one of the compactors involves a splice junction and the difference between the compactor sequences cannot be explained by simple substitutions. Note that both compactors should overlap to the same gene.

For human genome, for anchors that have not been classified as splicing, base pair change, and internal splicing, we further intersect compactor mapping positions with the 3'UTR coordinates, centromere satellite element coordinates in T2T CenSat database and repetitive element coordinates in RepeatMasker database and classify anchors to one of these categories accordingly.

Soft-clipping and realignment:

Compactors that have been aligned by STAR through softclipping can provide evidence for RNAs not part of the reference transcriptome such as circular RNAs, gene fusions, and strandcrosses. As a systematic approach for utilizing compactors with soft-clipping to infer such RNAs, we select those compactors with >20 soft-clipped bases in which the longest stretch of any nucleotide A, G, C, T is shorter than 6 and realign the softclipped part to the genome using STAR. We use the information from the original alignment of the compactor with the realignment of its softclipped part to infer if the compactor can provide evidence for circular RNA, gene fusion, and strandcross.

For a softclipped compactor to be classified as circular RNA, we look at the mapping positions and the strand orientation of the original compactor and softclipped part. Let m_C and m_S be the mapping positions for the compactor and its softclipped part, respectively, and also s_c and s_s be the strand orientations for the compactor and its softclipped part, respectively. Assuming that the softclipped part is at the start of the compactor, we classify a compactor as circular RNA, if it satisfies one of the following conditions and both alignments map to the same chromosome:

- (s_c == +) & (s_s == +) & m_s > m_c
- (s_c == -) & (s_s == -) & m_s < m_c

Similary, if the softclipped part is at the end of the compactor, one of the following conditions should be met by a compactor to be classify as a circular RNA:

- (s_c == +) & (s_s == +) & m_s < m_c
- (s_c == -) & (s_s == -) & m_s > m_c

To assign strandcross to a compactor, both alignments should have the same chromosomes but different strand orientations. Finally for gene fusion compactors, either the mapping chromosomes for the compactor and its softclipped part are different or they map to the same chromosome and strands but with a distance of at least 10^6 bases. Note that for a compactor to be assigned to one of these classes, both the compactor and its softclipped part should be uniquely mapped by STAR.

Supervised Compactor Generation

NOMAD-called anchors from a single donor-tissue can be missing from the NOMAD calls in other donor-tissues. This can be caused by NOMAD's downsampling of input FASTQs and by inconsistencies in an anchor's signal between donor-tissues. To investigate anchor sequences called in one donor-tissue but not another, we used NOMAD-called anchors as seeds for compactor generation in donor-tissues where NOMAD did not call the anchor.

snRNA-RNU6

There were 670 unique compactors called as snRNA-RNU6 by RFAM but 586 of these contained a homopolymer of length 5 or larger and were filtered out. We performed compactor-assignment-by-association: which further annotated 59 unique non-homopolymer compactors as RNU6 for a total of 143. Further filtering to RNU6 compactors perfectly represented by a sequence read resulted in 71 compactors.

NOMAD+ IG-compactor comparison to BraCeR

To test whether NOMAD+ was detecting IG-compactors with perfect sequence similarity to BraCeR contigs from the same cell, we calculated the minimum Hamming distance between all IG-compactors and all BraCeR contigs. IG-compactors are defined as NOMAD+ compactors which map to an immunoglobulin heavy, light, or kappa chain gene by STAR, allowing for mismatches and soft-clipping. We further define IG-anchors as anchors which have a plurality of compactors (at least 20%) mapping to immunoglobulin genes. IG-anchors allow us to annotate-by-association compactors which are unmapped by STAR, but have the same anchor as predominantly IG-mapping compactors. The minimum Hamming distances of all IG-anchor compactors were calculated against all BraCeR contigs downloaded for the same donor from the Tabula Sapiens AWS bucket. *s3://czb-tabula-sapiens/Pilot*/immune-repertoire-analysis/bracer*, where the * is 2 for donor 2 etc.

Setting BraCeR contigs as ground truth, we estimate the true-positive and false-positive rates (TPR and FPR) of NOMAD+ by constructing a square binary matrix M, identically indexed on rows and columns by the cell-ids that were run through both NOMAD+ and BraCeR. The i,j entry in the M matrix is 1 if there is a perfect alignment, Hamming distance 0, of at least one compactor from cell-i to any of the BraCeR contigs from cell-j, otherwise M[i,j] = 0. From this table we define the TPR as the sum of the diagonal entries of M, M[i,i], divided by the number of diagonal entries, which is again the number of cells run through both BraCeR and NOMAD+. Similarly, the FPR is the sum of off-diagonal entries divided by the number of off-diagonal entries. Note that if a single compactor has perfect alignment to multiple contigs, or vice-versa. each of those pairs will contribute to either the TPR or FPR.

We define another category of IG-compactors as “interesting” IG-compactors which are a more stringently filtered subset of IG-compactors. We require that the compactor is (i) IG-compactor (ii) IG-anchor (iii) has 10 or more compactors unique per anchor (iv) has a total anchor abundance greater than 100 (v) must be STAR aligned (vi) has a minimum Hamming distance larger than 30 to BraCeR contigs, and finally (vii) has either more than 20 soft-clipped bases, split-mapping, or has more than 5 mutations with respect to the reference. These interesting IG-compactors were used to test if NOMAD+ was extending past BraCeR by discovering BCR recombination events in cells which did not have BCR contigs. When calculating the TPR and FPR above we ensured that the IG-compactors and BraCeR contigs were only counted if they had perfect sequence similarity, but for this analysis to test NOMAD+ sensitivity beyond the BraCeR calls, sequence comparisons were not made. Instead a cell was called BraCeR+/NOMAD+ if it contained at least one BraCeR contig and at least one interesting IG-compactor with 5 or more reads in that cell.

Data: Gene counts

Gene count tables for SS2 data were downloaded from the Tabula Sapiens AWS bucket.

Classification of anchors by alignment to reference databases

Anchors were assigned to categories by their alignment to a set of reference databases. Bowtie2 was used to align anchors to references; default parameters were used. Anchors are categorized by their top annotation, reported according to the following priority: false positive sequences (such as Univec, Illumina adaptors), Rfam, transposable elements (Dfam), spacers, and the human genome (hg38).

Aggregation of anchors abundance over cell types

To quantify anchor abundance on the cell type level, we use a compactor output file, which contains per-sample counts for each anchor-compactor. The sample identities are converted to cell type values, with the use of cell-level metadata. Counts are then summed, to provide aggregated counts per anchor-cell type. Anchors are converted to their compactor classification values and further aggregated via summation, to provide counts per classification-cell type.

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 bases 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.

Computation of compactor p-values

We additionally utilize the same contingency table test on the compactor x sample contingency table, to generate a statistically valid p-value bound testing whether the distribution of compactors is the same across samples or not. We utilize a similar procedure to that for generating the initial p-values on the target x sample count matrices, selecting 50 pairs of random c and f, and taking the minimum p-value across these random c and f after applying Bonferroni correction. The analysis techniques are identical to those used in the original NOMAD paper (Meyer et al. 2022). We additionally utilize alternating maximization-based c and f, where p-values are derived using a sample splitting approach, recently derived in (Bharav et al, 2022).

