ReadZS detects developmentally regulated RNA processing programs in single cell RNA-seq and defines subpopulations independent of gene expression

Elisabeth Meyer[1,2,*], Roozbeh Dehghannasiri[1,2,*], Kaitlin Chaung[1,2,*], Julia Salzman[2,1†]

Author affiliation
[1]Department of Biochemistry, Stanford University, Stanford, CA 94305
[2]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305

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

Abstract:

Post-transcriptional regulation of RNA processing (RNAP), including splicing and alternative polyadenylation (APA), controls eukaryotic gene function. Conservative estimates based on bulk tissue studies conclude that at least 50% of mammalian genes undergo APA. Single-cell RNA sequencing (scRNA-seq) could enable a near complete estimate of the extent, function, and regulation of these and other forms of RNA processing. Yet, statistical methods to detect regulated RNAP are limited in their detection power because they suffer from reliance on (a) incomplete annotations of 3' untranslated regions (3' UTRs), (b) peak calling heuristics, (c) analysis based on measurements collapsed over all cells in a cell type (pseudobulking), or (d) APA-specific detection. Here, we introduce ReadZS, a computationally-efficient, and annotation-free statistical approach to identify regulated RNAP, including but not limited to APA, in single cells. ReadZS rediscovers and substantially extends the scope of known cell type-specific RNAP in the human lung and during human spermatogenesis. The unique single-cell resolution and statistical properties of ReadZS enable discovery of new evolutionarily conserved, developmentally regulated RNAP and subpopulations of lung-resident macrophages, homogenous by gene expression alone.

Introduction

Differential RNA processing (RNAP) of the same "gene" can regulate gene function and controls RNA localization, stability, protein production, and translation efficiency (Di Giammartino et al., 2011; Floor and Doudna, 2016; Tushev et al., 2018; Wilusz et al., 2001). Despite this, most studies including of single-cell RNA-seq (scRNA-seq) reduce cell measurements to gene counts because differential RNAP, including kinetic rates of intron splicing, alternative polyadenylation sites (APA), and 3' untranslated region (3' UTR) use, are difficult to predict and quantify (Lusk et al., 2021; Mayr, 2019). While some poly(A) sites are annotated, a comprehensive annotation is still unavailable (Di Giammartino et al., 2011; Floor and Doudna, 2016; Mayr, 2019; Tushev et al., 2018; Wilusz et al., 2001).

Droplet-based single-cell sequencing (e.g., 10X Chromium) is designed to prime on poly(A) stretches of RNA, which are prevalent in introns and at the 3' end of most cytoplasmic RNAs (Zhang et al., 2019). Thus, in principle, 10X Chromium (10X) is an ideal technology to measure the rate of intron processing and inclusion of 3' poly(A) sites. However, technical limitations of scRNA-seq such as low capture efficiency and high dropout rates have led to the prevailing view that RNA is too sparsely sampled to measure alternative RNAP at single-cell level without imputation (Gao et al., 2021) or pseudobulking. Pseudobulking -- aggregating

reads from all cells within a cell type makes it impossible to measure heterogeneity within a pre-annotated cell type.

ReadZS is a new statistical approach that is annotation-free, is nonparametric and is a true single cell measure of differential RNAP that reveals biology missed by other methods. Published algorithms rely on peak-calling and referencing existing annotations (Shulman and Elkon, 2019; Ye et al., 2020); however, due to incompleteness of annotation, this reliance limits the power of these methods to discover novel regulation of RNAP and fully utilize single-cell resolved measurements. “Peaks” are seen in 10X data due to preferred priming at a single 3’ end, which produces a distribution of insert lengths that is approximately normal, due to tagmentation (Islam et al., 2014). Peak-calling-based methods assign reads to one of several peaks, corresponding to 3’ UTR sites, and then measure the enrichment of peaks in different cell types or conditions. However, if two “peaks” originating from priming from sites within 1 to 2 standard deviations of each other overlap (Huntsman Cancer Institute, 2021), peak callers may not distinguish them, limiting power to discover differential RNAP. Further, biochemical error processes can cause failures of a strict parametric modeling of peaks.

To our knowledge, all published methods inferring RNAP from scRNA-seq require peak-calling to identify regions of increased read density along the genome (Gao et al., 2021) and references therein). These methods either require annotation or analyze differential enrichment of at most two peaks within a 3’ UTR and have further restrictions on interpeak differences as well as manual curation of training data (Shulman and Elkon, 2019; Agarwal et al., 2021; Patrick et al., 2020). Current methods also perform pairwise tests of differential RNAP, losing statistical power by requiring (n choose 2) tests for n cell types (Gao et al., 2021). Connected to this idea, no published algorithm takes a statistical approach to subclustering cells or has been shown to reproducibly detect subclusters based on RNAP profiles within a single cell type. scDaPars (Gao et al., 2021) finds subclusters in a highly heterogeneous population of deeply-profiled cells via Smart-seq2 (SS2) and across a 96-hour time course of human endoderm development, but does not provide statistical quantification of cluster existence, and is limited to analysis based on reads from the 3’ UTR alone, a shortcoming we show has biological significance.

ReadZS enables annotation-free detection of RNA processing in scRNA-Seq

ReadZS is a computationally efficient, truly single-cell measure of RNAP. It overcomes biases and the reduced statistical power inherent in annotation-dependent and peak calling approaches. It can detect differential RNAP at single-cell resolution that are regulated in any number of cell types, identify subclusters of cells on the basis of RNAP alone, and find regulated RNAP as a function of developmental time. While ReadZS can identify differential RNAP that visually appear as “peaks”, it can also detect distributional shifts that are not (Fig. 1B). ReadZS is applicable to 10X and other 3’ capture scRNA-Seq methods as well as SS2, though we focus on 10X in this manuscript. While the ReadZS method does not rely on annotation, after significant windows are called by ReadZS, their positions are intersected with annotation files to allow assignment of regulated RNAP events to a 3’ UTR, gene body or unannotated region in order to enhance interpretability and downstream analysis.

The major innovations of ReadZS include: (1) no reliance on exon, isoform, or gene annotation: detection of differential RNAP including in 3’ UTRs and in introns; (2) a purely statistical approach to analyzing read distributions that bypasses “peak calling” and can detect read distributional shifts that are not defined by “peaks” with an ad hoc minimum interpeak distance, or limited to only cases with two peaks per gene (Shulman and Elkon, 2019); (3) a truly single-cell-resolved score that can be used to further elucidate function and single-cell biology of RNAP via integrating with other analytical scRNA-seq analysis such as continuous correlation analysis with single-cell resolved measures such as pseudotime or as input to subclustering algorithms; (4) a way to prioritize windows on the basis of effect sizes and

quantifiable FDR for each set of calls (5) a very efficient workflow: processing 12 BAM files, totalling 271 GB, required 2 hours 27 minutes of run time and 485GB total memory on a high-performance computing cluster. For convenience and reproducibility, the ReadZS pipeline has been implemented in Nextflow (Di Tommaso et al., 2017), a workflow management platform, with all needed packages and libraries pre-installed and can be accessed on github: https://github.com/salzmanlab/ReadZS.

ReadZS first partitions each chromosome into genomic windows and then summarizes the distributions of 10X reads across the genome by giving a lower score cells with reads closer to the downstream end of a window, and a higher score with reads closer to the upstream end. This is achieved by reducing each uniquely mapped read to a rank within a genomic window across a cell population, treating each strand separately (Methods, Fig. 1A) and ignoring metadata. Ranks within each window are normalized by subtracting their population mean and dividing by their standard deviation, defining a so-called read residual. The ReadZS value (or just "ReadZS") per cell per genomic window is defined by summing and scaling read residuals (Methods). Large negative (respectively, positive) ReadZS values mean that a cell's reads within a window are skewed upstream (resp. downstream) compared to the population average (Fig 1A). In this paper, we analyze 5kb windows, a length sufficient to capture variation in 3'UTR length, although this parameter is user-defined and flexible. A user can choose any genomic feature such as all annotated introns, genes, or UTRs to analyze with ReadZS.

The ReadZS values for a given genomic window follow a normal distribution centered at 0 under the null hypothesis that each cell has a statistically exchangeable (Durrett, 2019) read distribution per window (Methods). Moreover, ReadZS is scaled such that if two or more subpopulations of cells exist within a sample (Methods), the expected value of the ReadZS will converge to a value that is a function of the cell population, independent of sequencing depth.

The interpretable single-cell-resolved scalar value of the ReadZS means that its multivariate relationships with other covariates, such as pseudotime, can be evaluated without using cell type classification. Thus, ReadZS can detect regulated RNAP events that vary continuously with any measured covariate, such as space or time. When metadata is available, the cell-type-level distributions of ReadZS values can be used to test whether median ReadZS scores per cell type and window are exchangeable (Fig. 1A). After multiple hypothesis testing correction, the pipeline calls windows that are differentially processed across any number of cell types, with no need to pre-specify pairs of cell types to compare. This single-test to detect differences among n cell types increases power compared to pairwise differential testing as O(n^2) fewer tests are required. Further, the range in median ReadZS by cell type defines an "effect size" which can be used to systematically prioritize genomic windows for subsequent analysis.

We now present a technical study of ReadZS based on real scRNA-Seq data sets: (1) ReadZS provides highly significantly concordant calls of RNAP across different biological replicates ; (2) a post-facto peak calling on genomic windows with significant ReadZS variation between cell types shows the inferred the poly(A) priming sites are enriched with known 3' UTR ends, while discovering a significant number of sites that cannot be explained by priming from annotated 3' UTRs; (3) it rediscovers and extends known regulation of 3' UTR length in human and mouse spermatogenesis; and (4) its results are consistent with a recently published algorithm, Sierra (Patrick et al., 2020).

ReadZS calls are consistent across biological replicates

Benchmarking bioinformatics algorithms on real data is preferred to doing so in simulated data (Engström et al., 2013) due to multifactorial biochemical errors introduced during library preparations. Therefore, we performed a high throughput benchmarking of ReadZS by applying it to 10X data of non-tumor samples from three participants in the Human Lung Cell Atlas (HLCA), together encompassing 57 cell types in the lung and blood (Travaglini et al.,

2020). We chose this dataset because it is deeply curated and thought to define all existing subtypes of cells in the lung. Thus, it provides an opportunity to test if manual and computational cell type definition can be further refined by analysis of RNAP.

Consistent with (Travaglini et al., 2020), in this manuscript, we used participant 3 (P3), the most deeply sequenced individual, as the primary participant and P1 and P2 to validate our discoveries on P3. We ran the ReadZS pipeline from each participant separately. To identify if highly expressed genes had evidence of RNAP, we required >10 counts in 20 cells in at least two cell types required to calculate the ReadZS. ReadZS was calculable in 454 windows (across 419 genes, according to RefSeq Gencode annotations) in P3; 89 windows (19.6%, in 89 genes) were called as having significantly different RNAP between cell types (FDR < 0.05, Methods, supp. Table Y). Similar proportions of significant windows were found in the two other participants (19.2% and 18.9% of all ReadZS-calculable windows for P2 and P1 were significant, respectively). We focused our analysis on these calls, but note that lowering thresholds for calculating ReadZS to 5 counts in 10 cells, results in 2578 windows (across 2160 genes) and 274 windows (10.6%) called as significant likely due to decreased statistical power resulting from fewer reads and cells. In P1 and P2, respectively, at lower count thresholds, 87 (resp. 297) windows, 8.03%, (resp., 11.1%) were significant out of 1084 (resp. 1672) calculable.

Because both global and different sampling depths across cell types could impact concordance analysis for called windows, we restricted to windows with calculable ReadZS in both participants for each pairwise comparison. Restricting to the 245 windows with calculable ReadZS in P3 and P2, 27 were significant in both individuals (hypergeometric p-value = 2.08E-06). The P3-P1 and P2-P1 comparisons showed similarly high overlap (p<0.005; Supp. Table X).

