Title

Statistical analysis supports pervasive RNA subcellular localization and alternative 3' UTR regulation.

Abstract

Targeted low-throughput studies have previously identified subcellular RNA localization as necessary for cellular functions including polarization, and translocation. Further, these studies link localization to RNA isoform expression, especially 3' Untranslated Region (UTR) regulation. The recent introduction of genome-wide spatial transcriptomics techniques enable the potential to test if subcellular localization is regulated in situ pervasively. In order to do this, robust statistical measures of subcellular localization and alternative poly-adenylation (APA) at single cell resolution are needed. Developing a new statistical framework called SPRAWL, we detect extensive cell-type specific subcellular RNA localization regulation in the mouse brain and to a lesser extent mouse liver. We integrated SPRAWL with a new approach to measure cell-type specific regulation of alternative 3' UTR processing and detected examples of significant correlations between 3' UTR length and subcellular localization. Included examples, *Timp3*, *Slc32a1*, *Cxcl14*, and *Nxph1* have subcellular localization in the brain highly correlated with regulated 3' UTR processing that includes use of unannotated, but highly conserved, 3' ends. Together, SPRAWL provides a statistical framework to integrate multi-omic single-cell resolved measurements of gene-isoform pairs to prioritize an otherwise impossibly large list of candidate functional 3' UTRs for functional prediction and study. SPRAWL predicts 3' UTR regulation of subcellular localization may be more pervasive than currently known.

Introduction

Subcellular protein localization is widely accepted to be critical to function [1], but the biological cause and consequence of RNA localization within a cell remains poorly understood. Targeted studies on subcellular RNA localization have identified motivating examples such as enrichment of beta-actin RNA to lamellipodia in motile chicken embryonic myoblasts [2]. Approximately 80% of total actin mRNA localized to the lamellipodia, and specific disruption of localization, but not expression, of the mRNA resulted in decreased cell motility [3,4]. In a larger scale study it has been estimated that 70% of mRNAs are spatially localized in *Drosophila* embryogenesis [5]. Other well-known and recently identified examples of RNA subcellular localization with functional consequences include lipid droplets [6] and TIS granules [7]. In these instances, either alternative splicing or alternative 3' Untranslated Region (UTR) usage results in differentially localizing RNA sequences.

While the vast majority of 3' UTR isoform functions remain unknown and incompletely annotated, emerging evidence points to an abundance of cell-type specific regulation [8] where different 3' UTRs may even have opposite functions. Cd47, for example, expresses a long-isoform 3' UTR that results in a peripherally localized protein product protecting against phagocytosis, but can also express a short-isoform 3' UTR that results in a cytoplasmic protein product with the same amino-acid sequence that does not confer the same phagocytotic protection [9]. Control of RNA subcellular localization through RNA isoform choice may help pinpoint functions for alternative RNA isoforms and UTRs in eukaryotes.

Spatial transcriptomics has seen rapidly increasing interest as methods become increasingly powerful and affordable, and as consortiums form to consolidate efforts [10], however work remains primarily focused on gene expression. Techniques such as MERFISH [11], and its commercialization Vizgen, as well as SeqFISH+ [12] utilize sequential multiplexed imaging to localize hundreds to thousands of RNA species across a tissue with subcellular resolution. Along with RNA-capture based spatial transcriptomics techniques [13–15], these spatial datasets have primarily been used to analyze the distribution of cell-types within a tissue via gene expression. At a finer scale, RNA distribution within cells has been understudied despite an established history of biologically important case studies [16]. The limited approaches that have been used to detect subcellular localization patterns from high throughput, high resolution spatial datasets rely on co-stains and/or heuristics without statistical formalism [17–19].

Statistical approaches to identify subcellular RNA localization are lacking in the field. As an example, a SeqFISH+ dataset was used to descriptively identify peripherally and centrally localizing genes in different mouse cortex cell-types. Thresholding was used to find the strongest signals of localizing genes, but weaker spatial patterns were

potentially missed and the analysis did not control the false discovery rate (FDR) [12]. MERFISH has been used to detect subcellular localization patterns in cell-lines, identifying differences in neuron soma, axon, and dendrite compartment transcriptomes using the Wilcoxon rank-sum test and Moran's I [20], but this analysis has not been applied to tissue samples where neuronal projections cannot be segmented nor has it identified continuous, non-compartment based, subcellular localization [21]. While proximity-tagging and sequencing approaches such as APEX-seq [22,23] have generated high-plex datasets for RNA localization within subcellular compartments, these methods require genetically modified cell-lines, and cannot be readily applied to tissue. Further, no study has attempted to test whether isoform regulation can explain subcellular localization at the gene level in massively multiplexed FISH datasets.

To address the limitations of prior approaches, we introduce Subcellular Patterning Ranked Analysis With Labels (SPRAWL) as a transparent and statistical approach to detect RNA subcellular patterning from multiplexed imaging datasets. SPRAWL assigns interpretable per-gene/per-cell scores to detect RNA localization patterns between cell-types and biological replicates. SPRAWL scores can be analyzed in closed forms and provide FDR control. SPRAWL currently identifies continuous peripheral, central, radial, and punctate localization patterns of a gene within a cell-type which are significantly more extreme than expected by chance in either direction of effect. SPRAWL can be extended to detect user-defined patterns and represents a general framework for unbiased discovery of RNA subcellular localization patterns from multiplexed imaging datasets. This unbiased integrative approach identifies genes with potential cis-regulatory sequences, and prioritizes candidates for experimental follow-ups.

Results

SPRAWL robustly detects subcellular localization

SPRAWL is a non-parametric single-cell resolved measure of RNA subcellular localization that is robust against confounding variables of cell size, and RNA expression level, while providing effect-size and statistical significance measures (Figure 1 and Supplemental Figure 1). In this work, SPRAWL peripheral, central, punctate, and radial two-sided metrics have been used to analyze spatial datasets comprising a total of 22 experiments over 4 mice processed by three different research groups and two technologies [12,24,25]. Applying SPRAWL to these datasets revealed: (1) gene/cell-type localization patterns have high correlation between biological replicates; (2) differential subcellular localization patterns of the same gene in different cell-types; and (3) differential subcellular regulation corresponding with cell-type differential 3' UTR length from associated scRNAseq datasets [26] for 26 genes including *Timp3*, *Slc32a1, Cxcl14* and *Nxph1*.

The SPRAWL workflow preprocesses spatial datasets into an HDF5 format that facilitates rapid data retrieval and standardization (Figure 1: Input and Preprocessing). SPRAWL gene/cell scores are calculated using each peripheral, central, radial, and punctate metric to identify more extreme subcellular patterning than expected by chance if the gene-identity of each RNA spot were randomly permuted. The peripheral metric identifies extremely proximal or distal genes in relation to the cell-membrane. The central metric is similar but measures in relation to the cell centroid. The radial score identifies RNA species with spots that tend to aggregate in one angular sector of the cell. Finally the punctate score represents the degree to which RNA spots from a given gene are clustered together, scores closer to 1 indicate self-colocalizing genes. The peripheral and central metrics have exact p-values under the null, while the radial and punctate scores both rely on permutation testing to create a null distribution for each gene in each cell (Methods). Gene/cell-type scores are then calculated as the mean SPRAWL gene/cell scores to provide an aggregate measure of RNA localization per gene per cell-type with statistical rigor.

Reduction of complex spatial patterns into two-sided, one-dimensional metrics produces interpretable and comparable RNA localization scores. An additional benefit of SPRAWL scores is their direct integration with other statistical methods: scores can be analyzed through the lens of various metadata such as cell type, or correlated with other measures such as RNA 3' UTR regulation or splicing state. Artificial intelligence and machine learning approaches are avoided to allow for direct control of FDR. SPRAWL is a publicly available Python package that can be installed using pypi and has also been implemented in Nextflow [27] and Docker for reproducible analyses at large scale in high-performance or cloud computing environments. SPRAWL source code and documentation are available at https://github.com/The-Salzman-Lab/SPRAWL.