Abundance of third most abundant target:

We can provide statistically valid p-value bounds for observing large counts of the third most abundant target under a 2-target null hypothesis.

For M reads, with sequencing error rate epsilon (each basepair independently undergoes a substitution error with probability epsilon, uniformly erroring to one of the other 3 basepairs). $D_{KL}(p,q)$ denotes the Kullback-Liebler divergence between two independent Bernoulli random variables with heads probabilities p and q. Here, we bound the probability of observing more than T counts of the third most abundant length k target.

$$2\sum_{\ell=1}^{k}\binom{k}{\ell}3^{\ell}\exp\left(-MD_{\mathrm{KL}}\left(\frac{T}{M}\,\middle\|\,(1-\epsilon)^{k-\ell}(\epsilon/3)^{\ell}\right)\right)$$

The derivation for this is provided in the supplemental methods, and we can see that this bound decays very quickly as a function of T. For example, for M=50 and epsilon=0.01, T=4 yields a p-value of .01, T=5 yields a p-value of 1E-5, and T=7 yields a p-value of 1E-10. This can naturally be extended to bounding the probability of observing many counts for the j+1st target, given that only j targets truly exist without errors.

Robustness to biochemical sampling error models

After biochemical sampling, counts in single cell RNA-seq are often overdispersed. Thus, counts are often modeled as poisson with a random (gamma distributed) mean, which is equivalent to the widely used negative binomial distribution. This has been studied in DESeq2 (Love, Huber, and Anders 2014), where the authors show that the overdispersion can be well modeled as a function of the sequencing depth, where low sequencing depth leads to higher dispersion than would be expected under the poisson null. We simulate NOMAD's robustness to overdispersion, showing that under the overdispersion modeled by DESeq2, NOMAD provides much better control of the false discovery rate against this biological null (as opposed to the statistical

null), whereas a classical chi-squared test immediately begins to aggressively reject once the null is no longer satisfied (dispersion > 0). For example, for 20 observations per sample, with 10 equally likely targets and 20 samples, NOMAD still controls the FDR below 5% while the chi-squared test has an FDR of approximately 87%. Simulations in Supplement.

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.

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

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 Baharav, Ivan Zheludev, and Julia Salzman. 2022. "A Statistical, Reference-Free Algorithm Subsumes Myriad Problems in Genome Science and Enables Novel Discovery." *bioRxiv : The Preprint Server for Biology*, June. https://doi.org/10.1101/2022.06.24.497555.

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

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.
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.
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.
Hoyt, Savannah J., Jessica M. Storer, Gabrielle A. Hartley, Patrick G. S. Grady, Ariel Gershman, Leonardo G. de Lima, Charles Limouse, et al. 2022. “From Telomere to Telomere: The Transcriptional and Epigenetic State of Human Repeat Elements.” *Science* 376 (6588): eabk3112.
Hubley, Robert, Robert D. Finn, Jody Clements, Sean R. Eddy, Thomas A. Jones, Weidong Bao, Arian F. A. Smit, and Travis J. Wheeler. 2016. “The Dfam Database of Repetitive DNA Families.” *Nucleic Acids Research* 44 (D1): D81–89.
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.
Jiang, Ruochen, Tianyi Sun, Dongyuan Song, and Jingyi Jessica Li. n.d. “Statistics or Biology: The Zero-Inflation Controversy about scRNA-Seq Data.” https://doi.org/10.1101/2020.12.28.424633.
Kalvari, Ioanna, Eric P. Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasinska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, et al. 2021. “Rfam 14: Expanded Coverage of Metagenomic, Viral and microRNA Families.” *Nucleic Acids Research* 49 (D1): D192–200.
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.
Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and

Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550.
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. n.d. "Single Cell Transcriptomics Reveals the Hidden Microbiomes of Human Tissues." 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.
Meyer, Elisabeth, Kaitlin Chaung, Roozbeh Dehghannasiri, and Julia Salzman. 2022. "ReadZS Detects Cell Type-Specific and Developmentally Regulated RNA Processing Programs in Single-Cell RNA-Seq." *Genome Biology* 23 (1): 226.
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.
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.
Sullivan, Jeremy M., Alexander A. Borecki, and Sharon Oleskevich. 2010. "Stem and Progenitor Cell Compartments within Adult Mouse Taste Buds." *The European Journal of Neuroscience* 31 (9): 1549–60.
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.
Tavor Z. Baharav, David Tse, Julia Salzman. n.d. "NOMAD: An Efficient Finite-Sample Valid Test for Contingency Tables with Applications to Computational Genomics." *In Preparation*.
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.
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.
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.

Figure 1.