We next measured the concordance in the directionality of ReadZS between the same cell types across individuals by evaluating consistency of ordered median ReadZS values, using a multivariate metric based on the Spearman footrule (Methods). The concordance of directionality of the ReadZS value per cell type was highly significant compared to chance ordering of cells (p < 0.005 for P1-P2, P2-P3, and p < 0.01 for P1-P3).

Statistically-identified read peaks in ReadZS-significant windows are enriched for known poly(A) sites and predict new APA sites

We further evaluated the biological relevance of ReadZS by calculating the fraction of ReadZS-significant windows that can be explained by poly(A) priming at annotated 3' UTRs. Given the 3' bias of 10X sequencing, ReadZS should detect differential RNAPs in 10X data that are enriched at annotated 3' UTRs. Indeed: 79 (88.8%) of 89 ReadZS-significant windows in P3 overlap with at least one 3' UTR annotation (Supp. Table 4).

To further assess the biological and statistical properties of significant windows, we performed a statistical post-processing step with Gaussian Mixture modeling (GMM) on significant windows to define regions of high read density in these windows (Fig 1A) . The GMM summarizes elevated read density within a window by modeling it as a mixture of Gaussians and the means of the components can be defined as the peak locations. Using these means, we computed summary statistics for the distances between the Gaussian means (intuitively, peak locations) and the closest downstream annotated 3' UTR.

We quantified the distance between the means of the fitted GMM and the nearest annotated downstream 3' UTR end, conditioning on this distance being <2kb because of known intronic priming. In the 10x data from both a timeline of human spermatogenesis (Hermann et al., 2018), a dataset we analyze and discuss below, and HLCA P3 data, the median distance from the GMM means to the nearest annotated 3' UTR is 286 bp, consistent with the average insert length of ~350bp in 10X libraries (Fig 1D; Huntsman Cancer Institute, 2021). This supports the idea that ReadZS-significant windows and the GMM approach to identify peaks primarily recover annotated 3' UTRs though no annotation was used in any step. In addition,

~25% of GMM-called means are farther than 580 bp from the nearest downstream 3' UTR end, supporting that restricting analysis to annotated 3' UTRs would limit discovery power (Fig1E). One of many examples illustrating this is a predicted novel 3' UTR used in CATIP in lung macrophages (Fig.1D). CATIP has the highest ReadZS effect size in P3 with relaxed filters, suggesting this 3' UTR has cell-type specific function.

ReadZS rediscovers and extends known regulation of RNA processing

To illustrate ReadZS discoveries in primary tissue, we examined the five highest cell type-specific effect sizes (defined as the range of medians of ReadZS across cell types within an individual) in HLCA P3 (Travaglini et al., 2020). The highest effect size reflects two 3' UTRs in overlapping genes within a single 5kb genomic window, a rare event in the human genome: *PTPRCAP*, a transmembrane phosphoprotein, and *CORO1B*, an actin-binding protein that controls cell motility (Fig 1D). For this genomic window, we illustrate differential ReadZS values using the only two cell types with sufficient reads (≥10 counts in ≥20 cells) in P3 to calculate median ReadZS values (see Methods). CD4+ memory/effector T cells dominantly express *PTPRCAP* whereas lung macrophages dominantly express *CORO1B*. This difference creates a dramatic shift in read distributions, demonstrating that ReadZS indeed detects genomic windows with large cell type-specific differences in read distribution.

Windows overlapping the genes *RPLP1*, *NEAT1*, and *SRSF7* were among the top 10 significant windows as ranked by effect size. In *RPLP1*, a component of the 60s subunit of the ribosome, significant intronic reads are cell type-specifically enriched in CD8+ memory/effector T cells relative to proliferating basal cells (Supp. Fig 1). *NEAT1*, a long noncoding RNA, is involved in nuclear paraspeckle assembly and undergoes extensive splicing and APA, but its isoforms have unknown functions (Dong et al., 2018). One peak detected by the GMM postprocessing of the ReadZS coincides with an annotated end and one not annotated-- this could reflect either unannotated APA or internal priming resulting from alternative splicing (Supp. Fig 1). A similar phenomenon is observed in *SRSF7*, a master splicing regulator implicated in tumor progression, with many splice isoforms (Königs et al., 2020). CD4+ T cells exhibit an unannotated GMM-called peak that could be evidence of unannotated alternative splice variants or unannotated APA (Supp. Fig 1). Because P3 and P2 have different distributions of cell types and read depths within those cell types, these windows did not have sufficient reads in the same cell types in P2 to compare read distributions.

In *CALM*1, two "peaks" called by the GMM each correspond to an annotated 3' UTR, which are differentially represented in proximal ciliated epithelial cells and macrophages (Fig. 1D). *CALM1* regulates calcium signaling and is known to undergo APA; in mouse, its long isoform is primarily expressed in neural tissue, and its 3' UTR has been shown to control localization and be functionally essential (Bae et al., 2020). ReadZS automatically extends this finding in both P3 and P2 to reveal significant regulation of 3' UTR length of *CALM1*, specifically that proximal ciliated cells have highest use of the longest 3' UTR in *CALM1*, consistent with the idea that the long isoform of *CALM1* is related to excitatory cell function (Bae et al., 2020). Differential RNAP in *KLF6*, a tumor suppressor regulating transcription (Narla et al., 2001) shows read coverage changes that suggest alternative regulation of the 3' end of *KLF6* in ciliated cells and macrophages compared to other cell types such as alveolar fibroblasts (both P2 and P3). According to the gene annotation of *KLF6*, these reads support use of unannotated 3' UTRs which are predicted to change the protein coding potential of *KLF6*, albeit at lower frequency than the dominant priming site. Because these variants modify the 3' UTR, they have unknown impacts on translation and thus protein abundance. Together, these examples illustrate the unique power of ReadZS to identify regulation including those that do not correspond to annotated 3' UTR sites.

ReadZS has complementary power compared to other algorithms

To the best of our knowledge, no published method is comparable to ReadZS, which can predict novel APA sites and detect alternative intronic processing using only 10X data and is completely agnostic to annotation. We still view it as important to illustrate how ReadZS compares to other methods, including instances of previously unreported regulated RNAP detected by ReadZS . We compared ReadZS to a state-of-the-art method, Sierra (Patrick et al., 2020) that uses pseudobulk analysis to detect differential transcript usage (DTU) including from single-cell data fibroblasts in injured and uninjured mouse hearts (Farbehi et al., 2019). Sierra was used to measure 3' UTR length changes between actively cycling fibroblasts (F-Cyc, F-Act, or F-CI) and resting fibroblasts (F-SL and F-SH) and found 631 genes exhibiting DTU (though with unknown FDR).

We performed ReadZS analysis and found 353 significant windows, across 299 genes (according to RefSeq annotations; FDR < 0.05; Methods). Surprisingly, over 90% of these genes were not called or reported by Sierra. Restricting to 631 genes with DTU reported by Sierra, 148 had sufficient per-cell read coverage to calculate the single-cell-resolved ReadZS and 26 (17%) of these genes were called by ReadZS. Out of the 7 genes the authors the authors investigated via RT-qPCR, only windows intersecting Cd47 and Col1a2 had sufficient read depth (≥ 5 counts in ≥ 10 cells) in at least two fibroblast populations to calculate median ReadZS values, and both were called significant by ReadZS. Despite the limitation of shallow read depth, ReadZS discovers many cases of regulated 3' UTR changes missed by Sierra. Examples of new discoveries by ReadZS include *Rpl13a*, missed by Sierra despite having two cleanly separated APA peaks, and *Rab2a,* where multiple APA sites in the final exon create overlapping peaks which we expect to be missed by peak-calling-based methods. This analysis illustrates that ReadZS is a complementary approach that recovers genes found by other algorithms and reveals biology they miss. We attempted to run ReadZS on three lung adenocarcinoma cell lines used to benchmark scDAPars (Gao et al., 2021), but we could not as the metadata only contained cell barcodes with counts, and no cluster identities were given.

We now go on to focus on novel biology and the increased power of the ReadZS method. Because the ReadZS score is calculated at the single-cell level, it has lower power to detect differential RNAP at the level of pseudobulk. However, ReadZS has power against alternatives where Sierra or similar methods lack it: As discussed, what ReadZS can detect is not restricted to the 3' UTR. It has further power as follows: 1) Multiple APA sites within the same exon of a gene can create overlapping peaks in the read coverage, which cannot be quantified by a peak-calling method (e.g., Fig. 1B) 2) because ReadZS is annotation-free, peak-free and a true single-cell measure of differential RNAP, it can discover subpopulations of cells in fine-grained, manually curated and populations of cells profiled with 10X homogenous by gene expression 3) it can automatically discover RNAP regulated monotonically as a function of pseudotime. Importantly, discoveries in 2) and 3) include shifts in read distributions detected by ReadZS that would be missed by all published methods (Gao et al., 2021; Patrick et al., 2020; Wu et al., 2021) for APA detection based on peak-calling.

The ReadZS detects RNAP-defined subpopulations within highly curated cell types

ReadZS is a single-cell-resolved statistical measure of RNAP, enabling it to reveal subpopulations of cells with distinct RNAP regulation in seemingly highly homogenous cell populations defined by gene expression. To our knowledge, as discussed in the introduction, no existing algorithm is capable of identifying such regulation, which potentially includes differential RNAP outside of the 3' UTR, in a homogenous cell population. Similarly, no such approach provides a statistical assessment or measure of falsely discovered subclusters.

We now describe a statistical approach that uses ReadZS to identify reproducible subpopulations of lung-resident macrophages with the strongest signal in six distinct genes. First, post-facto analysis shows these subpopulations cannot be explained by numerical artifacts of low sampling depth. Second, cluster identity across the six genes is highly

dependent, suggesting a biological origin. Although Seurat (Stuart et al., 2019) does not have calculable false discovery or false negative estimates, it also identifies the six ReadZS-identified genes as markers of the subpopulations.

The ReadZS has a closed-form null distribution that is approximately normal with mean 0 under a null hypothesis that reads are sampled from a homogenous population of cells; when two subpopulations exist, if one is a minority, theory implies that ReadZS will be a mixture distribution of two Gaussians, one with mean near 0 and one with a non-zero mean. Using this as a starting point, we performed automatic, model-based clustering of ReadZS for each pair of window and cell type separately to test if subpopulations within the highly curated HLCA cell types may exist. We fit a Gaussian mixture model (GMM) with an integrated complete-data likelihood (ICL) criterion to identify the number of components and parameters that define the mixing components (Methods). ICL (Biernacki et al., 2000) is a function of the likelihood and penalizes both the number of parameters and the mean entropy of mixing components. It is used to decrease the chance of overfitting and favors more separated mixing components.

We focused subpopulation analysis on macrophages because they were the deepest sequenced cell population and because they are difficult to profile with SS2 technology (Travaglini et al., 2020). Six windows (out of ~9K) overlapping the genes *BCL2A1*, a cell death regulator, and *RPL35*, *COX6C*, *RPL26*, *ATP51E* and *GNG5* showed statistically significant evidence of cell type-specific subclusters within macrophages in both P3 and P2 (Fig. 2, Supp. Figure 3). Subclusters in these windows (both P2 and P3) were in the top 15% of windows evidencing mixed ReadZS populations, measured by Bhattacharya distance, a distance between the two mixing distributions. All windows show the same read distributions in both individuals and share a signature of increased read density in one subpopulation of macrophages in the ultimate intron of the gene (Fig. 2A-D). This is consistent with slower intron kinetics or random, premature APA in these genes. Thus, the fact that ReadZS measures RNAP of intronic reads enables it to detect these subpopulations: their definition hinges on the read density detected in the last intron of the genes rather than distributional shifts in the 3' UTR region.

We further tested whether subpopulations could have been spurious artifacts derived from sampling a homogenous population of cells at low depth. To evaluate this, we resampled reads from the observed genes and randomly assigned reads to cells conditional on observed counts and computed the number of cells per cluster (Methods). In simulations, far fewer cells were observed in the minor subcluster than in the real data for all windows in both individuals (p-value < 10^-5).