There exists no directly competing method to SPRAWL which is able to leverage highly multiplexed imaging datasets, requiring only RNA spot locations, cell-boundary estimates, and RNA identity of each spot. Many current software approaches aim to discretize RNA patterning into subcompartments and rely on co-stains which are not guaranteed to be present in every dataset. Other approaches use statistically opaque machine-learning based classifiers to assign RNA spots to pre-specified patterns [28]. As spatial transcriptomics methods are commercialized and become more

accessible, increasing numbers of public datasets will become available and can be processed by SPRAWL regardless of the tissue or study design.

Cell-type specific subcellular localization is regulated across replicates of three datasets

Sensitivity of SPRAWL to segmentation and cell-boundary locations was tested by computationally shrinking the cell-boundaries. Median peripheral scores per gene/cell-type were significantly correlated between original cell-boundaries and shrunk cell-boundaries with Pearson correlation coefficient of 0.85 on the mouse motor cortex datasets (Supplemental Figure 1e), suggesting empirically that SPRAWL would have low sensitivity to potential cell-segmentation errors.

We applied SPRAWL to the BICCN motor cortex (MOp) [24], Vizgen Brainmap, and Vizgen Liver datasets [25] which each contained either biological or technical replicates. The median SPRAWL gene/cell-type scores were significantly positively correlated between replicates within all three datasets for all four spatial metrics having significant Pearson correlation with coefficients larger than 0.8, Spearman correlation coefficients larger than 0.72, at an alpha level of 0.05 (Figure 1: Results).

In the MOp dataset, SPRAWL detects hundreds of significantly patterned gene/cell-type groups. The MOp dataset imaged 252 genes through multiplexed barcoding, including 10 negative-control barcodes, and profiled nearly 300,000 cells from the mouse motor cortex [24]. Biological replicates were present from two mice (Figure 1: Results top row) with 6 slices taken from each animal. Conservative filtering of cells and cell-types (see Methods: SPRAWL Filtering) resulted in 220 unique genes and 19 distinct cell-types, observed in 1999 gene/cell-type combinations. After BH multiple hypothesis testing correction over both replicates, 1511 (75.6%) gene/cell-type pairs were significant by the peripheral metric, 1492 74.6%) by the central metric, 1475 (73.8%) by the radial metric, and 1448 (72.4%) by the punctate metric. Spatial patterning was extensive and consistent between replicates with more than 77.8% of the gene/cell-type pairs having the same direction of effect, positive or negative, between the two replicates. 176 of 220 (80%) unique genes were found to be significantly spatially regulated in at least one cell-type in all metrics, but not necessarily the same cell-type in all metrics. Similarly all 19 cell-types were observed to be significant with at least one gene in each metric (Supplemental Table 1).

To provide further support of the surprisingly high pervasiveness of subcellular RNA patterning we tested the specificity of SPRAWL on permuted data and the included Blank MERFISH codes. Correlations between the permuted data depicted in orange (Figure 1: Results) were insignificant, as to be expected from the empirical null. Additionally, the null datasets did not have a bias towards positive radial or punctate scoring. The 10 included BICCN MOp negative control Blank barcodes provided additional validation of the specificity of SPRAWL. These Blank barcodes, which do not correspond to actual genes and are therefore not expected to have spatial patterning, were determined to be significantly spatially regulated in only the radial and punctate metric. Out of the 248 unique genes that had statistically significant patterning in at least one cell-type, 191 had smaller BH-corrected p-values than the smallest Blank gene corrected p-value for the punctate metric. For the radial metric, 232 of the 241 unique significant genes had a smaller p-value than the minimum Blank gene corrected p-value. Similarly 121 genes in the peripheral and 153 genes in the central metrics had a smaller p-value than the minimum Blank gene corrected p-values. Adjusting the p-value thresholds to filter out Blank genes would still retain the vast majority of non-Blank gene hits.

We analyzed two other datasets, Vizgen Brain and Vizgen Liver. The Vizgen Brainmap dataset contains nine MERFISH experiments from three coronal sections of a mouse brain. Each section contains three adjacent cryotome slices from the same animal that are considered pseudo-biological replicates (Figure 1: Results middle row). Approximately 70,000 cells and 649 genes, of which 165 were Blank negative controls, were imaged. Cell-type annotations were not provided for this dataset, and instead a simple clustering of cells by gene count from the spatial data was performed using scanpy [29] that resulted in 42 cell-type proxies (Methods, Brainmap clustering). Analysis of the three brain slices resulted in 158 (54.7%), and 159 (55.0%), 139 (48.1%), and 156 (54.0%) unique genes significant in at least one cell-type for the peripheral, central, radial, and punctate analysis respectively. For the peripheral metric, 2,535 of 2,877 (88.1%) gene/cell-type groups present in all three tissue slices had significant Benjamini-Hochberg corrected p-values ($p < 0.05$). A similar 87.7% of gene/cell-types were significant according to the centrality metric. For the radial metric, 1,194 of 2,877 (51.6%) gene/cell-type groups were significant, while the punctate metric identified 2,196 of 2,877 (76.3%) of the gene/cell-type pairs as significant.

In the same Vizgen Brainmap dataset, all slices from all sections were pairwise significantly correlated for the peripheral, radial, and punctate metrics with a minimum Pearson correlation coefficient of 0.55. Cell-type SPRAWL correlation results were insensitive to different cell-type clustering parameters (Supplemental Figure 2), suggesting that the agreement between biological replicates found by SPRAWL is robust to different granularities of clustering; a desirable trait since cell-type clustering approaches vary widely. In these datasets without human-curated cell-type labels, few

gene/cell-type pairs are called by SPRAWL potentially supporting its specificity. No Blank genes were called significant in any cell-type pairings across all replicates.

The Vizgen Liver dataset consists of two mice, each with two replicates for a total of four MERFISH experiments (Figure 1: Results bottom row). Spatial data was collected on more than 1 million liver cells across all four datasets and 589 distinct genes were imaged, of which 127 were Blank genes. No Blank genes were called significant in any cell-type pairings across all replicates, again showing SPRAWL specificity. As with the Vizgen Brainmap dataset, no cell-type annotations were provided and naive clustering was performed to generate pseudo-annotations. SPRAWL called 385 unique genes and 53 or 54 unique pseudo cell-types for each metric/mouse combination filtering out gene/cell-type groups with fewer than 20 cells. SPRAWL detected 112 (29.1%) peripheral, 112 (29.1%) central, 118 (30.6%) radial, and 134 (34.8%) punctate genes significant in at least one cell-type. Median SPRAWL scores per gene/cell-type were highly correlated between the biological replicates with Pearson correlation coefficients of 0.80, 0.74, 0.62, 0.77 for the peripheral, central, radial, and punctate metrics respectively. The peripheral metric identified 1,399 of 1,642 (85.2%) of gene/cell-types present in both mice with median RNA spot count >=5 to be significant with similar percentages of 85.1%, 51.4%, and 77.4% in the central, radial, and punctate metrics.