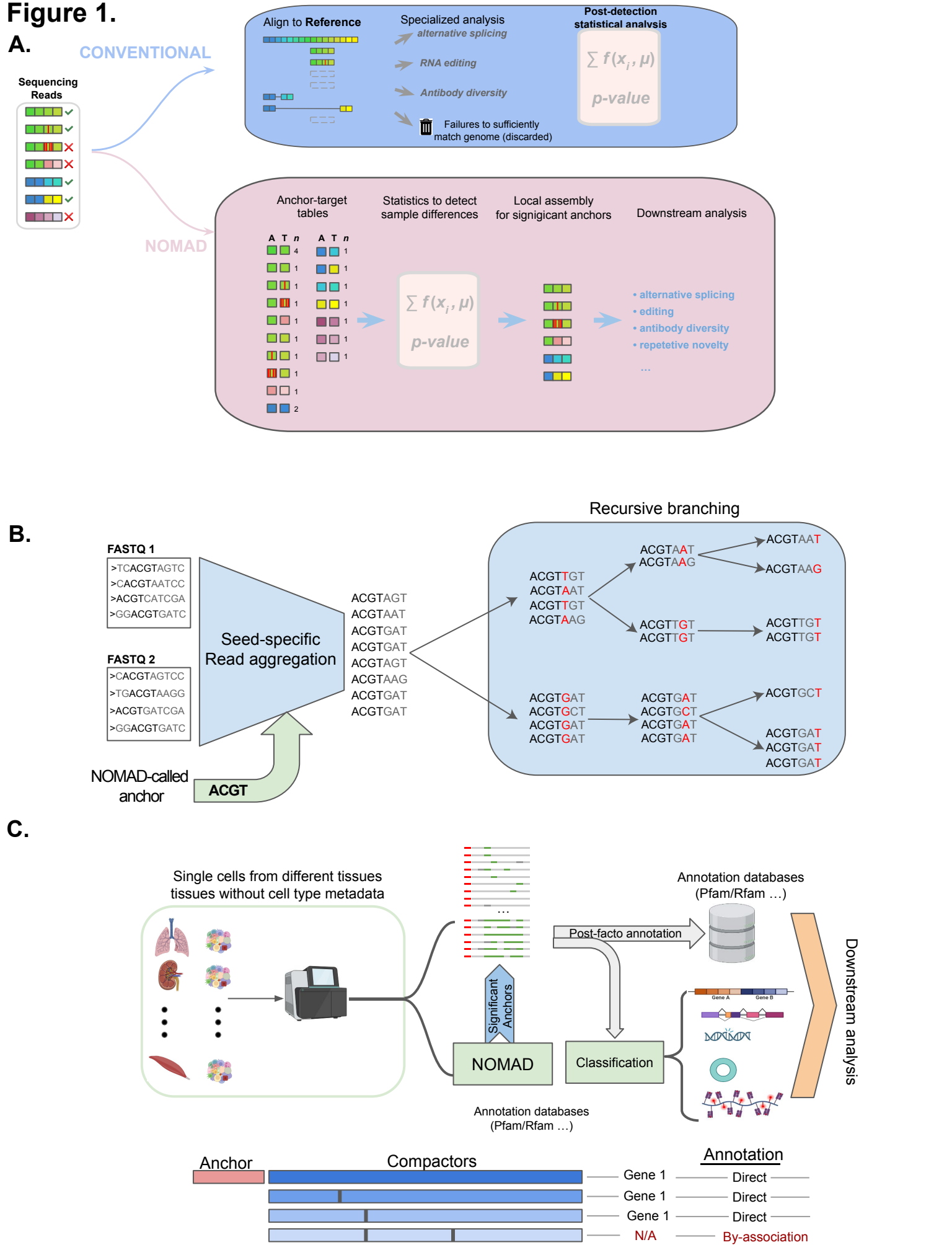

A.
CONVENTIONAL
Sequencing Reads
Align to Reference
Specialized analysis
alternative splicing
RNA editing
Antibody diversity
Failures to sufficiently match genome (discarded)
Post-detection statistical analysis
∑ f(x_i, μ)
p-value
NOMAD
Anchor-target tables
Statistics to detect sample differences
Local assembly for signigicant anchors
Downstream analysis
A T n
∑ f(x_i, μ)
p-value
• alternative splicing
• editing
• antibody diversity
• repetetive novelty
...
B.
FASTQ 1
>TCACGTAGTC
>CACGTAATCC
>ACGTCATCGA
>GGACGTGATC
FASTQ 2
>CACGTAGTCC
>TGACGTAAGG
>ACGTGATCGA
>GGACGTGATC
Seed-specific Read aggregation
NOMAD-called anchor
ACGT
ACGTAGT
ACGTAAT
ACGTGAT
ACGTGAT
ACGTAGT
ACGTAAG
ACGTGAT
ACGTGAT
Recursive branching
ACGTTGT
ACGTAAT
ACGTTGT
ACGTAAG
ACGTAAT
ACGTAAG
ACGTTGT
ACGTTGT
ACGTAAT
ACGTAAG
ACGTTGT
ACGTTGT
ACGTGAT
ACGTGCT
ACGTGAT
ACGTGAT
ACGTGAT
ACGTGCT
ACGTGAT
ACGTGAT
ACGTGCT
ACGTGAT
ACGTGAT
ACGTGAT
C.
Single cells from different tissues
tissues without cell type metadata
Significant Anchors
NOMAD
Post-facto annotation
Annotation databases
(Pfam/Rfam ...)
Classification
Gene A
Gene B
Downstream analysis
Annotation databases
(Pfam/Rfam ...)
Anchor
Compactors
Annotation
Gene 1
Direct
Gene 1
Direct
Gene 1
Direct
N/A
By-association

Figure 2.