Correlation between subcluster assignments for each window further supports that ReadZS identifies true subpopulations within gene-expression-homogenous macrophages. Subcluster assignments are correlated (p-value < 10^-5, Fig. 2E,F, Supp. Figure 5). Finally, we ran Seurat (Stuart et al., 2019) on ReadZS within macrophages (Supp. Figure 4). Seurat blindly identified these six genes as marker genes for a subpopulation of cells from both individuals, though also identified 5 other subclusters of macrophages which are likely false positives and lack statistical quantification (Supp. Fig. 4). In summary, ReadZS identifies a reproducible, statistically quantified biological signal in a homogenous population of macrophages, possibly ones that have higher rates of intron retention, or have slower splicing kinetics in their ultimate intron, that could not be identified with gene expression profiling (Travaglini et al., 2020).

Single-cell resolution of ReadZS reveals evolutionarily conserved, developmental post-transcriptional regulation

Global 3' UTR shortening during spermatogenesis is a conserved but incompletely understood post-transcriptional regulatory program (Bao et al., 2016; Li et al., 2016). We tested if ReadZS could detect global changes in 3' UTR length in mouse and human spermatogenesis (Hermann et al., 2018) without human-guided annotation. For each genomic window, we

calculated the correlation between estimated pseudotime and ReadZS value (Fig. 3A). In human, restricting to the 562 windows overlapping annotated 3' UTRs (Methods), 93 windows had significant correlation to pseudotime (|Spearman's rho| > 0.3, corrected p-value < 0.05). 14 out of 93 (15%) windows were positively correlated, consistent with 3' UTR lengthening, and 79 (85%) were negative-- consistent with global shortening. In mouse, restricting to the 310 windows overlapping annotated 3' UTRs, 3 (1%) significant windows had signs consistent with 3' UTR lengthening and 307 (99%) had signs consistent with shortening. This finding is consistent with work showing that 3' UTRs globally shorten during mouse spermatogenesis (Bao et al., 2016; Li et al., 2016; Shulman and Elkon, 2019) ; we are not aware of studies that have reported this phenomenon, or the genes we identify as regulated during spermatogenesis in humans.

To test if there is evolutionary conservation of mammalian genes undergoing regulated changes in 3' UTR length, we matched and found 374 genomic windows (314 shared gene names) annotated with the same gene name using RefSeq annotations (Methods). 56 of 374 (15%) of window pairs were significantly correlated with pseudotime in both human and mouse, significantly more overlap than expected by chance (hypergeometric p-value = 0.006). 42 out of 56 pairs had the same sign, corrected for gene direction (hypergeometric p-value = 1.27E-6). For example, ZFAND6, a zinc finger protein implicated in the pathophysiology of diabetes but not studied in spermatogenesis (Ndiaye et al., 2017), shows high conservation across vertebrates in the 3' UTR. Indeed both human and mouse exhibit similar patterns of 3' UTR shortening in spermatogenesis (Fig. 3B). Mouse read distributions further support an un-annotated 3' UTR. The significant overlap in magnitude and direction of significant correlations between human and mouse supports the finding that ReadZS detects unreported, evolutionarily conserved regulation of RNAP during spermatogenesis.

ARPP19, a gene known to be a mitotic regulator but with unreported 3' APA regulation (Virshup and Kaldis, 2010), has the largest negative correlation with pseudotime in human, reflecting 3' UTR shortening. The second highest magnitude correlation is in *S100A10*, a gene studied in the immune system but with unknown function in sperm (Miles and Parmer, 2010). ReadZS detects a shift in RNAP over time, but peaks in the detected window are overlapping and thus would likely be missed by a peak caller (Fig 3C). ReadZS also discovers 3' UTR lengthening of *OAZ1*, a gene implicated in ovarian function but with un-reported regulation in sperm (Fig. 3C) (Kang et al., 2017). Other examples include windows with overlapping or multiple peaks, e.g. windows covering the 3' UTRs of TSSK1B and SLC25A37 (Supp. Fig. 6).

Manual curation of spermatogenic transitions can categorize sperm into developmental categories: spermatocytes, spermatids, and mature sperm. These stages have been pseudobulked to enable differential APA analysis (Wu et al., 2021). However, because the ReadZS value is computed at a single-cell level, it potentiates discovery of fine-scale developmental transitions such as pseudo-temporal trends within immature sperm. To illustrate this capability, we used ReadZS to study differential RNAP within narrow windows of pseudotime. We correlated the ReadZS values to pseudotime, restricted to the earliest 25% of time in human (Fig. 3C). Out of 1433 windows with calculable correlation in that pseudotime interval, 38 windows had significant correlation to pseudotime (|Spearman's rho| > 0.3, corrected p-value < 0.05). 32 windows had significant correlation within the first 25% of pseudotime but not over the entire range of time. These windows include *MED21*, a component of the mediator complex involved in transcriptional regulation which shows general lengthening, and potentially, equalizing use of longer 3' UTRs over early spermatogenesis. Like *OAZ1*, *MED21* contains overlapping peaks and peaks at unannotated 3' UTR sites, which may hinder a peak-calling algorithm (Fig 3C). Such discoveries highlight the power of a single-cell measure of RNAP that can discern RNAP within a "cell type," including events early in spermatogenesis.

Conclusion

In summary, ReadZS is a novel, transparent, robust, and annotation-free statistical algorithm to detect regulated RNAP in high throughput single-cell RNA sequencing data. Applying it to primary cells reveals new biology of RNAP including in regions outside and within the 3' UTR regulation missed by peak calling algorithms. We anticipate that further analysis of the ReadZS will facilitate further functional inference for regulated RNAP in single cells, including 3' UTR use.

The computational pipeline is efficient and lightweight and can be easily integrated into any existing single-cell pipeline. The window size used in this manuscript, and the use of poly(A) primed data are not necessary for the methodology developed here. For example, windows could be chosen adaptively or based on a subset of features of interest. In addition, 5' capture technology, SS2 data or even scATAC-seq (Buenrostro et al., 2015) could all be used as inputs to ReadZS because the algorithm operates on read distributions that need not be peaks. For example, we would expect the ReadZS could detect differential 5' UTR use, intron retention, or exon inclusion when windows include these features and neighboring features that are not differentially processed (e.g., a constitutive exon or UTR). We anticipate that ReadZS should also be a powerful analytic tool for data such as generated by single nucleus sequencing data and or derived from split-pool tagging (Agarwal et al., 2021). A final area of future work will be to test whether more cell types and states can be defined when the ReadZS value itself or in conjunction with gene expression are used to perform clustering analysis or trajectory inference.

Acknowledgments: We thank the Salzman lab for useful discussion, especially Julia Oliveri and Robert Bierman for comments that enhanced the clarity of the document. We thank Sarthak Satpathy for early prototypes of processed bam files that were used in early versions of ReadZS.

Funding: E. M. is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1656518 and a Stanford Graduate Fellowship. R.D. is supported by the Cancer Systems Biology Scholars Program Grant R25 CA180993 and the Clinical Data Science Fellowship Grant T15 LM7033-36. J.S. is supported by the National Institute of General Medical Sciences Grants R01 GM116847 and R35 GM139517 and NSF Faculty Early Career Development Program Award MCB1552196.

Methods:

Creating counts tables from 10X BAM

The read Z-score (ReadZS) summarizes the transcription state of a genomic window in a single cell. It is calculated using only reads that fully align to the genome with no gapping, so it excludes spliced reads. 10 Reads were aligned using STAR (v 2.7.5.a) (Dobin et al., 2013) with default parameters except for chimSegmentMin = 10 and chimJunctionOverhangMin = 10. UMI demultiplexing and cellular barcode identification and correction for 10X data was performed using UMI-tools (Smith et al., 2017). BAM files are opened with Samtools and reads were filtered based on the CIGAR string "<length(SEQ)>M" and MAPQ score 255 to only allow uniquely-mapping exact and full-length matches. The reads were then split by chromosome and strand. The reads were then deduplicated, by removing cells with any duplicated UMIs or UMIs aligning to more than one unique position. The reads were then collapsed using the identifier column and counted at each position.

Each chromosome is split up into equal-sized windows with size inputted by the user - 5kb windows were used in every analysis unless stated otherwise. Each read is assigned a stranded window (window position and genomic strand) based on the read's position and strand. The

tables of reads and counts are then separated by chromosome. If there are multiple samples or files within the experiment, the counts tables from the same chromosome from different samples are concatenated together, e.g., all reads from chromosome 1 from any file are in the same counts file. For HLCA data, the data was divided by participant for ReadZS calculation in order to avoid any batch effects between individuals. For the Sierra, human spermatogenesis, and mouse spermatogenesis datasets, all data was analyzed together.

ReadZS Z-score calculation

To calculate the ReadZS value for a window in a particular cell *i*, the genomic positions of the window across all cells are assigned an increasing rank, with the most upstream position assigned rank 1. Only positions appearing in the data (i.e., positions with at least one read mapped there) are assigned a rank value and there is no gap in ranks between one position and the next, even if the positions are far apart.

For each genomic window, we first find the mean rank for this window across all cells, μ. For each position along the window with rank *r,* the number of aligned genomic reads that overlap with this position, m_j, is counted. The sum of the weighted ranks is divided by the total number of reads within that window across all cells, *N*:

$$\mu = \frac{\sum_i \sum_r r \; m_{ir}}{N}$$

Then the standard deviation of the weighted rank for the window, σ, is given by:

$$\sigma = \sqrt{\sum_i \sum_r (r - \mu)^2 \frac{m_{ir}}{N}}$$

For each window, the weighted ranks in individual cells are found by multiplying each rank r by the number of reads at that rank in cell i, m_ir.

Then, these weighted ranks are renormalized by subtracting the mean weighted rank, μr, and dividing by the standard deviation of the ranks, σ. To further normalize for read depth, we divide by *N_i*, the total number of reads from the window for cell *i*. This gives us the Z-score for the window in cell *i*:

$$z_i = \frac{\sum_r (m_{ir} r - \mu r)}{\sigma N_i}$$

Then, the expectation of z_i will be zero and the variance will be 1/N_i.

Pseudocode/steps for calculating ReadZS:

- Within each window, assign ranks to positions
- Count number of reads per position
- Count total number of reads per window
- Calculate weighted mean rank for each window
- Calculate standard deviation of mean rank for each window
- Calculate total number of reads for each cell
- Calculate scaled z-score (ReadZS) for each window & cell = (sum of weighted ranks - mean weighted rank) / (standard deviation * total reads per cell)

- Remove rows with infinite ReadZS

Identification of windows with regulated RNA processing: median ReadZS and its p-value for each window/cell type pair

When cell type metadata is available, cells can be assigned a cell-specific annotation (e.g., lung macrophage). For each window, a median ReadZS is then calculated within each cell type. To include a particular combination of a window and cell type, we required a minimum of 20 cells with at least 10 counts in that window-cell type combination. For the Sierra data, we reduced these minimum requirements to 10 cells with at least 5 counts to account for the lower read depth. To systematically prioritize windows for further follow on studies, they can be ranked according to the range of median ReadZS values across all cell types. To find which genes and 3' UTRs intersect these windows, we intersected the window positions with annotation bed files of genes and 3' UTRs requiring an overlap of at least 25% to annotate a window with that gene or 3' UTR (Supplementary Methods).