The vast majority of RNA species have significant positive self-radial and self-punctate clustering

As expected under a model of active localization, over 99% of significant gene/cell-type groups have positive radial and punctate scores, revealing a significant and general tendency of RNAs to colocalize with other RNAs of the same gene both by euclidean distance (punctate metric), and angular dispersion (radial metric). In comparison, the SPRAWL peripheral metric in the BICCN MOp dataset identifies 828 significant gene/cell-type pairs, of which 56.1% are more peripheral and 43.9% are less peripheral than expected under the null of random gene-label assignments. Similarly, the SPRAWL central metric identifies 45.2% of significantly positive scoring gene/cell-type pairs. Empirical CDF plots of SPRAWL metric scores provide an alternate view for the same phenomenon (Supplemental Figure 1a,1b). SPRAWL detects 112 of 252 genes (44.4%) as positively punctate and radial in all cell-types which express them including extreme genes, such as Claudin 5 (*Cldn5*) which has a median SPRAWL punctate and radial score of 0.85 and 0.84 respectively (Figure 2b purple ticks) as well as VEGFR-1 (*Flt1*) which has a median SPRAWL punctate and radial score of 0.83 for both metrics. A random selection of *Cldn5* and *Flt1* cell images illustrates high colocalization of RNA spots (Figure 2c). *Cldn5* protein product is the primary integral membrane protein component of tight junctions in mouse brain and knockouts result in postnatal death [30]. It is currently not known in the literature whether *Cldn5* RNA localization is regulated. *Flt1* is a transmembrane tyrosine kinase receptor that binds vascular endothelial growth factor (VEGFR) which also has shortened a soluble protein isoform [31]. We note that imaging errors resulting in calling a single RNA molecule as two nearby molecules could artificially inflate the radial and punctate scores and potentially inflate significance calls.

SPRAWL detects genes with opposite and cell-type dependent RNA localization

We defined opposite-directionality genes as those that have the pattern of being significantly positively scoring in one cell-type, while being significantly negatively scoring in another cell-type for the same metric. Significant spatial patterning of a gene in only a subset of cell-types suggests differences in either cis or trans-acting regulatory factors. For the BICCN dataset out of 252 genes, 92 (36%) peripheral, 96 (38%) central, 2 (1%) radial, and 10 (4%) punctate genes are opposite-directionality, implying cell-type specific regulation (Supplemental Table 1). An additional class of genes have cell-type dependent, but not opposite-directionality patterning. These genes are significant in at least one cell-type, but insignificantly localized in at least one other cell-type and account for approximately 55% of genes in peripheral and central metrics, and 20% for the radial and punctate metrics across all datasets. Observation of cell-type specific regulation of subcellular localization generates testable hypotheses for follow-up experiments to validate causal links and possible functions between 3' UTR length and localization.

Subcellular RNA localization is enriched for correlations with 3' UTR length

Alternative 3' UTRs and splice isoforms are known to impact function of mRNA, including localization [4]. The presumed mechanism of isoform-specific localization involves differential ability of trans-acting factors to bind alternative isoforms due to changes in the sequence composition of an isoform, for example shortening of the UTR that changes the affinity of an RNA binding protein to that transcript. Several hundred RNA Binding Protein (RBP), have been identified as controlling transcript localization, and many are likely or have already been shown to have cell-type specific expression, including at the isoform level [32]. Examples of such RBPs include members of the RNA-transport granule [33] , providing a model for why RNAs may be cell type specifically localized as a function of their isoform. RNA binding proteins can localize RNAs through trans-interactions, for example by miRNA binding and cell-type specific differential localization of

a gene may be independent of its isoform, due to differential expression of a trans-acting factor responsible for localization.

We coupled a recent statistical method to measure 3' UTR length called the ReadZS [8] with SPRAWL to identify genes with spatial localization correlated with 3' UTR regulation [34]. We used ReadZS to statistically quantify 3' UTR lengths at single-cell resolution, and then computed the median ReadZS score by cell-type and gene on cell-type-matched 10Xv3 scRNAseq datasets from the BICCN consortium [35]. Spatial localization SPRAWL scores and ReadZS 3' UTR lengths were correlated by gene/cell-type (Figure 3a). Twenty-six genes were detected as having significant SPRAWL/ReadZS correlation after BH multiple hypothesis correction at an FDR level of .05, a 2-fold enrichment compared to what is expected by chance (see Methods: Correlation analysis between SPRAWL and ReadZS). No significant gene/metric pairs were detected from the CZB mouse kidney/liver dataset which was the only other dataset with matched scRNAseq, which could be explained by this dataset having fewer and coarser cell-type definitions.

Timp3 3' UTR length predicts peripheral localization

In the BICCN data, *Timp3* has the largest observed variation in estimated 3' UTR length between cell-types (Methods), with most divergent read-buildup between layer-6 inferior temporal (L6 IT) and somatostatin-expressing (Sst) neurons reflecting at least two dominant 3' UTRs differing in length by > 2 kilobases (Figure 3b). These 3' UTR read densities were consistent across mouse biological sequencing replicates within 10X scRNAseq experiments. Only one UTR is annotated, though a gene antisense to *Timp3*, *Sync3* on the minus strand, overlaps its transcriptional radius. Observed reads can be confidently attributed to *Timp3* as *Sync3's* nearest exon is ~5kb from *Timp3's* UTR and plus-strand mapping reads alone were analyzed.

Timp3 protein is a secreted metalloprotease inhibitor that has been implicated in multiple diseases ranging from cardiopathies to macular dystrophies [36,37], but subcellular RNA localization patterns have not been reported. Elevated timp3 gene expression [38] is linked to compromised cerebral blood flow, and the RNA is experimentally validated to be a target of microRNA (miRNA) regulation [39]. *Timp3* is significantly peripheral in L6 IT neurons; while being insignificantly peripherally localized in Sst cells. SPRAWL and ReadZS 3' UTR scores had a significant negative correlation of R=-0.68 and p << 0.05 Pearson BH-corrected pval (Figure 3c). *Timp3's* longer, annotated 3' UTR isoform is significantly less peripherally localized as compared to shorter isoforms which are unannotated (Figure 3d).

We studied whether *Timp3's* 3' UTR was more globally regulated in endothelial and other cell-types through scRNAseq and in different biological contexts in both mouse and human cells. We extended the analysis to *Timp2* because Mouse and human *Timp3* and *Timp2* all have high sequence similarity (Figure 3f). ReadZS also detected statistically significant *Timp3* 3' UTR length shifts between cell-types from the Tabula Sapiens consortium [40] in the lung and other tissues (Supplemental Figure 3a). Further, we found both *Timp2* and *Timp3* UTR length to be regulated in lung tissue slices of endothelial cell-types (scRNAseq and ReadZS, Figure 3e, dotted red line). *Timp2* and *Timp3* 3' UTRs become shorter, during culture. Since SPRAWL identified a highly negative correlation between *Timp3* peripheral subcellular localization and 3' UTR length, the subcellular localization of *Timp3* is predicted to shift to a more peripheral distribution. In conjunction with 3' UTR length shortening, gene expression of *Timp3* decreases over this time course in all cell-type groups (Supplemental Figure 3b).

Both *Timp2* and *Timp3* show high conservation within their UTRs. Conservation is particularly high near the two dominant alternative 3' UTR regions, all but one of which are un-annotated (Figure 3b last track). These regions could contain alternative end processing or regulatory sequences, and include annotated binding sites for miR-181c-5p and miR-221-3p and RBPs Cirbp, Cpsf6, and Celf1. The isoforms differentially include these regions, releasing the shorter isoforms from regulatory pressures by these elements, including the experimentally validated miR-21 that binds in the 3' UTR of human *Timp3* [41]. In this study, high expression of miR-21 led to repression of *Timp3* and pathogenic activation of angiogenesis.

Together, we hypothesize that TIMP3 may have both secreted and non-secreted isoforms, with a precedent set by *Cd47*; and further, that this regulation is controlled by alternative 3' UTR isoform lengths that determine subcellular RNA localization through interaction with RBPs and microRNAs that specifically bind the longer isoform. This example illustrates the power of SPRAWL for unsupervised discovery of subcellular localization and its integration with isoform-resolved, annotation-free analysis of scRNA-seq to generate testable biological hypotheses regarding isoform-specific regulation and function.