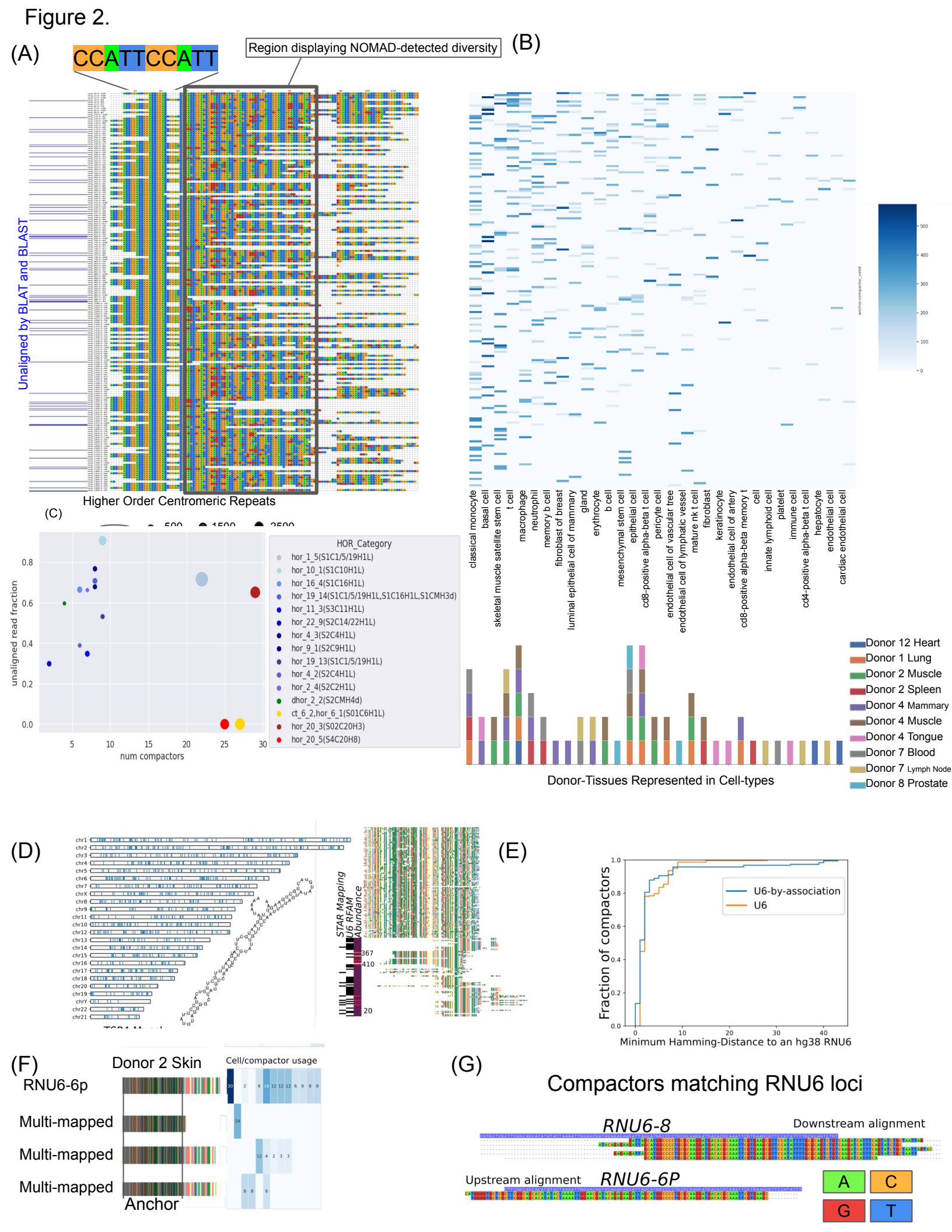

(A)
CCATTCCATT
Region displaying NOMAD-detected diversity
Unaligned by BLAT and BLAST
Higher Order Centromeric Repeats
(B)
classical monocyte
basal cell
skeletal muscle satellite stem cell
t cell
macrophage
neutrophil
memory b cell
fibroblast of breast
luminal epithelial cell of mammary
gland
erythrocyte
b cell
mesenchymal stem cell
epithelial cell
cd8-positive alpha-beta t cell
pericyte cell
endothelial cell of vascular tree
endothelial cell of lymphatic vessel
mature nk t cell
fibroblast
keratinocyte
endothelial cell of artery
cd8-positive alpha-beta memory t cell
cell
innate lymphoid cell
platelet
immune cell
cd4-positive alpha-beta t cell
hepatocyte
endothelial cell
cardiac endothelial cell
Donor 12 Heart
Donor 1 Lung
Donor 2 Muscle
Donor 2 Spleen
Donor 4 Mammary
Donor 4 Muscle
Donor 4 Tongue
Donor 7 Blood
Donor 7 Lymph Node
Donor 8 Prostate
Donor-Tissues Represented in Cell-types
(C)
unaligned read fraction
num compactors
HOR_Category
hor_1_5(S1C1/5/19H1L)
hor_10_1(S1C10H1L)
hor_16_4(S1C16H1L)
hor_19_14(S1C1/5/19H1L,S1C16H1L,S1CMH3d)
hor_11_3(S3C11H1L)
hor_22_9(S2C14/22H1L)
hor_4_3(S2C4H1L)
hor_9_1(S2C9H1L)
hor_19_13(S1C1/5/19H1L)
hor_4_2(S2C4H1L)
hor_2_4(S2C2H1L)
dhor_2_2(S2CMH4d)
ct_6_2,hor_6_1(S01C6H1L)
hor_20_3(S02C20H3)
hor_20_5(S4C20H8)
(D)
STAR Mapping
U6 RFAM
Abundance
(E)
Fraction of compactors
Minimum Hamming-Distance to an hg38 RNU6
U6-by-association
U6
(F)
Donor 2 Skin
Cell/compactor usage
RNU6-6p
Multi-mapped
Multi-mapped
Multi-mapped
Anchor
(G)
Compactors matching RNU6 loci
RNU6-8
Downstream alignment
Upstream alignment
RNU6-6P
A
C
G
T

Figure 3.

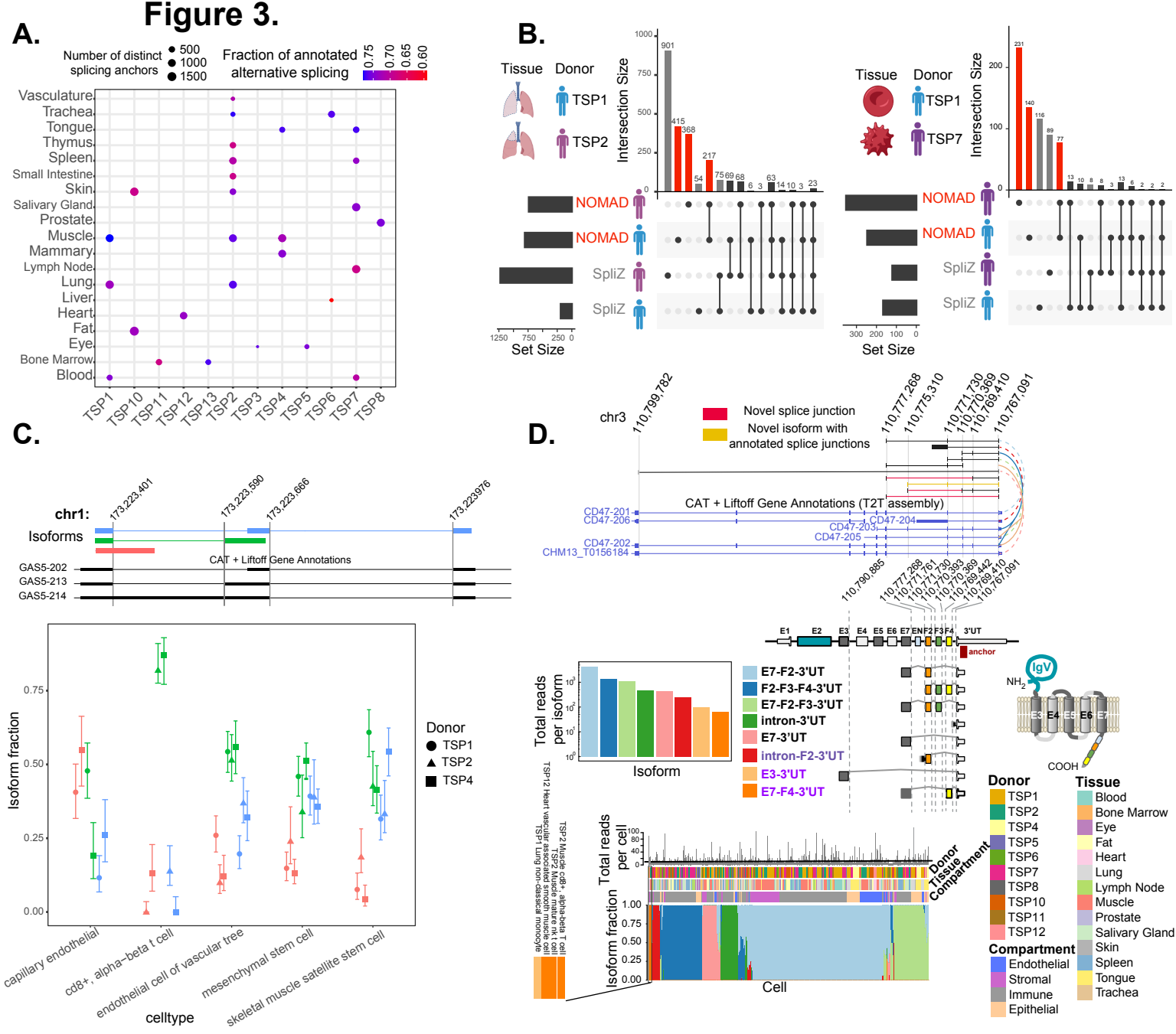

A.
Number of distinct splicing anchors
500
1000
1500
Fraction of annotated alternative splicing
0.75
0.70
0.65
0.60
Vasculature
Trachea
Tongue
Thymus
Spleen
Small Intestine
Skin
Salivary Gland
Prostate
Muscle
Mammary
Lymph Node
Lung
Liver
Heart
Fat
Eye
Bone Marrow
Blood
TSP1
TSP10
TSP11
TSP12
TSP13
TSP2
TSP3
TSP4
TSP5
TSP6
TSP7
TSP8
B.
Tissue
Donor
TSP1
TSP2
Intersection Size
NOMAD
NOMAD
SpliZ
SpliZ
Set Size
Tissue
Donor
TSP1
TSP7
Intersection Size
NOMAD
NOMAD
SpliZ
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
TSP1
TSP2
TSP4
capillary endothelial
cd8+, alpha−beta t cell
endothelial cell of vascular tree
mesenchymal stem cell
skeletal muscle satellite stem cell
celltype
D.
chr3
110,799,782
Novel splice junction
Novel isoform with annotated splice junctions
110,777,268
110,775,310
110,771,730
110,770,369
110,769,410
110,767,091
CAT + Liftoff Gene Annotations (T2T assembly)
CD47-201
CD47-206
CD47-204
CD47-203
CD47-205
CD47-202
CHM13_T0156184
110,790,885
110,777,268
110,771,761
110,771,730
110,770,393
110,770,369
110,769,442
110,769,410
110,767,091
E1
E2
E3
E4
E5
E6
E7
ENF2
F3
F4
3'UT
anchor
Total reads per isoform
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
E3
E4
E5
E6
E7
Total reads per cell
Donor
Tissue
Compartment
Isoform fraction
Cell
Donor
TSP1
TSP2
TSP4
TSP5
TSP6
TSP7
TSP8
TSP10
TSP11
TSP12
Compartment
Endothelial
Stromal
Immune
Epithelial
Tissue
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
Total reads per isoform
Isoform
Novel RPS24 isoform with only 3bp microexon included
Total reads per cell
Compartment
Celltype
Isoform fraction
Cell
epithelial cells expressing predominantly 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

Figure 4.