To evaluate whether a window has median ReadZS values that are more extreme than expected by chance, i.e., represent some biological change in RNA processing, we adopted an approach from (Chung and Romano, 2013). For each window, we have *I* cell types, with cell type *i* having n_i cells, a median $theta_{n,i}$, and a sample variance $sigma^2_{n,i}$. The null hypothesis is that there is no difference in the median ReadZS between cell types. We can compute the following test on the medians for one window:

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

This test statistic can also be computed multiple times on versions of the data where the cells' cell type labels have been permuted to create a permutation distribution for one window. However, according to (Chung and Romano, 2013), this permutation distribution converges to the chi-squared distribution with *k-1* degrees of freedom; therefore, instead of starting with permutations, we first calculate the p-value for each window by comparing it to the chi-squared distribution. This saves the compute time by allowing us to quickly filter out most windows that are not significant. Then, only if the window has p_chi^2 < 0.05, a permutation distribution is computed by permuting cell type labels. We have used 100 permutations. Then, the cumulative distribution function of the permutation distribution is given by:

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

So to quantify whether the median is extreme in either direction, we can calculate the p-value as

$$p_{\mathrm{perm}} = 2\min(\mathrm{cdf}_{\mathrm{perm}}, 1 - \mathrm{cdf}_{\mathrm{perm}})$$

The p-values calculated from permutations are then Benjamini-Hochberg corrected to yield the final list of significant windows.

Identification of windows with regulated RNA processing: correlation between ReadZS and pseudotime

In the human and mouse spermatogenesis datasets, we calculated the correlation

between ReadZS and pseudotime for all windows with at least five reads from that window in at least 300 cells. We called a window as significant if it had |Spearman's rho| > 0.3 and Bonferroni-corrected p-value < 0.05.

3' UTR shortening in human and mouse spermatogenesis

To examine changes in the 3' UTR length, we intersected all the significantly correlated genomic windows in the human and mouse spermatogenesis datasets with RefSeq 3'UTR annotations obtained from the UCSC Table Browser, requiring an overlap of at least 25% to annotate a window with that 3' UTR. We then determined the "sign-corrected correlation value" by multiplying the Spearman's correlation coefficient by -1 if the genomic window was on the minus strand. That way, a window with a negative correlation to pseudotime always indicates a skew towards more upstream reads for that gene, i.e., 3' UTR shortening if the window covers a 3' UTR region. We first considered genomic windows with any 3' UTR annotations: in human, we found 79 windows (85%) out of 93 with negative sign-corrected correlations, consistent with 3' UTR shortening; in mouse, we found 307 (99%) out of 310 windows with negative sign-corrected correlations. Since a genomic window may contain several exons or even several genes, we also considered genomic windows with 3' UTR annotations but without any 5' UTR or exon annotations. Among those windows, we found 6 (86%) out of 7 in human and 66 (96%) out of 69 in mouse had negative sign-corrected correlations, consistent with overall 3' UTR shortening.

Overlap between human and mouse windows with significant correlation

To test whether there were genes exhibiting similar changes in RNA processing over spermatogenesis in both mouse and human, we selected the windows from both data sets with calculable correlation to pseudotime (requiring a minimum of 5 counts per window per cell in at least 300 cells), and intersected the two data sets to find windows with matching RefSeq gene names in mouse and human. Among window pairs with the same gene name in human and mouse (374 in total), we found 56 window pairs where both windows were significantly correlated with pseudotime, and 40 of those had negative correlation values for both windows (after correcting for gene direction). We used a hypergeometric test to calculate whether these overlaps of significance and correlation sign were more extreme than expected by chance, using the R function phyper.

Concordance of ReadZS between pairs of data sets

In order to assess whether the windows called as significant by ReadZS pipeline actually have consistent cell type-specific regulation of RNA processing, we created a test statistic that measures concordance in ReadZS values between data sets. For a genomic window called as significant, a cell type "A" with a higher median ReadZS value will have reads skewed upstream relative to a cell type "B" with a lower median ReadZS. If these differences in read distribution reflect real differences in biology between these cell types, we expect cell type A to consistently have a higher median ReadZS than cell type B in different biological replicates. For each genomic window called as significant, we expect the cell types to follow the same ranking as determined by their median ReadZS values.

Accordingly, we created the following test statistic to measure concordance between two data sets:

$$x = \sum_{j=1}^{N} \frac{1}{m_j} \sum_{i=1}^{m_j} \left| R_{ij} - R_{ij}' \right|,$$

where R_ij is the rank out of m_j cell types of the ith cell type in genomic window j, R_ij' is the rank for the same cell type and window but in a second data set, and N is the total number of genomic windows. If most windows have similar rankings of cell types in the two data sets, the

differences in ranks between the data sets will tend to be small, resulting in a smaller value for x. We simulated a null distribution for x for each pair of data sets by calculating x 5000 times using permuted ranks: for each simulated calculation, we first randomly permuted the ranks of cell types in the second data set, and then we used the intact first data set and the permuted second data set to compute x. For each pairwise comparison of data sets, we were then able to calculate a p-value by comparing the real value of x against the simulated null distribution.

Subpopulation analysis using ReadZS

Under the null hypothesis, the distribution of the ReadZS values of a gene for the cells within a cell type follow a normal distribution with mean 0. However, if there are subpopulations of cells within a cell type with distinct RNA processing profiles, the ReadZS distribution would be multimodal and could be better modeled by a Gaussian mixture model (GMM). We obtain the optimal number of components in the GMM, which corresponds to the number of subpopulations with distinct ReadZS profiles, as the knee point in the integrated complete-data likelihood (ICL) curve across different numbers of components. We apply the ICL criterion to the ReadZS distribution of each pair of gene and cell type and for each pair where ICL finds at least two components, we assign cells to one of the subclusters via fitting a Gaussian mixture model. We further check for distinct subclusters by computing the Bhattacharya distance between the components. If the distance is <0.5, we reduce the number of components by one and again fit a GMM. We stop if there is only one component remaining or the distance between components is at least 0.5.

We tested whether the identified subclusters are driven by real biological signals or could arise as artifacts of low sampling depth in scRNA-Seq. To do so, we defined cluster 1 as the major cell cluster with ReadZS mean nearest to 0 and then simulated the probability of observing at least as many cells in cluster 2 as follows:

We fitted a Gaussian mixture model with 2 components to the overall read distribution within the window across the cell type and defined any read upstream of the cutpoint of the cluster assignment as an “outlier” read. In our simulations, we assigned any cell with at least one “outlier” read as belonging to cluster 2. This classification rule is essentially consistent with the way that cells have been assigned to subclusters based on their ReadZS values as after examining the read distributions for the identified subclusters in macrophages for each marker gene and in both P2 and P3 participants, we noticed that only cells in subcluster 2 contain more upstream reads.

We then permuted reads and assigned them to the cells conditioned on their read counts and iterated the permutation for 10^5 times. We computed the empirical p-value for each subcluster corresponding to a marker as the fraction of simulations with at least K cells in the first subcluster. Our simulations showed extremely significant p-values (< 10^-5) for all 6 genes in both P3 and P2.

Peak calling using Gaussian mixture model

For each significant window called by ReadZS, we performed peak calling by fitting a Gaussian mixture model to the distribution of the reads from the entire dataset across that window. The identical methodology to the subclustering analysis was performed but on read positions rather than ReadZS.

Pipeline implementation using Nextflow

To allow for reproducible and parallelizable results, the ReadZS pipeline is written in Nextflow. Nextflow (Di Tommaso et al., 2017) is an open-source workflow management system that integrates command-line and scripting tools to analyse large-scale datasets. The ReadZS workflow takes in BAM alignment files from 10X or SS2, and it performs processing and calculation steps on all of the files in parallel. The workflow then outputs tables with cell type

medians and their associated p-values. The workflow also allows users to input dataset-specific parameters, such as cell annotation files, genome window files, and the columns used to define ontology (cell type or other grouping). To further enhance portability, the entire workflow can be run on a high-performance computing platform or on a cloud computing platform.

Calculating Distance to 3' UTR Annotations

For human samples, the Gencode GFF3 files were used for distance calculations and plotting. For mouse samples, RefSeq GFF3 files were used. To extract 3'UTR regions in bed format, the GFF3 file was filtered for `feature type = 'three_prime_UTR'`, with the `ID` field used as the 3'UTR identifier. To extract gene regions in bed format, the GFF3 file was filtered for `feature type = 'gene'`, with the `gene_name` field used as the gene identifier.

To determine the `num_3UTR_300bp_downstream` column, a bed file was created for each window's start and end positions, shifted 300bp downstream relative to the strand of the window. To determine 3' UTR ends, the 3' UTR bed file was filtered for regions on the same strand as the window, and the start position or end position was used to create a separate bed file for minus and plus strands, respectively. The command, `bedtools intersect -c -s` command was used to find the number of overlapping 3' UTR ends in each of the shifted windows.

To determine the `window_has_gene` column, a bed file was created for each window. The command `bedtools intersect -c` was used to determine if there were any annotated genes intersecting the window. To determine the `peak_has_600bp_downstream_gene` column, a bed file was created for the region of each peak and 600bp downstream from the peak, relative to the strand of the window. If the peak was at position less than 600 and the window strand was 'minus', then the bed file was created for the region from 0 to the peak. To determine if each shifted window intersects with any annotated genes, the command `bedtools intersect -c -s` command was used for the shifted window and the annotated genes bed file.

To determine the closest upstream and downstream 3'UTR ends, a bed file was created for each peak, and `bedtools closest` was used to determine the 3' UTR ends that were the least distant from each peak. For peaks located in a 'plus' stranded window, the closest upstream 3' UTR end and its distance were determined from the output of `bedtools closest -c` of the peak bed file and the strand-respective 3' UTR ends bed file, with the `ignore downstream` flag to only capture upstream regions. The closest downstream 3' UTR end was determined with the same command, but with the `ignore upstream` flag to only capture downstream regions. For peaks located in a 'minus' stranded window, the same commands were used, but with the `ignore upstream` flag used for upstream regions and the `ignore downstream` flag used for downstream regions, in order to account for reverse strandedness.

Plot Generation

To investigate windows that showed a large range in median ReadZS values, histograms were plotted showing the number of counts at each genomic position. To plot each window, every ontology (i.e. cell type or other grouping) for that window was sorted by its median ReadZS value. The top 2 and bottom 2 ontologies were then chosen to be plotted for each window. For each ontology, pass-filter reads were extracted if they came from that window and from the cell barcodes associated with that ontology. These reads were then deduplicated, with positions rounded to a bin size of 10. Each position was then counted, by summing the number of reads at each position per ontology in that window. The count was then normalized by the total number of counts per ontology in that window, to produce a percent score.

To plot this percent score, the counts were read into Gviz (Hahne and Ivanek, 2016), and positions without count values were imputed with 0. Each ontology was used to create a data track, with a x-axis range of the window start and window end. A respective genome GFF file was used to plot the gene region track, with the GFF 'transcript' feature excluded, for visual clarity.

CDF and Histogram Generation

The GMM annotated peak tables (Supplementary Table 4) was used to create the overlaid CDF and histogram of peak distances closest to a downstream 3'UTR end. For each dataset, the table was filtered for `peak_has_600bp_downstream_gene == True` and `df.downstream_3UTR_dist < VALUE`, where VALUE is some bound on the x axis. Unless otherwise stated, all plots are made with `bins = 100` and VALUE=[200, 800000]. The histograms and CDF plots were made with the matplotlib `hist()` function with `density=True`. The CDF plots were also plotted with `cumulative=True` and `histtype='step'`. The quantiles were calculated with the pandas `quantile()` function, to determine the 25th, 50th, and 75th quantiles. The quantile cutoffs are visualised by the red-dotted lines.

Code Availability

The ReadZS pipeline is available through a GitHub repository: https://github.com/salzmanlab/ReadZS.

Data Availability

The human lung scRNA-seq data used here was generated through the Human Lung Cell Atlas project (Travaglini et al., 2020) and is accessible through European Genome-phenome Archive (accession number: EGAS00001004344). Human and mouse unselected spermatogenesis data was downloaded from the SRA databases with accession IDs SRR6459190 (AdultHuman_17-3), SRR6459191 (AdultHuman_17-4), and SRR6459192 (AdultHuman_17-5) for human, and accession IDs SRR6459155 (AdultMouse-Rep1), SRR6459156 (AdultMouse-Rep2), and SRR6459157 (AdultMouse-Rep3) for mouse. Fibroblast data from injured and uninjured mouse hearts, as analyzed in (Patrick et al., 2020), was generated by (Farbehi et al., 2019) and is available on ArrayExpress under identifier E-MTAB-7376. For GENCODE annotations, we used v37 for human and vM26 for mouse.