Slc32a1, Cxcl14, and *Nxph1* 3' UTR length predicts sub-cellular localization

Beyond *Timp3* discussed above, SPRAWL detects 36 unique genes and 84 pairs of gene/metric (peripheral, radial, punctate, or central) combinations as having significant correlations with 3' UTR length, detected by ReadZS, from

the BICCN dataset. From this list *Slc32a1*, *Cxcl14*, and *Nxph1* were chosen for discussion as representatives of the central, radial, and punctate metrics respectively. All have significant evidence for cell-type differential expression of un-annotated 3' UTRs and an unusually high degree of 3' UTR conservation. Figure 4 depicts the SPRAWL scores and predicted 3' UTR lengths for *Slc32a1*, *Cxcl14*, and *Nxph1* in multiple cell-types. Representative low and high scoring cells for each gene/cell-type pair were chosen randomly after filtering for SPRAWL scores less than -0.2 and greater than 0.2 respectively, having 5 or more RNA spots of the gene of interest.

Slc32a1, synonymously *VIAAT* or *VGAT*, is a marker of GABAergic neurons and was found to be differentially central by cell-type (Figure 4a). *Slc32a1* is an integral membrane protein residing in synaptic vesicles where it uptakes glycine and gamma-aminobutyric acid (GABA) [42]. *Slc32a1* is currently annotated to have 2 exons in the UCSC genome browser mm39 [43], but was at one point thought to have 3 exons and exhibit alternative splicing near the 3' UTR without known biological significance [44]. SPRAWL central score and ReadZS have significant correlation (Pearson R=-0.94, corrected $p << 0.05$). Differential central localization of *Slc32a1* RNA between cell-types is of potential interest due to the protein product's known role of localizing to synaptic vesicles in neurons which would yield the highly non-central distribution observed in the L6 CT and L5 IT neurons.

Cxcl14 3' UTR length and SPRAWL radial score were significantly correlated (Pearson R=0.9 corrected $p << 0.05$); cell-types with longer 3' UTRs have radially-clustered RNA subcellular localization, while the unannotated shorter 3' UTRs have middling SPRAWL radial scores near zero. Only one Cxcl14 3' UTR isoform is annotated, but ReadZS analysis predicts a shift in length of about 600 bps (Figure 4b) [45]. The protein product of *Chemokine (C-X-C motif) ligand 14*, Cxcl14 or BRAK, is a small chemokine of length 99 residues in mouse and 111 in human and was originally found to be highly expressed in breast and kidney [46]. Cxcl14 is constitutively expressed in skin and keratinocytes and is a potent leukocyte recruitment factor [47] but has also more recently been observed as constitutively expressed throughout multiple brain regions where one of its functions is to regulate synaptic transmission [48]. According to the MERFISH dataset, Cxcl14 was lowly but consistently expressed with the full-length 3' annotated UTR in 429 L6 neurons with a median of 5 spots per cell while having higher expression in Vip-cells and astrocytes where a slightly shorter 3' UTR was expressed. We hypothesize that *Cxcl14* has differential 3' UTR usage associated with differential expression across these cell-types and that the novel short 3' UTR is less radially clustered than the annotated long 3' UTR.

Nxph1, neurexophilin-1, is a ligand of α-neurexin (α-*Nrxn*) and is expressed in inhibitory neurons [49]. The radial SPRAWL score of *Nxph1* is positively correlated with 3' UTR length (Pearson R=0.9, corrected $p << 0.05$ Figure 4c). *Nxph1* is a secreted protein that binds to multiple splice isoforms of α-*Nrxn* at synapses with varying specificity [50]. To our knowledge, neither differential 3' UTR lengths nor differential subcellular localization patterns have been previously described for *Nxph1* but this finding follows manual guided studies linking UTRs to dendritic targeting [51].

All three genes, *Slc32a1*, *Cxcl14,* and *Nxph1*, have predicted miRNA binding sites tiling their 3' UTRs suggesting possible mechanisms of differential 3' UTR post-transcriptional selection and regulation (Supplemental Figure 4a). We show an additional three genes with correlated spatial and 3' UTR length show similar patterns (Supplemental Figure 4b).

Conclusions

SPRAWL is a novel non-parametric approach for unbiased detection of subcellular RNA localization patterns in highly multiplexed imaging datasets, which are becoming increasingly available but remain focused on gene expression and cell-type localization in tissue. In mice and humans, > 96% of genes are alternatively spliced and UTR regulation is pervasive but poorly annotated [52,53]. In this study, SPRAWL identifies (1) cell-type-specific RNA localization as highly consistent across replicates, (2) genes with differential localization depending on cell-type (3) genes with differential localization correlated with estimated 3' UTR length, and (4) novel predicted UTRs. SPRAWL prioritizes functionally important isoform expression for further study such as *Timp3, Slc32a1, Cxcl14*, *and Nxph1* which have significant spatial and 3' UTR-usage correlation between cell-types. SPRAWL generates testable hypotheses of cis-regulatory elements that alter RNA localization. Trans-spatial events can be detected in the future by combining or applying SPRAWL to single-cell resolved protein abundances, such as CODEX [54] or MIBI [55] datasets. The localization scores generated by SPRAWL are versatile and can be computed for proteins rather than RNA. Furthermore, the SPRAWL framework can be used to implement different measures of subcellular localization. Some but not all statistically significant patterns detected by SPRAWL are "striking to the human eye," which has implications for whether human-guided or statistical-guided inferences are preferred.

Though the importance of correlation between SPRAWL subcellular localization and isoform expression, including *Timp3*, *Slc32a1, Cxcl14,* and *Nxph1*, was not explored in this work, we hypothesize a causal link between 3' UTR regulation, localization and potential protein function, as was observed for Actb, which could guide future

experimental efforts, as well as help pinpoint cell-type specific functions Sampling a handful of tissues and cell types, SPRAWL identified tens of RNA species with subcellular localization related to cell type. Many technical limitations suggest that this number is a significant underestimate: for one, MERFISH based approaches require probes to be pre-specified, and thus they (a) aggregate isoforms, missing cases where two co-expressed isoforms have dramatically different localization patterns; (b) miss isoforms that lack sequence contained in the probe set measurements. Further, single cell sequencing technology and analysis may be under-ascertaining RNA expression due to (i) sampling depth; (ii) poly-A capture bias and (iii) computational algorithms to analyze isoform-specific differences. Through the ReadZS we have collapsed UTR variation to a single scalar value [8,52,56] but we have not explored correlations with RNA splicing or other sequence variants, a topic of further research. Our findings support a model where 3' UTR regulation at the nucleotide level controls localization through function. If this is true, imaging-based technology like MERFISH will have limited power over discovery and in situ sequencing may be a preferred approach. Together, this suggests that isoform-specific localization may be widespread and confer functions that should be tested in future computational and experimental work.

SPRAWL provides an estimate of the pervasiveness of cell-type and 3' UTR regulated RNA localization. Limitations of the study include possible confounding by technical artifacts from probe hybridization, improper cell-segmentation, and bias in the gene panel selected for imaging. Additionally, our decision not to use nuclei boundaries blinds us to situations where an RNA may be highly peripheral, but still within the cell nucleus. This could mean UTR peripherality is confounded with dynamis of export, including transcription at the nuclear periphery. We have attempted to address these potential artifacts through hundreds of thousands of observations and by permutation where possible. Additionally, computationally shrinking cell-boundaries resulted in only minimal changes in SPRAWL scores. Future work on novel datasets using different segmentation approaches will provide further confidence that SPRAWL detects biologically relevant patterns. We believe the current implementation of SPRAWL is conservative and likely misses patterns due to optical crowding and low-abundance gene expression.