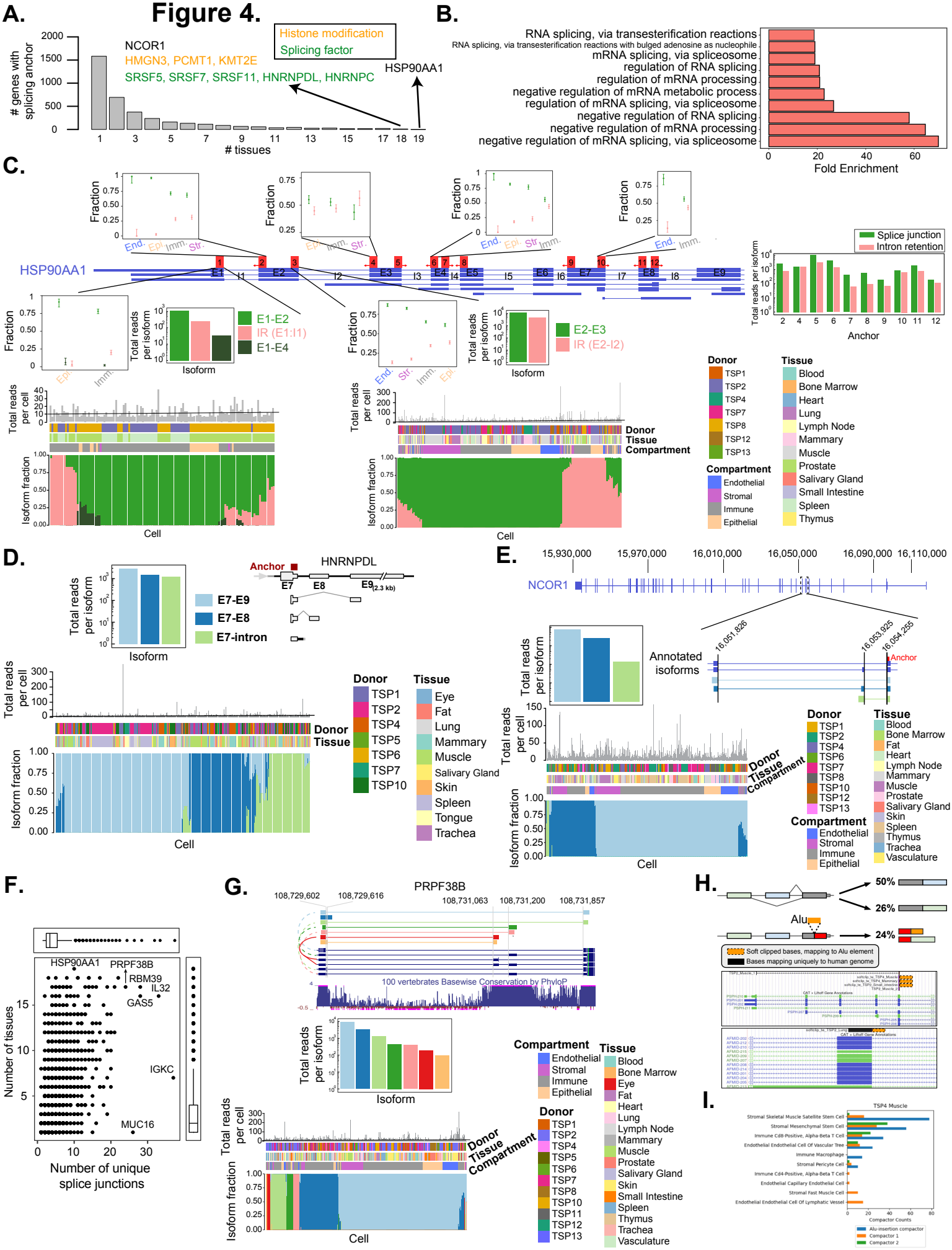
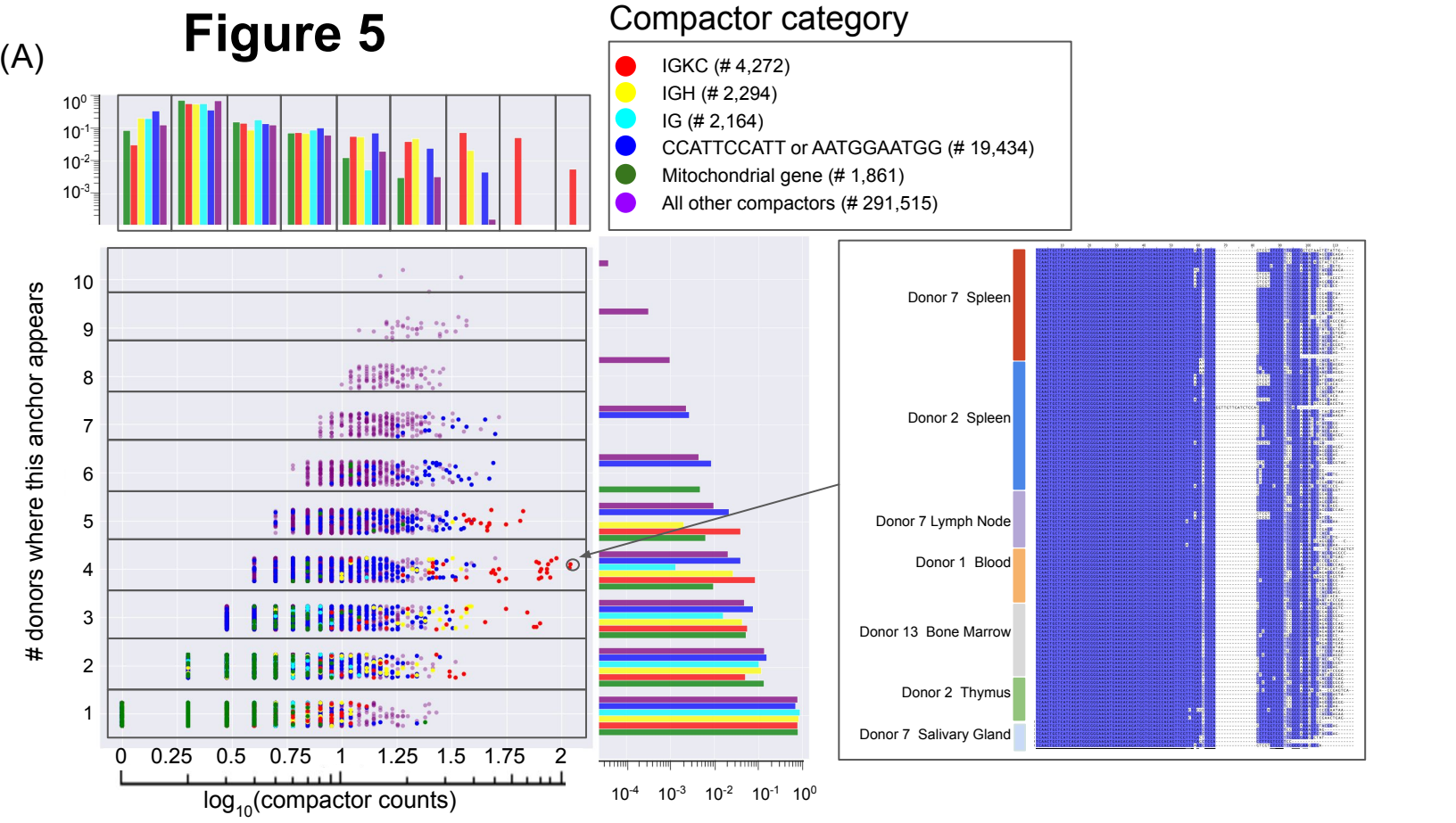
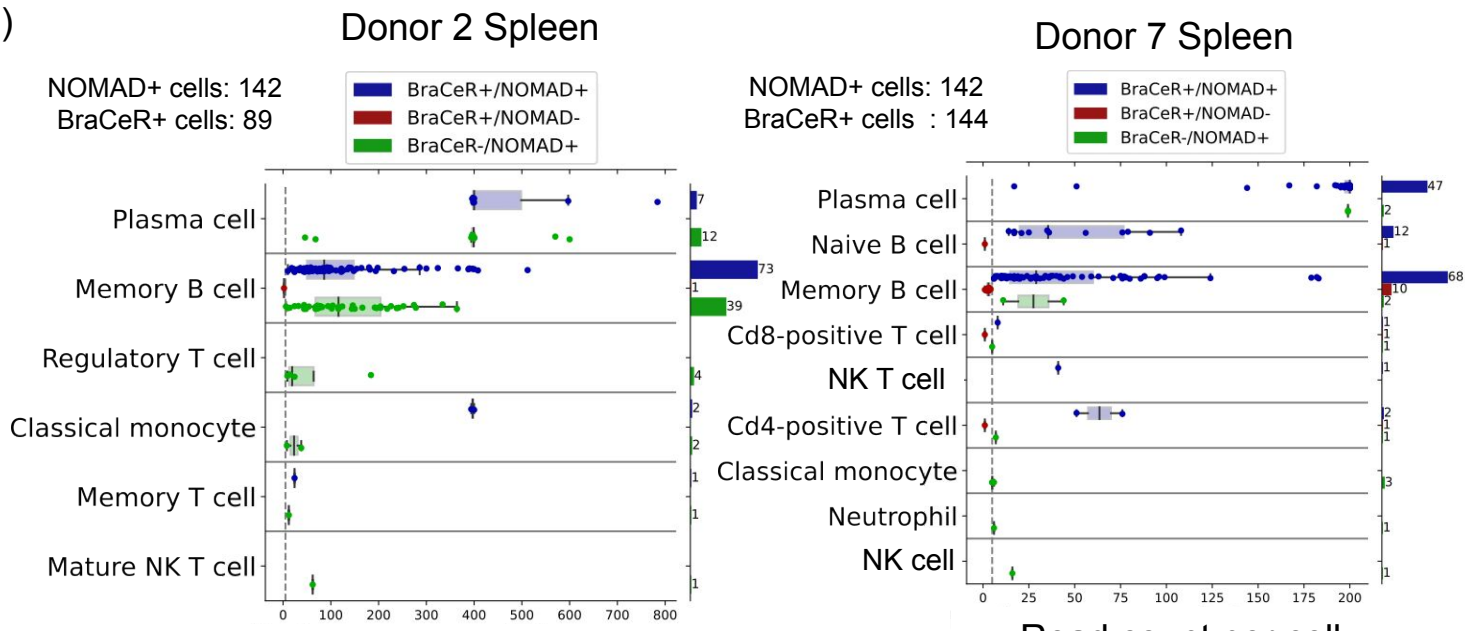