Figures:

Figure 1. Overview of the ReadZS.

(A) Read positions are ranked in equal-sized genomic bins, separated by read strand. Within each genomic window, the read distribution for each cell is summarized by a weighted, normalized function of read positions (Methods). With metadata, cell type-specific RNAP can be detected by finding windows with significantly different median ReadZS by cell type; subclustering on the basis of one or more genomic windows may reveal subpopulations of cells with different regulated RNAP. Continuous metadata such as pseudotime, enables discovery of multivariate relationships between ReadZS and metadata. A GMM-based clustering approach (Methods) can also be run to identify subpopulations within cell types. GMM-based peak detection is used to compare read distributions with annotated 3' UTRs.

(B) The ReadZS has unique power to discover regulation missed by peak-callers. Left: Rpl13a has two cleanly separated peaks but was not called by Sierra. Right: multiple APA sites in the final exon of Rab2a create overlapping peaks that hinder peak-calling-based methods.

(C) The ReadZS is technically reproducible across cell types in the 3 HLCA participants; p-values, computed via simulation (see Methods) show strong ReadZS concordance in all pairs.

(D) Above: read distributions of genomic windows with largest effect size in HLCA P3 when requiring minimum 10 counts in 20 cells (left) or minimum 5 counts in 10 cells (right), which overlap the genes CORO1B/PTPRCAP and CATIP1 respectively. Below: example windows called in both P3 and P2 as having cell type-specific differences in RNAP in CALM1 (left) and KLF6 (right); upper panel shows reads from P3, lower from P2. For KLF6, the relative rank of each cell type (ranked by highest to lowest median ReadZS) is shown for each participant. Peaks in significant windows called by the GMM (see text) are starred; peaks are called across all cell types. In CALM1, the peaks are 254 bp and 285 bp from the closest downstream 3' UTR.
(E) Histogram and cdfs of the distribution of distances from GMM-called peaks to closest downstream annotated 3' UTR in HLCA P3 and human spermatogenesis; lines denote the 25th, 50th, and 75th quantile, respectively. Distance distributions are compatible with expectation from 10x library construction.

Figure 2. ReadZS reveals subpopulations in macrophages. Gaussian mixture modeling of the ReadZS values reveals subpopulations in macrophages distinguished by RNAP in *BCL2A1*, *RPL35*, *COX6C*, *RPL26*, *ATP51E*, and *GNG5* in both P3 and P2 individuals (Suppl. Figure 3). Read distributions for (A), (B) *ATP51E* and (C), (D) *BCL2A1* and further statistical testing (Methods) reproducibly distinguishes cells in cluster 1 and cluster 2 in P2 and P3. ReadZS distribution for each window across macrophages shows two subpopulations. (B) and (D) show the empirical cumulative distribution functions for *ATP51E* and *BCL2A1*, respectively, in both donors. (E) High concordance between the subcluster assignments based on the identified marker for both participants. (F) The p values from the chi-squared test of independence are < 10^-6 between the 6 genes in both individuals.

Figure 3. Model-based identification of peaks in read distributions.
(A) The ReadZS detects a global trend of 3' UTR shortening in both human (left) and mouse (right) spermatogenesis datasets, as indicated by significant negative correlation between ReadZS and pseudotime. Significance is defined as abs(Spearman correlation) > 0.3 and Bonferroni corrected p < 0.05. Histogram bin width = 0.2.
(B) The ReadZS reveals evolutionarily conserved 3' UTR regulation in human and mouse. Left: windows containing the 3' end of ZFAND6 were significantly correlated with pseudotime in both human (Spearman = -0.34113, Bonferroni corrected p < 1E-39) and mouse (Spearman = -0.75702, Bonferroni corrected p < 1E-84). Vertical red lines indicate peak positions. Right: the 3' UTR region of ZFAND6 in human shows high conservation with other vertebrates (UCSC Genome Browser). Red lines correspond to peak positions from the left plot.
(C) The ReadZS discovers finescale developmental regulation of RNAP in human spermatogenesis, including in regions where neighboring APA sites create highly overlapping peaks. Left, top to bottom: windows with significant correlation between ReadZS and pseudotime within the first 0 to 25% of pseudotime: OAZ1, which is both significantly correlated over all pseudotime (Spearman = 0.13111, Bonferroni corrected p = 1.5E-06) and within the first 0 to 25% of pseudotime (Spearman = 0.58974, Bonferroni corrected p < 1E-55); and MED21, which is only significantly correlated when restricting the first 0 to 25% of pseudotime (Spearman = -0.53554, Bonferroni corrected p < 1E-23), not across all of pseudotime. Right, top to bottom: the windows with the two highest correlation values when calculated over all of

pseudotime: ARPP19 (Spearman = 0.74012, Bonferroni corrected p < 1E-223) and S100A10 (Spearman = 0.66626, Bonferroni corrected p < 1E-88). Red lines highlight peak positions.

Supplementary Figures:

Supplementary figure 1: Read distributions of genomic windows in P3 with significantly different cell type-specific RNA processing, overlapping the genes RPLP1, NEAT1, and SRSF7 (left to right). Peaks in significant windows called by the GMM (see text) are starred; peaks are called across all cell types.

Supplementary figure 2: Histogram and cdfs of the distribution of distances from GMM-called peaks to closest downstream annotated 3' UTR in HLCA P2 and mouse spermatogenesis; lines denote the 25th, 50th, and 75th quantile, respectively. Distance distributions are compatible with expectation from 10x library construction

Supplementary figure 3: Read distribution, ReadZS distribution, and the empirical cumulative distribution function across in each subcluster and HLCA donor for (A) *RPL26*, (B) *COX6C*, (C) *GNG5,* and (D) *RPL35*.

Supplementary figure 4: UMAP visualization of macrophages in both P3 and P2 based on the ReadZS values. Violin plots show the distribution of the ReadZS values for these 6 genes in cluster 5 versus other clusters.

Supplementary figure 5: Pearson's correlation coefficients between the ReadZS values across macrophages for each pair of marker genes and in each P3 and P2 individual are positive.

Supplementary figure 6: Read distributions of genomic windows in human spermatogenesis with significant correlation between ReadZS and pseudotime. Left: window overlapping TSSK1B (Spearman = -0.54894, Bonferroni corrected p < 1E-82); right: window overlapping SLC25A37 (Spearman = -0.51287, Bonferroni corrected p < 1E-35). Red lines are placed to highlight peaks in read distribution.

Bibliography

Agarwal, V., Lopez-Darwin, S., Kelley, D.R., and Shendure, J. (2021). The landscape of alternative polyadenylation in single cells of the developing mouse embryo. Nat. Commun. *12*, 5101.

Bae, B., Gruner, H.N., Lynch, M., Feng, T., So, K., Oliver, D., Mastick, G.S., Yan, W., Pieraut, S., and Miura, P. (2020). Elimination of Calm1 long 3'-UTR mRNA isoform by CRISPR-Cas9 gene editing impairs dorsal root ganglion development and hippocampal neuron activation in mice. RNA *26*, 1414–1430.

Bao, J., Vitting-Seerup, K., Waage, J., Tang, C., Ge, Y., Porse, B.T., and Yan, W. (2016). UPF2-Dependent Nonsense-Mediated mRNA Decay Pathway Is Essential for Spermatogenesis by Selectively Eliminating Longer 3'UTR Transcripts. PLoS Genet. *12*, e1005863.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. IEEE Trans Pattern Anal Mach Intell *22*, 719–725.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

Chung, E., and Romano, J.P. (2013). Exact and asymptotically robust permutation tests. Ann. Statist. *41*, 484–507.

Di Giammartino, D.C., Nishida, K., and Manley, J.L. (2011). Mechanisms and consequences of alternative polyadenylation. Mol. Cell *43*, 853–866.

Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nat. Biotechnol. *35*, 316–319.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dong, P., Xiong, Y., Yue, J., Hanley, S.J.B., Kobayashi, N., Todo, Y., and Watari, H. (2018). Long Non-coding RNA NEAT1: A Novel Target for Diagnosis and Therapy in Human Tumors. Front. Genet. *9*, 471.

Durrett, R. (2019). Probability: theory and examples (Cambridge University Press).

Engström, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., Rätsch, G., Goldman, N., Hubbard, T.J., Harrow, J., Guigó, R., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. Nat. Methods *10*, 1185–1191.

Farbehi, N., Patrick, R., Dorison, A., Xaymardan, M., Janbandhu, V., Wystub-Lis, K., Ho, J.W., Nordon, R.E., and Harvey, R.P. (2019). Single-cell expression profiling reveals dynamic flux of cardiac stromal, vascular and immune cells in health and injury. ELife *8*.

Floor, S.N., and Doudna, J.A. (2016). Tunable protein synthesis by transcript isoforms in human cells. ELife *5*.

Gao, Y., Li, L., Amos, C.I., and Li, W. (2021). Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression.

Genome Res.

Hahne, F., and Ivanek, R. (2016). Visualizing genomic data using gviz and bioconductor. Methods Mol. Biol. *1418*, 335–351.

Hermann, B.P., Cheng, K., Singh, A., Roa-De La Cruz, L., Mutoji, K.N., Chen, I.-C., Gildersleeve, H., Lehle, J.D., Mayo, M., Westernströer, B., et al. (2018). The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids. Cell Rep. *25*, 1650-1667.e8.

Huntsman Cancer Institute (2021). 10X Genomics 3' Gene Expression - University of Utah. https://uofuhealth.utah.edu/huntsman/shared-resources/gba/htg/single-cell/genomics-10x.php

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163–166.

Kang, B., Jiang, D., Ma, R., He, H., Yi, Z., and Chen, Z. (2017). OAZ1 knockdown enhances viability and inhibits ER and LHR transcriptions of granulosa cells in geese. PLoS ONE *12*, e0175016.

Königs, V., de Oliveira Freitas Machado, C., Arnold, B., Blümel, N., Solovyeva, A., Löbbert, S., Schafranek, M., Ruiz De Los Mozos, I., Wittig, I., McNicoll, F., et al. (2020). SRSF7 maintains its homeostasis through the expression of Split-ORFs and nuclear body assembly. Nat. Struct. Mol. Biol. *27*, 260–273.

Li, W., Park, J.Y., Zheng, D., Hoque, M., Yehia, G., and Tian, B. (2016). Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control. BMC Biol. *14*, 6.

Lusk, R., Stene, E., Banaei-Kashani, F., Tabakoff, B., Kechris, K., and Saba, L.M. (2021). Aptardi predicts polyadenylation sites in sample-specific transcriptomes using high-throughput RNA sequencing and DNA sequence. Nat. Commun. *12*, 1652.

Mayr, C. (2019). What are 3' utrs doing? Cold Spring Harb. Perspect. Biol. *11*.

Miles, L.A., and Parmer, R.J. (2010). S100A10: a complex inflammatory role. Blood *116*, 1022–1024.

Narla, G., Heath, K.E., Reeves, H.L., Li, D., Giono, L.E., Kimmelman, A.C., Glucksman, M.J., Narla, J., Eng, F.J., Chan, A.M., et al. (2001). KLF6, a candidate tumor suppressor gene mutated in prostate cancer. Science *294*, 2563–2566.

Ndiaye, F.K., Ortalli, A., Canouil, M., Huyvaert, M., Salazar-Cardozo, C., Lecoeur, C., Verbanck, M., Pawlowski, V., Boutry, R., Durand, E., et al. (2017). Expression and functional assessment of candidate type 2 diabetes susceptibility genes identify four new genes contributing to human insulin secretion. Mol. Metab. *6*, 459–470.

Patrick, R., Humphreys, D.T., Janbandhu, V., Oshlack, A., Ho, J.W.K., Harvey, R.P., and Lo, K.K. (2020). Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. Genome Biol. *21*, 167.