Acknowledgements

We'd like to acknowledge the Salzman lab members for helpful discussion and suggestions, especially Elisabeth Meyer, Roozbeh Dehghannasiri, and Tavor Baharav for text edits as well as Jonathan Liu from the Chan-Zuckerberg Biohub. We acknowledge George Emmanuel for help in initial data download and processing. We acknowledge Pehr Harbury and Mark Krasnow for feedback and Mark Krasnow and Catherine Blish for the human lung ex-situ culture time-course datasets. Some of the computing for this project was performed on the Sherlock cluster. 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. We would like to thank funding sources from the NCI (5F31CA243170-02), the NHGRI (1R56HG011231-01) and NIGMS (1R35GM139517-01).

Supplement

Methods

SPRAWL input data and preprocessing

SPRAWL takes as input processed datasets from MERFISH, Vizgen, and SeqFISH+ requiring cell-boundary and RNA spot x,y and gene label information. For MERFISH and Vizgen, this data is the product of applying MERlin [57] on the raw MERFISH microscopy images to align the images between sequencing rounds, call RNA spots, and perform cell segmentation using a seeded watershed approach described in a prior MERFISH work [58]. SeqFISH+ utilizes a similar approach to identify and decode RNA spots, but then simply defines the cell boundary as the convex hull around all points [12].

The MERFISH primary mouse cortex (MOp) dataset has 258 genes from coronal slices of the MOp from two mice as biological replicates [59]. Each mouse had 6 MERFISH experiments with 5-6 10 um sections processed together on the same coverslip. Each mouse had 32 total sections. Each 10 um thick section had 7 optical layers spaced 1.5 microns apart. The MERFISH brain MOp processed datasets include multiple z-slices for each cell. The data was downloaded from https://download.brainimagelibrary.org/cf/1c/cf1c1a431ef8d021/processed_data/

The SeqFISH+ dataset imaged 913 cells and 10,000 genes in the mouse cortex at a single z-slice [12]. The authors assigned each cell to one of twenty-six different cell-type annotations such as Endothelial, Interneuron, Astrocyte, etc. The dataset was downloaded from https://github.com/CaiGroup/seqFISH-PLUS/blob/master/sourcedata.zip?raw=true

The Vizgen MERFISH Mouse Brain Map (BrainMap) is a dataset of 649 total genes which include canonical brain cell type markers, GPCRs, and RTKs from a single mouse brain [25]. Three full coronal sections were processed along the rostral-caudal axis. Additionally, for each section, three adjacent slices were used as biological replicates with the underlying assumption that adjacent slices in the mouse brain have high similarities in cell-type composition and spatial organization. Each of the nine imaging datasets contain seven optical layers spaced 1.5 microns apart. Data is publicly available https://console.cloud.google.com/marketplace/product/gcp-public-data-vizgen/vizgen-mouse-brain-map

The Vizgen MERFISH Liver showcase contained 2 mouse liver samples each with two MERFISH experiments imaging 347 genes and over one million cells [25]. Cell-type annotations were not provided and instead cell-type proxies were determined by clustering the cells based on the MERFISH-determined RNA composition of each cell (Methods: Vizgen Brainmap and Liver showcase clustering to produce cell-type proxies). The dataset contains seven optical layers spaced 1.5 microns apart and data is publicly available from
https://info.vizgen.com/mouse-liver-data?submissionGuid=832a9f61-22d3-44c1-a2cf-838c166d9ac5

The CZB kidney/liver dataset contained a single mouse kidney and liver sample that were imaged using the Vizgen platform to detect the same panel of 307 genes in ~57,000 cells in the kidney and ~16,000 in the liver [60].
https://figshare.com/projects/MERFISH_mouse_comparison_study/134213

We have developed a simple HDF5 format to standardize the different data sources. In brief, data is stored in a cell-centric manner, consolidating RNA spots and cell boundaries into the same object. This flexible format is described in detail in the github repository https://github.com/r-bierman/SPRAWL and includes vignettes with example datasets. For MERFISH and Vizgen datasets, the RNA spots and cell boundaries were assigned locations in a global coordinate, but lacked cell assignments for each RNA spot. We have written simple and fast scripts to make these assignments using the python Rtree and shapely [61] packages. The github repository includes the nextflow pipelines used to transform the downloaded datasets to this HDF5 format.

SPRAWL methodology

SPRAWL preprocesses spatial datasets into a standardized HDF5 file that contains cell boundary, cell-type, and RNA location information generated from MERFISH/Vizgen and SeqFISH+ datasets (Figure 1a). Next per-gene/per-cells are calculated. For the peripheral metric, all RNA spots are ranked based on their minimum distance to the cell boundary (Figure 1b), then their means are used to generate a gene/cell-type score and p-value (Figure 1c). Scores near 1 indicate a gene is highly peripheral in a cell-type, while scores near -1 indicate a pattern of RNA molecules far from the cell-boundary. Intuitively, if a gene is not significant it will not be close or far from the cell boundary and its peripheral score will be near 0, and its p-value will be insignificant. The centrality metric is conceptually similar, where ranking is determined by minimum distance to the cell centroid and positive values indicate unexpectedly centrally-biased distributions. Empirically, the centrality and peripherality metrics are anti-correlated (Supplemental Figure 1b), but not perfectly, as it is possible for an RNA spot to be simultaneously close to the periphery and cell centroid with certain cell shapes such as a "dumbbell". Only the ranking step is different between the peripheral and central metrics; all downstream steps are identical.

Under the null hypothesis that a gene is not subcellularly patterned within a cell, the peripheral and central gene/cell scores have an expected value of 0 and a calculable variance that depends on the number of RNA spots. These statistical underpinnings of the gene/cell scores allow for identification of spatially significant patterning within gene/cell-types (Methods). Under the null which is each spot's gene identity is drawn uniformly from the set of gene/spots observed from the cell, gene/**cell** scores for k cells of a single cell-type and gene g are independent random variables $X_{g,1}$, $X_{g,2}$, $X_{g,3}$, ..., $X_{g,k}$ with expected values, $\mu_i = 0$ and variance σ_i. Independence in this case comes from the assumption that the scores of a given gene across different cells do not influence each other. Note that the scores of different genes within the same cell are not independent due to the ranking procedure. We define $Y = \text{mean}(X_{g,1}, X_{g,2}, X_{g,3}, ..., X_{g,k})$ as the SPRAWL gene/**cell-type** score and a z-score can be calculated under the null that within a cell, each spot's gene identity is exchangeable with the Lyapunov Central Limit Theorem [62] (Methods: SPRAWL gene/cell-type scoring).

The resulting values y are used to calculate two-sided p-values using the CDF of the standard normal. Multiple hypothesis testing from the numerous gene/cell-type pairs is controlled using the Benjamini-Hochberg correction [63].

SPRAWL peripheral and central metric definition

Each gene-cell pair is assigned a SPRAWL score by (1) ranking all RNA spots, (2) calculating median ranks per gene, and (3) normalizing by the expected median rank. Consider a single cell, with a single z-slice, that has n total RNA spots, and g unique genes with each gene having $m_1, m_2, m_3, \dots, m_g$ spots such that $\sum_i^g m_i = n$.

For the peripheral metric, let $d_1, d_2, d_3, \dots, d_n$ represent the minimum euclidean distance to the periphery of each RNA spot, for the central metric these distances are instead measured from the cell centroid. Each spot is assigned a rank from 1 to n such that the spot with rank 1 is $argmin(d_1, d_2, d_3, \dots, d_n)$ and the spot with rank n is $argmax(d_1, d_2, d_3, \dots, d_n)$ randomly breaking ties where needed.