A.
genes with splicing anchor
tissues
NCOR1
HMGN3, PCMT1, KMT2E
SRSF5, SRSF7, SRSF11, HNRNPDL, HNRNPC
HSP90AA1
Histone modification
Splicing factor
B.
RNA splicing, via transesterification reactions
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
mRNA splicing, via 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
Fraction
Total reads per isoform
Isoform
E1-E2
IR (E1:I1)
E1-E4
E2-E3
IR (E2-I2)
Splice junction
Intron retention
Anchor
Total reads per cell
Isoform fraction
Cell
Donor
Tissue
Compartment
TSP1
TSP2
TSP4
TSP7
TSP8
TSP12
TSP13
Endothelial
Stromal
Immune
Epithelial
Blood
Bone Marrow
Heart
Lung
Lymph Node
Mammary
Muscle
Prostate
Salivary Gland
Small Intestine
Spleen
Thymus
D.
Anchor
HNRNPDL
E7
E8
E9 (2.3 kb)
E7-E9
E7-E8
E7-intron
TSP5
TSP6
TSP10
Eye
Fat
Skin
Tongue
Trachea
E.
NCOR1
15,930,000
15,970,000
16,010,000
16,050,000
16,090,000
16,110,000
16,051,826
16,053,925
16,054,255
Annotated isoforms
Vasculature
F.
HSP90AA1
PRPF38B
RBM39
IL32
GAS5
IGKC
MUC16
Number of tissues
Number of unique splice junctions
G.
PRPF38B
108,729,602
108,729,616
108,731,063
108,731,200
108,731,857
100 vertebrates Basewise Conservation by PhyloP
H.
50%
26%
24%
Alu
Soft clipped bases, mapping to Alu element
Bases mapping uniquely to human genome
I.
TSP4 Muscle
Stromal Skeletal Muscle Satellite Stem Cell
Stromal Mesenchymal Stem Cell
Immune Cd8-Positive, Alpha-Beta T Cell
Endothelial Endothelial Cell Of Vascular Tree
Immune Macrophage
Stromal Pericyte Cell
Immune Cd4-Positive, Alpha-Beta T Cell
Endothelial Capillary Endothelial Cell
Stromal Fast Muscle Cell
Endothelial Endothelial Cell Of Lymphatic Vessel
Compactor Counts
Alu-insertion compactor
Compactor 1
Compactor 2

Figure 5

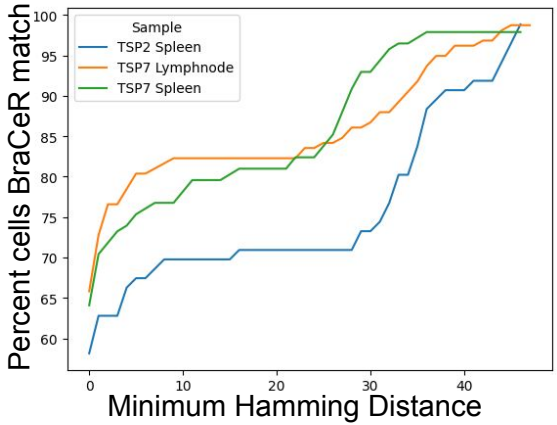
(A)