Shulman, E.D., and Elkon, R. (2019). Cell-type-specific analysis of alternative polyadenylation

using single-cell transcriptomics data. Nucleic Acids Res. *47*, 10027–10039.

Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. *27*, 491–499.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902.e21.

Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R.V., Chang, S., Conley, S.D., Mori, Y., Seita, J., et al. (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature *587*, 619–625.

Tushev, G., Glock, C., Heumüller, M., Biever, A., Jovanovic, M., and Schuman, E.M. (2018). Alternative 3' UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. Neuron *98*, 495-511.e6.

Virshup, D.M., and Kaldis, P. (2010). Cell biology. Enforcing the Greatwall in mitosis. Science *330*, 1638–1639.

Wilusz, C.J., Wormington, M., and Peltz, S.W. (2001). The cap-to-tail guide to mRNA turnover. Nat. Rev. Mol. Cell Biol. *2*, 237–246.

Wu, X., Liu, T., Ye, C., Ye, W., and Ji, G. (2021). scAPAtrap: identification and quantification of alternative polyadenylation sites from single-cell RNA-seq data. Brief. Bioinformatics *22*.

Ye, C., Zhou, Q., Wu, X., Yu, C., Ji, G., Saban, D.R., and Li, Q.Q. (2020). scDAPA: detection and visualization of dynamic alternative polyadenylation from single cell RNA-seq data. Bioinformatics *36*, 1262–1264.

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y., and Wang, J. (2019). Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. Mol. Cell *73*, 130-142.e5.

Figure 1

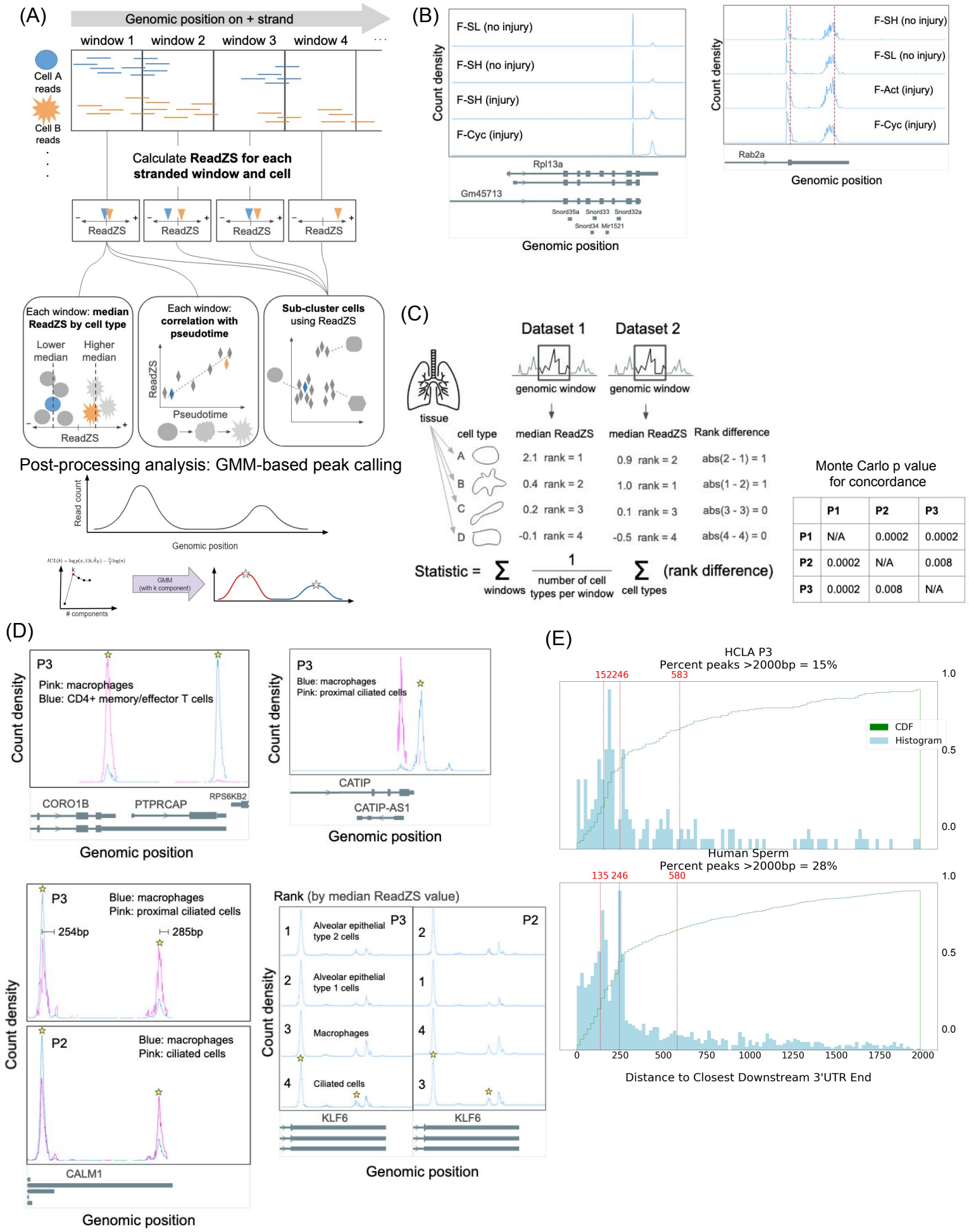

(A)
Genomic position on + strand
window 1
window 2
window 3
window 4
Cell A reads
Cell B reads
Calculate ReadZS for each stranded window and cell
ReadZS
Each window: median ReadZS by cell type
Lower median
Higher median
Each window: correlation with pseudotime
Pseudotime
Sub-cluster cells using ReadZS
Post-processing analysis: GMM-based peak calling
Read count
Genomic position
GMM (with k component)
components
(B)
F-SL (no injury)
F-SH (no injury)
F-SH (injury)
F-Cyc (injury)
Count density
Rpl13a
Gm45713
Snord35a
Snord33
Snord32a
Snord34
Mir1521
Genomic position
F-SH (no injury)
F-SL (no injury)
F-Act (injury)
F-Cyc (injury)
Rab2a
Genomic position
(C)
Dataset 1
Dataset 2
genomic window
genomic window
tissue
cell type
median ReadZS
median ReadZS
Rank difference
A 2.1 rank = 1 0.9 rank = 2 abs(2 - 1) = 1
B 0.4 rank = 2 1.0 rank = 1 abs(1 - 2) = 1
C 0.2 rank = 3 0.1 rank = 3 abs(3 - 3) = 0
D -0.1 rank = 4 -0.5 rank = 4 abs(4 - 4) = 0
Statistic = Σ windows 1 / number of cell types per window Σ cell types (rank difference)
Monte Carlo p value for concordance
P1 P2 P3
P1 N/A 0.0002 0.0002
P2 0.0002 N/A 0.008
P3 0.0002 0.008 N/A
(D)
P3
Pink: macrophages
Blue: CD4+ memory/effector T cells
CORO1B
PTPRCAP
RPS6KB2
Count density
Genomic position
P3
Blue: macrophages
Pink: proximal ciliated cells
CATIP
CATIP-AS1
Count density
Genomic position
P3
Blue: macrophages
Pink: proximal ciliated cells
254bp
285bp
P2
Blue: macrophages
Pink: ciliated cells
CALM1
Count density
Genomic position
Rank (by median ReadZS value)
P3
P2
1 Alveolar epithelial type 2 cells
2 Alveolar epithelial type 1 cells
3 Macrophages
4 Ciliated cells
KLF6
KLF6
Count density
Genomic position
(E)
HCLA P3
Percent peaks >2000bp = 15%
152246
583
CDF
Histogram
Human Sperm
Percent peaks >2000bp = 28%
135 246
580
0
250
500
750
1000
1250
1500
1750
2000
Distance to Closest Downstream 3'UTR End