The ranks are then grouped by gene to calculate the median ranks $t_1, t_2, t_3, \dots, t_g$. The peripheral/central SPRAWL gene/cell score x_i for $1 \leq i \leq g$, is the median rank t_i normalized by the expected median rank t_e which is $(n + 1)/2$ for all genes independent of m_i:

$$x_i = \frac{t_e - t_i}{t_e - 1}$$

Note that $-1 \leq x_i \leq 1$ since $min(t_i) = 1$ yields $x_i = 1$, and $max(t_i) = n$ yields $x_i = -1$.

To generalize the definition of the peripheral/central SPRAWL score in the case that a cell has multiple z-slices with a unique cell-boundary and set of spots for each, the distances d_i are calculated from each RNA spot to the cell-boundary/centroid in the same z-slice, then the ranks are assigned across all z-slices.

SPRAWL radial metric definition

The radial SPRAWL score is assigned to each gene-cell pair by performing gene-label swapping bootstrapping iterations and measures tendency of genes to be in one sector of a cell or to be radially dispersed.

Consider a single cell, with a single z-slice, that has n total RNA spots, and g unique genes with each gene having $m_1, m_2, m_3, \dots, m_g$ spots such that $\sum_i^g m_i = n$. We restrict to $m_i > 2$ since genes with a single spot do not conceptually have a radial bias.

Before permuting the gene labels, we randomly select a pair of RNA spots for each gene and measure the angle between them with respect to the cell-boundary centroid. Let θ represent the minimum angle formed by the three points of the location of RNA spot 1 (x_1, y_1), the cell centroid (x_c, y_c), and RNA spot 2 (x_2, y_2). The cell centroid (x_c, y_c) is approximated as the mean of all vertices in the cell boundary polygon. This process is repeated 10 times and averaged to calculate the mean observed angle of each gene.

The same process is repeated after randomly swapping gene labels but keeping the RNA spot locations the same. We perform 1000 bootstrap iterations. These mean permuted angles serve as the null distribution of mean angles which are used in conjunction with the mean observed angle to calculate both mean and variance. In the case that a cell has multiple z-slices, the mean cell centroid over all slices is used to calculate pairwise angles without regard to z-slice.

SPRAWL punctate score definition

The punctate SPRAWL score is conceptually identical to the radial score and also relies on bootstrapping. The punctate score is assigned to each gene-cell pair measuring euclidean distances instead of angles between randomly selected gene pairs. The null distribution is created using the same process as the radial score. In the case that a cell has multiple

z-slices, the scoring is performed by projecting all points onto the same (x,y) plane before measuring euclidean distances. This simplification can be readily replaced with true 3D pairwise distances.

Theoretical features of the SPRAWL peripheral score

While the punctate and radial metrics are calculated using bootstrapping and estimated statistics, the SPRAWL peripheral and central metrics have known properties under the null hypothesis that the gene of interest is not spatially regulated in the given cell. Under this null, the ranks of the gene of interest are chosen with equal probability. In an alternate hypothesis such as a gene being peripherally localized in a cell, RNA spots of the gene of interest will have a skewed probability of being assigned lower ranks, closer to the cell boundary. Under the null hypothesis $E[X_i] = 0$, since $E[T_i] = t_e$ for gene $1 \leq i \leq g$.

$Var[X_i]$ depends on the total number of RNA spots in a cell n, and the number of spots of the gene i, m_i. For example, in the extreme case where $m_i = n$, every spot in a cell is the gene of interest and $Var[X_i] = 0$. $Var[X_i]$ for any gene can be calculated under the null by iterating over all possible values of $x \in X$ since X is a discrete R.V.

$$Var[X] = \sum_x P(T = t_e + x(1 - t_e)) \cup t_e - x(1 - t_e))$$

When there are an odd number of gene spots m, t is the $(m + 1)/2$ rank order statistic, and under the null hypothesis where the ranks are chosen uniformly, the probability of the r-th order statistic taking the value t equals:

$$P(T = t) = \frac{\binom{t-1}{r-1}\binom{n-t}{m-r}}{\binom{n}{m}}$$

Where n is the total number of RNA spots, m is the number of RNA spots for the gene of interest, and $r = (m + 1)/2$.

$$P(R_0 = r) = \frac{\binom{r-1}{\frac{m+1}{2}-1}\binom{n-r}{m-\frac{m+1}{2}}}{\binom{n}{m}}$$

Calculating $Var[X]$ when m is even-values requires significantly more calculation. We still need to calculate $P(X^2 = x^2) = P(T = t_e + x(1 - t_e) \cup t_e - x(1 - t_e)$

But t is no longer an order statistic and does not have a closed form calculation. Instead t is the average of the "left of center" $(\frac{m}{2})$-th order-statistic L, and the "right of center" $(\frac{m}{2} + 1)$-th order statistic R. Then for a given x and corresponding t:

$$P(T = t) = P(\frac{L+R}{2} = t)$$

We calculate $P(T = t)$ by summing the probabilities of observing all possible pairs of L and R that sum to $2t$. We can think of starting L and R as close to t as possible, and then "walking" L and R away from t one rank at a time in lockstep summing over all i's such that $1 \leq t - 1 \leq n$ where n is the total number of spots in the cell:

$$P(\frac{L+R}{2} = t) = \sum_i P(L = t - i \cap R = t + i) = \sum_i P(R = t + 1 | L = t - 1) P(L = t - 1)$$

Omitted for clarity, the ceiling of $t - 1$ is taken and the floor of $t + 1$ above to account for non-integer t.

$P(R = r | L = l)$ has an intuitive interpretation that simplifies to an order statistic probability. Since we observe $L = l$ we know that the R-th order statistic must be one of the ranks between $l + 1$ and n inclusively. We can renumber these ranks to be between 1 and $n - l + 1$ and we are interested in the probability that the 1-st order statistic takes the value $r - l$ in the renumbering. This has the same closed form solution as described in the odd-valued m case.

Computing $Var[X]$ for even-valued m is $O\left(n^2\right)$ since we have to iterate over all possible medians, and then for each median we have to "walk" L and R outwards which is itself $O(n)$. In comparison, the computation of $Var[X]$ for odd-valued m is $O(n)$. Through various optimizations, multiprocessing, and caching, SPRAWL calculated $Var[X]$ in under an hour for all processed samples.

SPRAWL gene/cell-type scoring

Consider a cell-type with k cells with non-zero counts of a gene of interest where each cell is assigned a SPRAWL score $X_1, X_2, \ldots, X_k$. Note that the X_i are not i.i.d. due to having different $Var[X_i]$ resulting from different values of m and n as described above.

However, we do make the assumption that the X_i are independent, which has the biological interpretation that the localization of the gene of interest in one cell does not depend on its localization in another cell. Under this assumption we utilize the Lyapunov Central Limit Theorem [62] to estimate that

$$\lim_{k \to \infty} \frac{1}{\sqrt{\sum_{i=1}^{k} \sigma_i^2}} \sum_{i=1}^{k} (X_i - \mu_i) \to N(0,1) \text{ in distribution}$$

We approximate values of y for each gene/cell-type using the observed x_i and theoretical mean and variance whose calculation is described above. These y are used to calculate two-sided p-values from the CDF of the standard normal.

Multiple hypothesis testing over all gene/cell-type pairs is controlled using the Benjamini Hochberg correction [63] at a significance level of $\alpha = 0.05$.