(B)



(C)



(D)

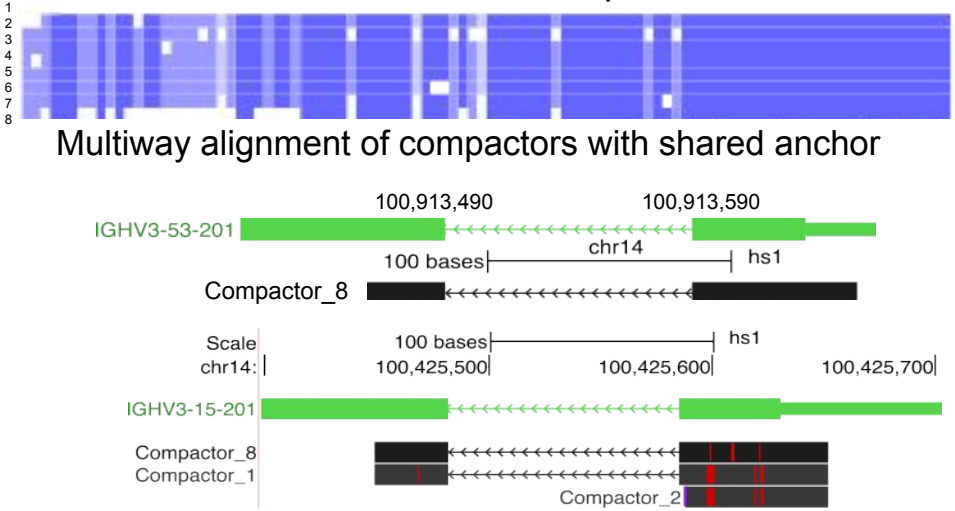
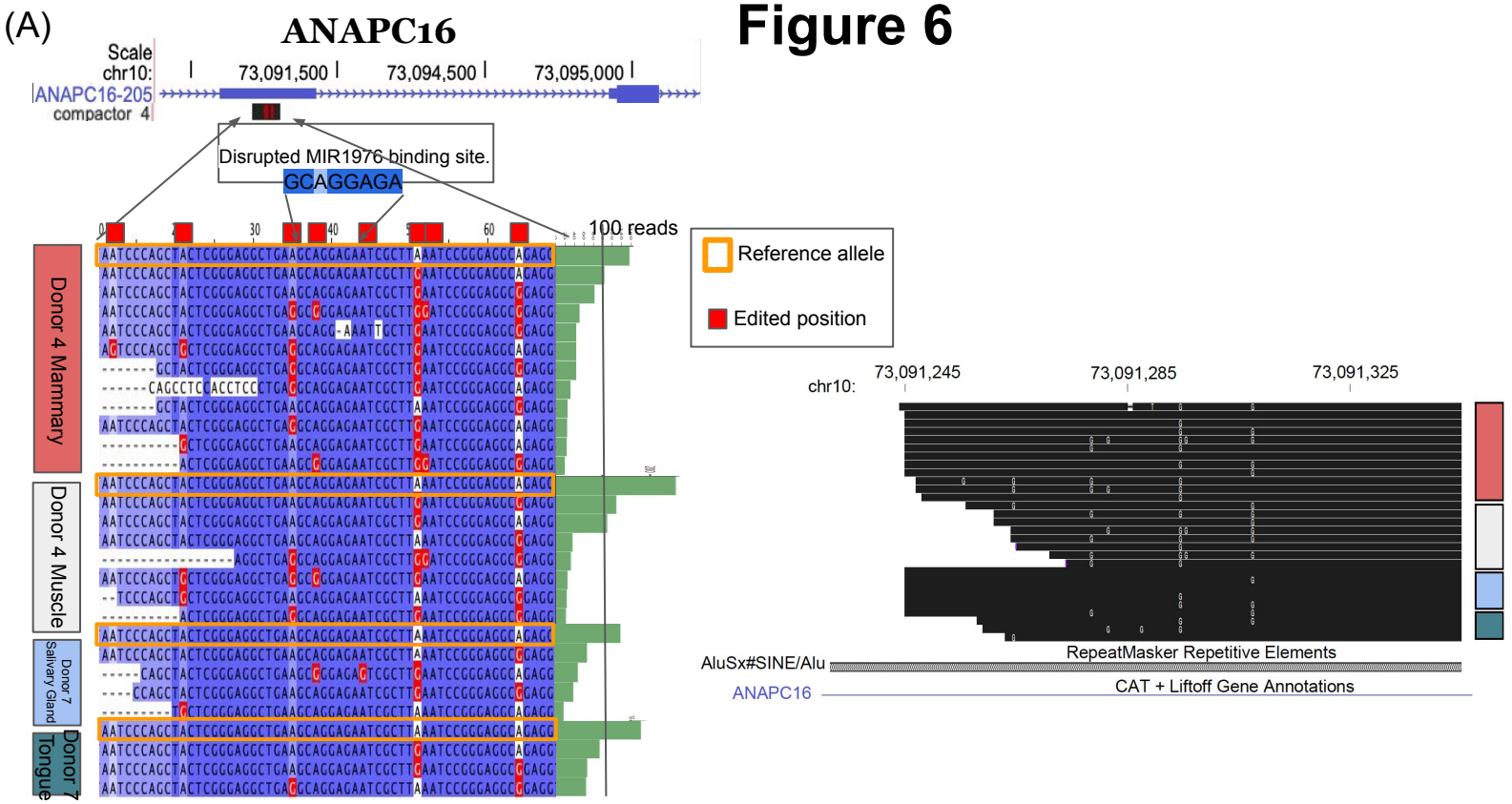
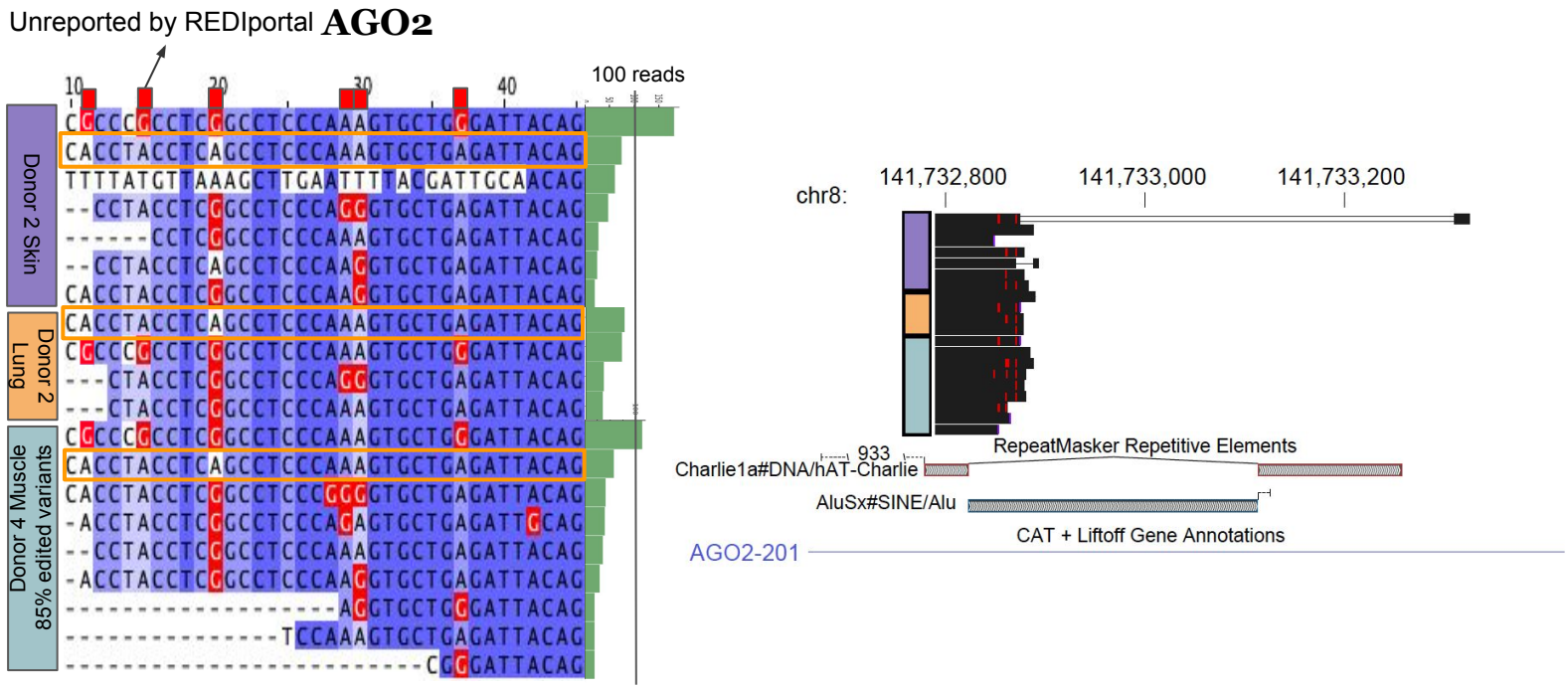