Figure 2

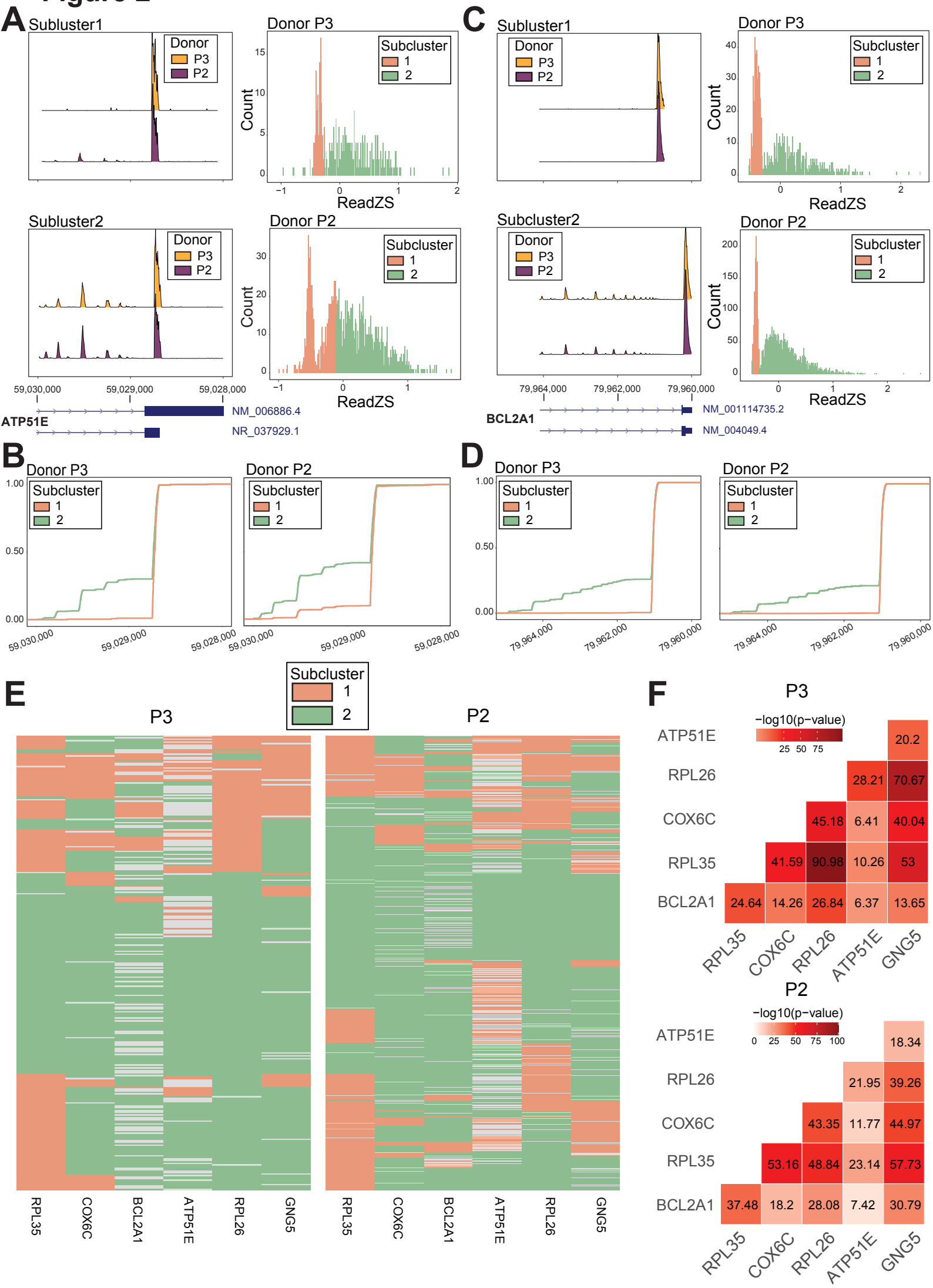
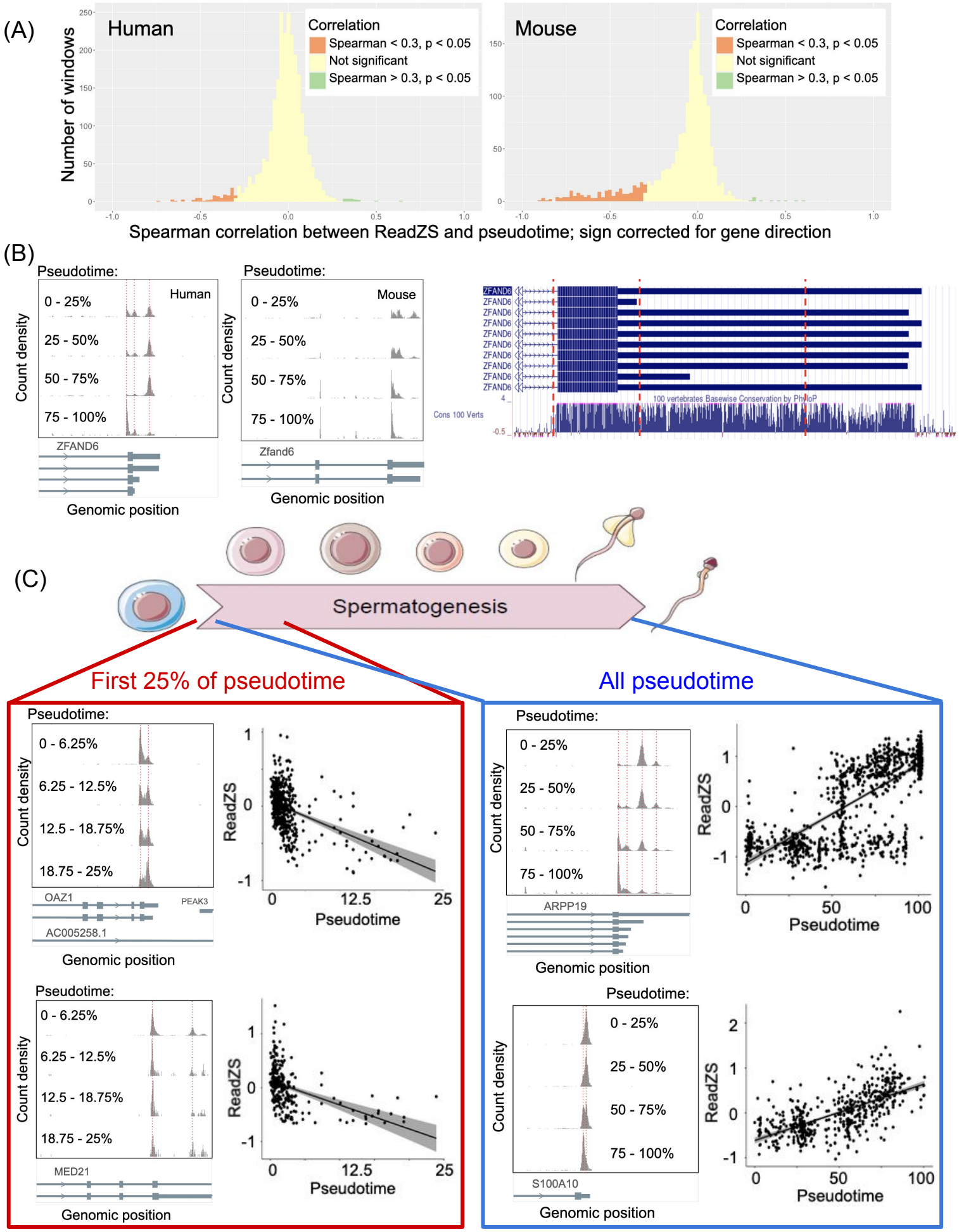
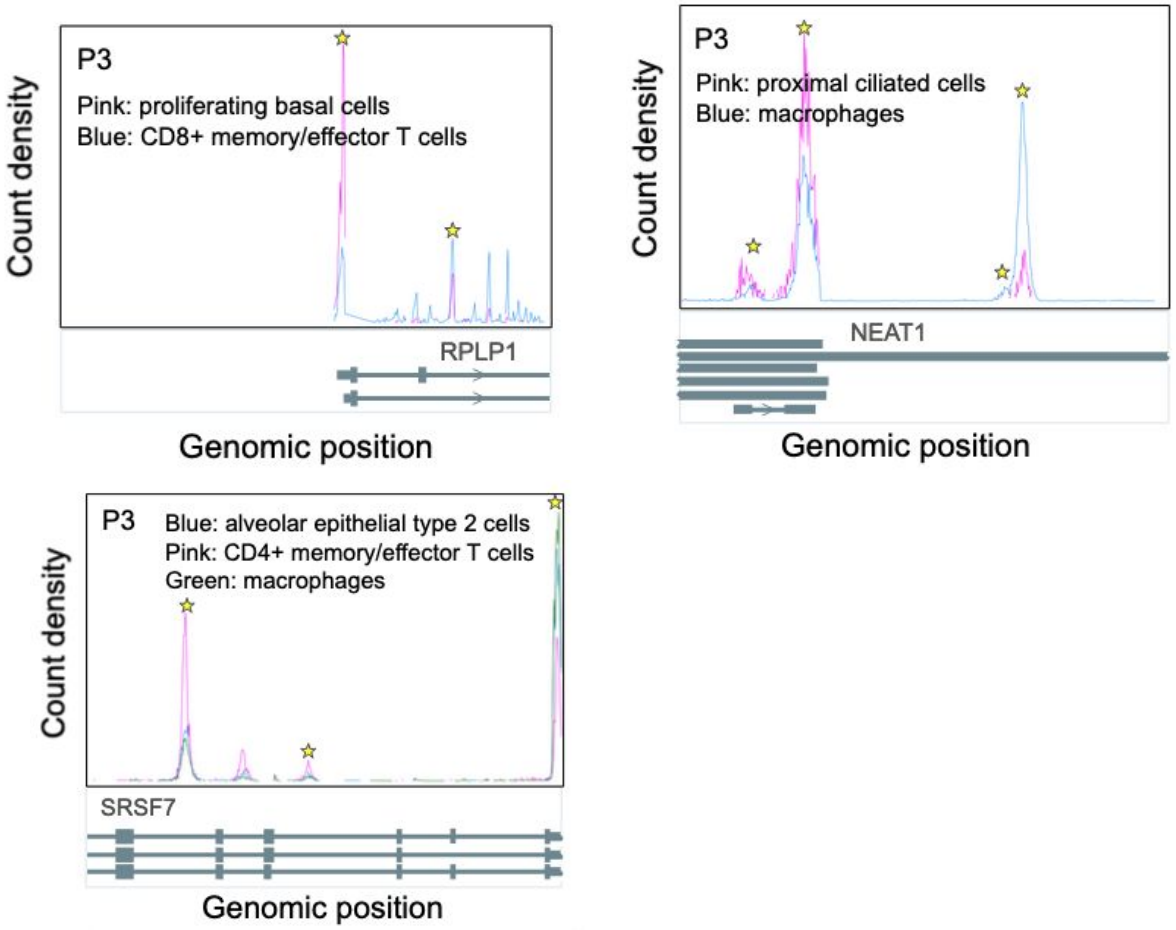

A
Subluster1
Donor
P3
P2
Donor P3
Subcluster
1
2
Count
ReadZS
Subluster2
Donor P2
59,030,000
59,029,000
59,028,000
ATP51E
NM_006886.4
NR_037929.1
C
Subcluster1
Subcluster2
79,964,000
79,962,000
79,960,000
BCL2A1
NM_001114735.2
NM_004049.4
B
D
E
P3
P2
RPL35
COX6C
BCL2A1
ATP51E
RPL26
GNG5
F
−log10(p−value)
ATP51E
RPL26
COX6C
RPL35
BCL2A1
20.2
28.21 70.67
45.18 6.41 40.04
41.59 90.98 10.26 53
24.64 14.26 26.84 6.37 13.65
18.34
21.95 39.26
43.35 11.77 44.97
53.16 48.84 23.14 57.73
37.48 18.2 28.08 7.42 30.79

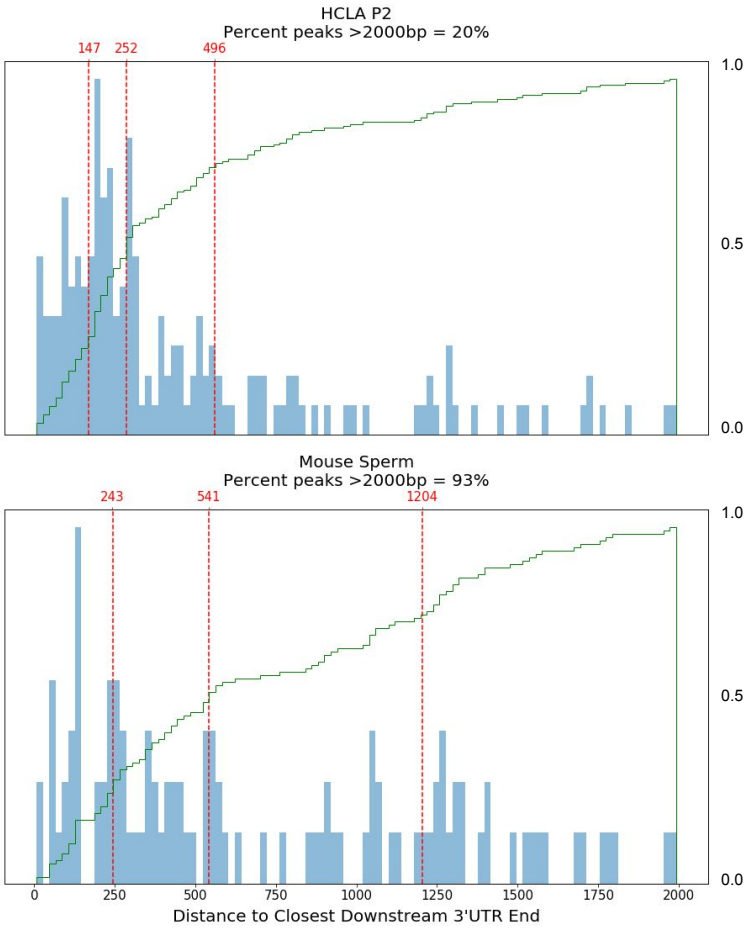
Figure 3


(A)
Human
Mouse
Correlation
Spearman < 0.3, p < 0.05
Not significant
Spearman > 0.3, p < 0.05
Number of windows
Spearman correlation between ReadZS and pseudotime; sign corrected for gene direction
(B)
Pseudotime:
0 - 25%
25 - 50%
50 - 75%
75 - 100%
Count density
ZFAND6
Zfand6
Genomic position
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
(C)
Spermatogenesis
First 25% of pseudotime
All pseudotime
0 - 6.25%
6.25 - 12.5%
12.5 - 18.75%
18.75 - 25%
OAZ1
PEAK3
AC005258.1
MED21
ARPP19
S100A10
ReadZS
Pseudotime