We calculate the effect size for each gene/cell-type as the mean gene/cell score $\sum_{i}^{k} \frac{x_i}{k}$

SPRAWL is highly specific in identifying genes with subcellular patterning conditional on cell boundaries

If the gene labels of RNA spots within cells of real datasets are permuted to remove any underlying spatial patterning (Methods), none of the metrics detect significant gene/cell-type patterning after Benjamini Hochberg (BH) multiple hypothesis correction with an FDR of 0.05 for any of the four datasets tested [63]. All metrics were observed to produce uniform p-values under this null dataset regardless of the number of cells per cell-type, as indicated by theory. The median score per gene/cell-type is dependent on the number of cells, with larger groups having median scores closer to zero (Supplemental Figure 1). The lack of any false positive calls under the permuted null is consistent with at an FDR of 0.05.

SPRAWL Filtering

For all datasets sparse cells and cell-types were filtered out by removing cells with fewer than 10 unique genes and/or fewer than 200 unique RNA spots. Gene/cell-type pairs with fewer than 20 cells were removed from consideration. Further filtering for the radial and punctate metrics requires removal of genes from cells that have only a single RNA spot. These spots are removed and then remaining spots can still be scored in this cell for other genes. All filtering steps are implemented as user-accessible parameters and have made SPRAWL more conservative, increasing the confidence of positive hits, but reducing the power to detect real localization differences that occur for lowly expressed genes and/or rare cell-types.

ReadZS usage and modifications

The ReadZS [8] detects read buildup differences between cell-types from single-cell RNA-seq datasets in an annotation independent manner using equal sized windows tiling the genome. We modified the ReadZS to analyze at the 3' UTR-level of just the ~250 genes imaged in the BICCN MOp dataset. The 10X scRNAseq data was processed individually for the 4 different mouse donors while the SS2 cells across 45 donors were processed as a single sample due to limited cell counts per mouse.

Correlation analysis between SPRAWL and ReadZS for MERFISH MOp datasets
For a given SPRAWL gene and spatial metric, the median ReadZS score of that gene for each cell-type was correlated against the median SPRAWL score over the same cell-types. For positive-strand genes, larger ReadZS score indicates longer 3' UTR isoforms, and vice versa for negative-strand genes. A proxy for 3' UTR length was defined as the distance between the annotated start of the 3' UTR and the RNA mapping position. The span in estimated 3' UTR lengths was measured as the difference between the longest and shortest median cell-type 3' UTR proxy lengths.

Vizgen Brainmap and Liver showcase clustering to produce cell-type proxies
Neither the Vizgen MERFISH Mouse Brain Map nor Liver showcase datasets provided cell-type annotations. We decided to roughly cluster the cells into groups to serve as a proxy for cell-type. The Leiden clustering method was used to find well-connected clusters in all of the filtered 90% highest spot-count cells using Scanpy python package [29]. First each dataset was normalized so that each cell had 10,000 spots, then the top 40 principal components were used to build the neighborhood graph with 10 neighbors and perform the Leiden clustering. This resulted in 22 clusters for the Brainmap dataset and 100 clusters for the Liver dataset. The fraction of cells in each cluster was consistent across biological replicates for the Vizgen Liver (Supplemental Figure 2) and Vizgen Brainmap (data not shown) indicating that cells were primarily clustering by type, and not by batch. To estimate the batch effect, we calculated the probability that two cells originated from the same biological replicate given that they were in the same cluster and compared this to the overall probability that two cells are from the same biological replicate. All clusters were within 0.05 of the overall probability of two cells sharing a batch.

Simulations to benchmark SPRAWL sensitivity and specificity
Null simulated datasets were created from the MERFISH BICCN spatial dataset by randomly permuting the RNA-spot gene labels within each cell across the entire dataset. The cell-boundaries, RNA-spot counts, and RNA (x,y,z) coordinates were preserved in the null dataset.

Identification of RBP and miRNA binding to Timp3 3' UTR
The RNAInter v4.0 RNA interactome repository was used to search for RBPs and miRNAs with experimental evidence of binding in the 3' UTR of the Mus musculus *Timp3*, *Slc32a1*, *Cxcl14*, and *Nxph1* genes [64]. Target regions for RBPs were taken from RNAInter, while miRNA binding sites were generated and cross-checked against TargetScan release 8.0 [65] and miRWalk [66]. Only miRNAs shared by RNAInter, TargetScan, and miRWalk results with experimental evidence were considered.

RNAs with signal peptides do not have significantly central or peripheral localization
We hypothesized that RNAs encoding a signal recognition peptide (SRP) for translation on the rough endoplasmic reticulum would be nuclear localized and would therefore be more centrally localized than genes without signal peptides. We predicted the presence of SRPs using DeepSig [67] with protein sequences downloaded from Gencode release M28 protein coding transcripts fasta for all genes present across the MOp, Vizgen Brainmap, and SeqFISH+ cortex datasets. For genes with multiple protein isoforms, the longest isoform was selected for SRP prediction. In all datasets the per-gene per-cell peripheral and central scores were not significantly different according to a Kolmogorov Smirnov test (Supplemental Figure 5a).

Genes enriched in single-nucleus RNAseq are marginally correlated with periphery score
We tested whether nuclear-localizing genes would be assigned higher SPRAWL central periphery scores utilizing both the 10X single-cell RNAseq (scRNAseq) as well as 10X single-nucleus RNAseq (snRNAseq) from the BICCN consortium (BRAIN Initiative Cell Census Network (BICCN) 2021). The single-cell sequencing data was first normalized to the number of counts per gene per cell per one million (TPM) reads for both the cell and nuclear datasets. The median gene/cell-type TPM for both sequencing datasets was determined, and the nuclear-fraction score was determined to be snRNAseq-TPM / (snRNAseq-TPM + scRNAseq-TPM). The median periphery score per gene/cell-type was correlated against the median snRNAseq-TPM, scRNAseq-TPM , and nuclear-fraction. In all comparisons, the correlation coefficients were small in magnitude, but were significantly positive for the snRNAseq, indicating a link between X tendency and peripherality, and significantly negative in the nuclear-fraction analyses, indicating a link between the gene's enrichment in nuclear reads and its distance from the cell periphery. The small effect size was detectable due to the approximately 8,000 gene/cell-type data points and provides weak support for the hypothesis. We investigated which genes, if any, are differentially nuclear-enriched across cell-types by sequencing and concordantly by peripheral score and

discovered *Wipf3* (Supplemental Figure 5b) and *Slc30a3*, which were highly negatively correlated with mean Pearson correlation coefficients of -0.86 and -0.93 across MERFISH MOp samples. Surprisingly, *Satb2* was also discovered to be significant, but had a highly positive mean Pearson correlation coefficient of 0.95. All genes were determined to be significant after Benjamini Hochberg multiple hypothesis correction.