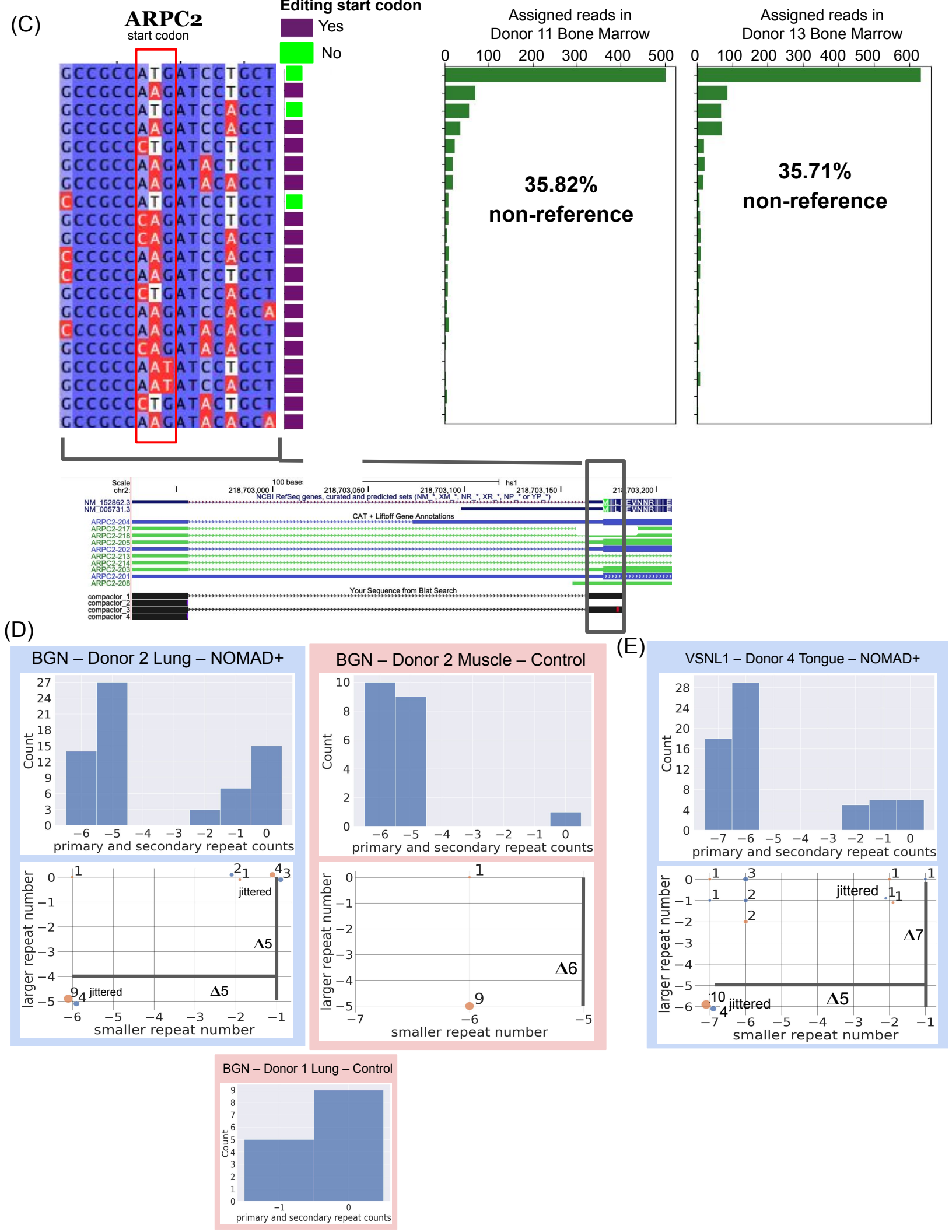
Figure 6

(A)


ANAPC16
Scale
chr10:
73,091,500
73,094,500
73,095,000
ANAPC16-205
compactor 4
Disrupted MIR1976 binding site.
GCAGGAGA
100 reads
Donor 4 Mammary
Donor 4 Muscle
Donor 7 Salivary Gland
Donor 7 Tongue
Reference allele
Edited position
chr10:
73,091,245
73,091,285
73,091,325
RepeatMasker Repetitive Elements
AluSx#SINE/Alu
ANAPC16
CAT + Liftoff Gene Annotations

(B)


Unreported by REDIportal
AGO2
100 reads
Donor 2 Skin
Donor 2 Lung
Donor 4 Muscle 85% edited variants
chr8:
141,732,800
141,733,000
141,733,200
RepeatMasker Repetitive Elements
933
Charlie1a#DNA/hAT-Charlie
AluSx#SINE/Alu
CAT + Liftoff Gene Annotations
AGO2-201


(C)
ARPC2
start codon
Editing start codon
Yes
No
Assigned reads in Donor 11 Bone Marrow
35.82% non-reference
Assigned reads in Donor 13 Bone Marrow
35.71% non-reference
CAT + Liftoff Gene Annotations
Your Sequence from Blat Search
(D)
BGN – Donor 2 Lung – NOMAD+
BGN – Donor 2 Muscle – Control
(E)
VSNL1 – Donor 4 Tongue – NOMAD+
BGN – Donor 1 Lung – Control
Count
primary and secondary repeat counts
larger repeat number
smaller repeat number
jittered
Δ5
Δ6
Δ7