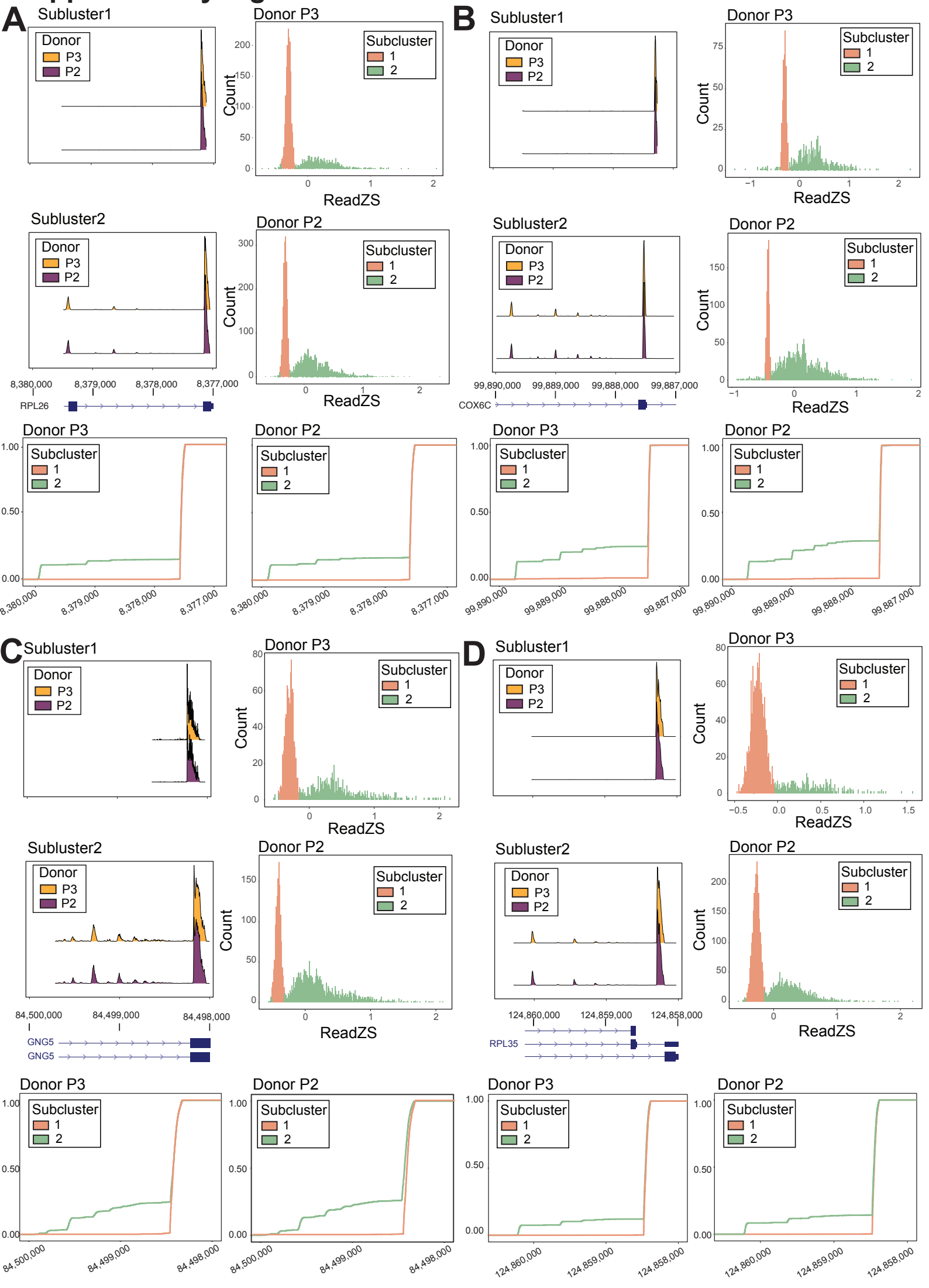
Supplementary Figure 1



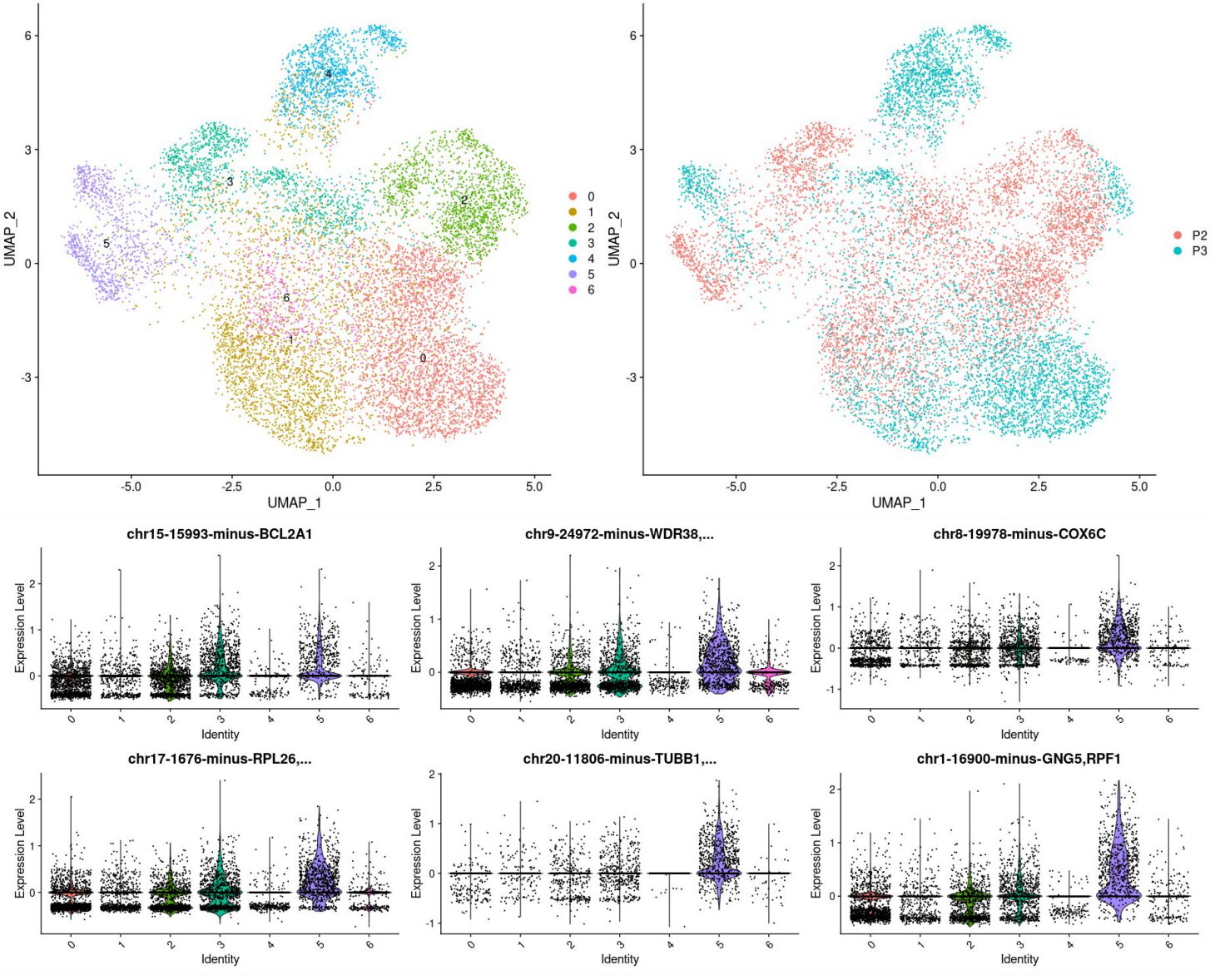
Supplementary Figure 2



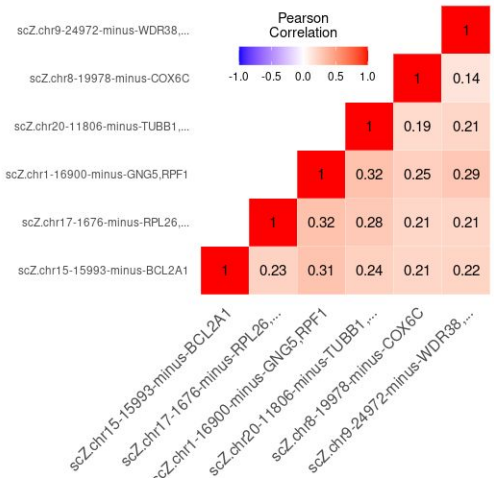
Supplementary Figure 3

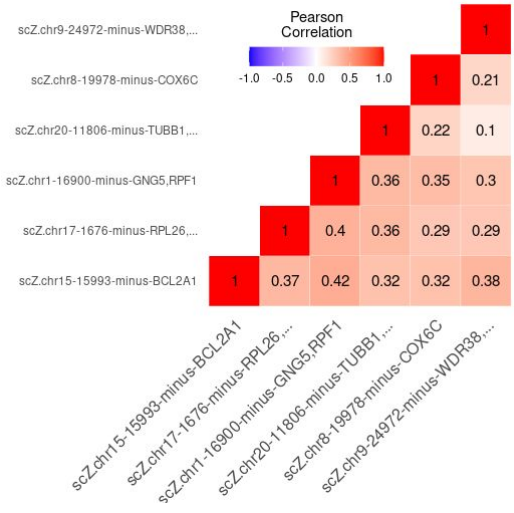
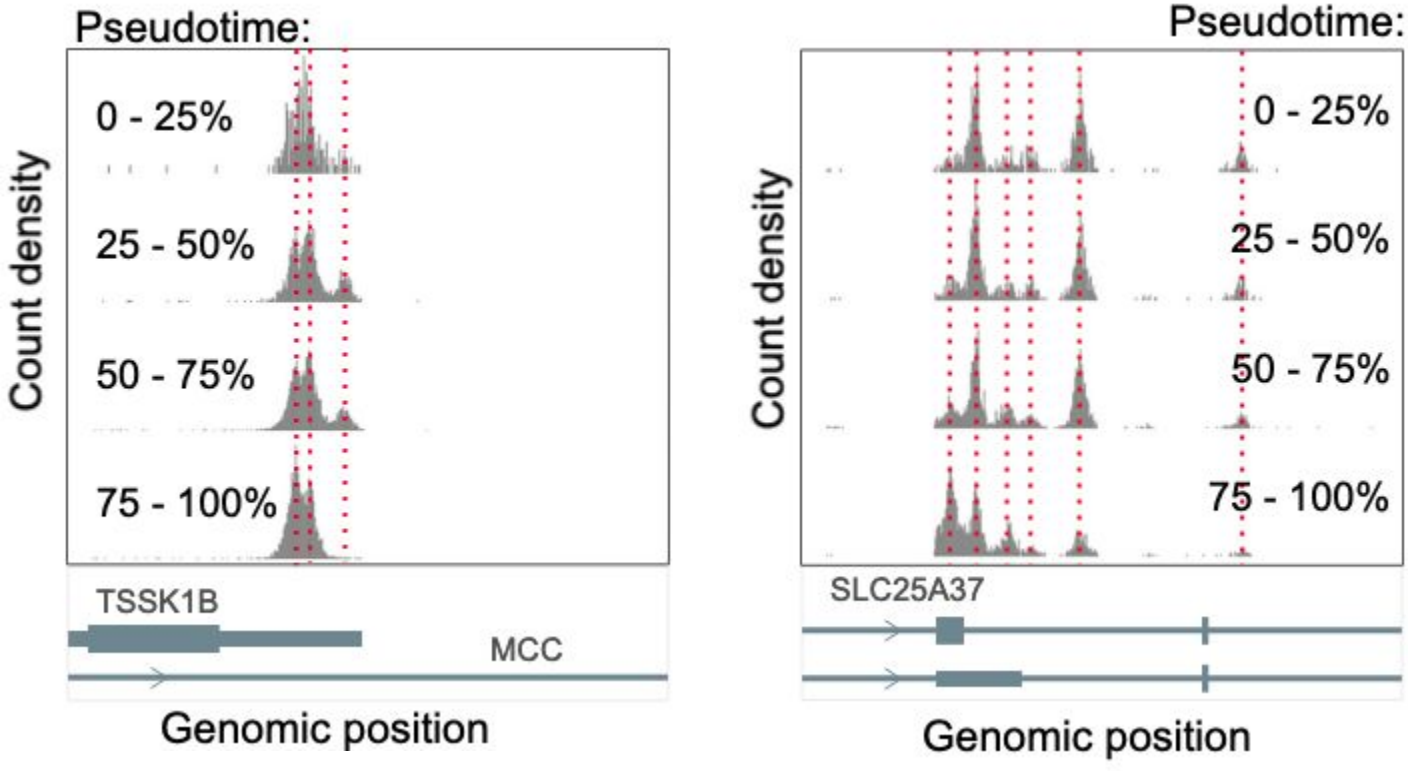


Supplementary Figure 4



Supplementary Figure 5





Supplementary Figure 6