Bibliography

1. Hung M-C, Link W. Protein localization in disease and therapy. *J. Cell Sci.* 2011;124(Pt 20):3381–3392.
2. Lawrence JB, Singer RH. Intracellular localization of messenger RNAs for cytoskeletal proteins. *Cell*. 1986;45(3):407–415.
3. Kislauskis EH, Zhu X, Singer RH. beta-Actin messenger RNA localization and protein synthesis augment cell motility. *J. Cell Biol.* 1997;136(6):1263–1270.
4. Kislauskis EH, Zhu X, Singer RH. Sequences responsible for intracellular localization of beta-actin messenger RNA also affect cell phenotype. *J. Cell Biol.* 1994;127(2):441–451.
5. Lécuyer E, Yoshida H, Parthasarathy N, et al. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*. 2007;131(1):174–187.
6. Saka HA, Valdivia R. Emerging roles for lipid droplets in immunity and host-pathogen interactions. *Annu. Rev. Cell Dev. Biol.* 2012;28:411–437.
7. Ma W, Mayr C. A Membraneless Organelle Associated with the Endoplasmic Reticulum Enables 3'UTR-Mediated Protein-Protein Interactions. *Cell*. 2018;175(6):1492-1506.e19.
8. Meyer E, Dehghannasiri R, Chaung K, Salzman J. ReadZS detects developmentally regulated RNA processing programs in single cell RNA-seq and defines subpopulations independent of gene expression. *BioRxiv*. 2021;
9. Berkovits BD, Mayr C. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature*. 2015;522(7556):363–367.
10. Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. *Science*. 2015;347(6220):1260419.
11. Moffitt JR, Hao J, Wang G, et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc Natl Acad Sci USA*. 2016;113(39):11046–11051.
12. Eng C-HL, Lawson M, Zhu Q, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature*. 2019;568(7751):235–239.
13. Stickels RR, Murray E, Kumar P, et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* 2021;39(3):313–319.
14. Su G, Qin X, Enninful A, et al. Spatial multi-omics sequencing for fixed tissue via DBiT-seq. *STAR Protocols*. 2021;2(2):100532.
15. Ståhl PL, Salmén F, Vickovic S, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. 2016;353(6294):78–82.
16. Holt CE, Bullock SL. Subcellular mRNA localization in animal cells and why it matters. *Science*. 2009;326(5957):1212–1216.
17. Tang Q, Nie F, Kang J, Chen W. mRNALocater: enhance the prediction accuracy of eukaryotic mRNA subcellular localization by using model fusion strategy. *Mol. Ther.* 2021;
18. Samacoits A, Chouaib R, Safieddine A, et al. A computational framework to study sub-cellular RNA localization. *Nat. Commun.* 2018;9(1):4584.
19. Xue Z-Z, Wu Y, Gao Q-Z, Zhao L, Xu Y-Y. Automated classification of protein subcellular localization in immunohistochemistry images to reveal biomarkers in colon cancer. *BMC Bioinformatics*. 2020;21(1):398.
20. Moran PAP. Notes on continuous stochastic phenomena. *Biometrika*. 1950;37(1–2):17–23.
21. Xia C, Fan J, Emanuel G, Hao J, Zhuang X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc Natl Acad Sci USA*.

2019;116(39):19490–19499.
22. Fazal FM, Han S, Parker KR, et al. Atlas of Subcellular RNA Localization Revealed by APEX-Seq. *Cell*. 2019;178(2):473-490.e26.
23. Padrón A, Ingolia N. Analyzing the Composition and Organization of Ribonucleoprotein Complexes by APEX-Seq. *Methods Mol. Biol.* 2022;2428:277–289.
24. Zhang M, Eichhorn SW, Zingg B, et al. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature*. 2021;598(7879):137–143.
25. Vizgen. Data Release Program - Vizgen.
26. Yao Z, Liu H, Xie F, et al. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*. 2021;598(7879):103–110.
27. Di Tommaso P, Chatzou M, Floden EW, et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 2017;35(4):316–319.
28. Mah CK, Ahmed N, Lam D, et al. Bento: A toolkit for subcellular analysis of spatial transcriptomics data. *BioRxiv*. 2022;
29. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 2018;19(1):15.
30. Nitta T, Hata M, Gotoh S, et al. Size-selective loosening of the blood-brain barrier in claudin-5-deficient mice. *J. Cell Biol.* 2003;161(3):653–660.
31. Jin J, Sison K, Li C, et al. Soluble FLT1 binds lipid microdomains in podocytes to control cell morphology and glomerular barrier function. *Cell*. 2012;151(2):384–399.
32. Hentze MW, Castello A, Schwarzl T, Preiss T. A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* 2018;19(5):327–341.
33. Kanai Y, Dohmae N, Hirokawa N. Kinesin transports RNA: isolation and characterization of an RNA-transporting granule. *Neuron*. 2004;43(4):513–525.
34. Booeshaghi AS, Yao Z, van Velthoven C, et al. Isoform cell-type specificity in the mouse primary motor cortex. *Nature*. 2021;598(7879):195–199.
35. BRAIN Initiative Cell Census Network (BICCN). A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature*. 2021;598(7879):86–102.
36. Weber BH, Vogt G, Pruett RC, Stöhr H, Felbor U. Mutations in the tissue inhibitor of metalloproteinases-3 (TIMP3) in patients with Sorsby's fundus dystrophy. *Nat. Genet.* 1994;8(4):352–356.
37. Schrimpf C, Xin C, Campanholle G, et al. Pericyte TIMP3 and ADAMTS1 modulate vascular stability after kidney injury. *J. Am. Soc. Nephrol.* 2012;23(5):868–883.
38. Capone C, Dabertrand F, Baron-Menguy C, et al. Mechanistic insights into a TIMP3-sensitive pathway constitutively engaged in the regulation of cerebral hemodynamics. *eLife*. 2016;5:.
39. Fiorentino L, Cavalera M, Mavilio M, et al. Regulation of TIMP3 in diabetic nephropathy: a role for microRNAs. *Acta Diabetol.* 2013;50(6):965–969.
40. Tabula Sapiens Consortium*, Jones RC, Karkanias J, et al. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*. 2022;376(6594):eabl4896.
41. Hu J, Ni S, Cao Y, et al. The Angiogenic Effect of microRNA-21 Targeting TIMP3 through the Regulation of MMP2 and MMP9. *PLoS ONE*. 2016;11(2):e0149537.
42. Gasnier B. The SLC32 transporter, a key protein for the synaptic release of inhibitory amino acids. *Pflugers Arch.* 2004;447(5):756–759.
43. Lee BT, Barber GP, Benet-Pagès A, et al. The UCSC Genome Browser database: 2022 update. *Nucleic Acids Res.* 2022;50(D1):D1115–D1122.
44. Ebihara S, Obata K, Yanagawa Y. Mouse vesicular GABA transporter gene: genomic organization, transcriptional regulation and chromosomal localization. *Brain Res. Mol. Brain Res.* 2003;110(1):126–139.
45. Bässler EL, Ngo-Anh TJ, Geisler HS, Ruppersberg JP, Gründer S. Molecular and functional characterization of

acid-sensing ion channel (ASIC) 1b. *J. Biol. Chem.* 2001;276(36):33782–33787.
46. Hromas R, Broxmeyer HE, Kim C, et al. Cloning of BRAK, a novel divergent CXC chemokine preferentially expressed in normal versus malignant cells. *Biochem. Biophys. Res. Commun.* 1999;255(3):703–706.
47. Westrich JA, Vermeer DW, Colbert PL, Spanos WC, Pyeon D. The multifarious roles of the chemokine CXCL14 in cancer progression and immune responses. *Mol. Carcinog.* 2020;59(7):794–806.
48. Banisadr G, Bhattacharyya BJ, Belmadani A, et al. The chemokine BRAK/CXCL14 regulates synaptic transmission in the adult mouse dentate gyrus stem cell niche. *J. Neurochem.* 2011;119(6):1173–1182.
49. Born G, Breuer D, Wang S, et al. Modulation of synaptic function through the α-neurexin-specific ligand neurexophilin-1. *Proc Natl Acad Sci USA*. 2014;111(13):E1274-83.
50. Wilson SC, White KI, Zhou Q, et al. Structures of neurexophilin-neurexin complexes reveal a regulatory mechanism of alternative splicing. *EMBO J.* 2019;38(22):e101603.
51. Mayford M, Baranes D, Podsypanina K, Kandel ER. The 3'-untranslated region of CaMKII alpha is a cis-acting signal for the localization and translation of mRNA in dendrites. *Proc Natl Acad Sci USA*. 1996;93(23):13250–13255.
52. Olivieri JE, Dehghannasiri R, Salzman J. The SpliZ generalizes "percent spliced in" to reveal regulated splicing at single-cell resolution. *Nat. Methods*. 2022;19(3):307–310.
53. Olivieri JE, Dehghannasiri R, Wang PL, et al. RNA splicing programs define tissue compartments and cell types at single-cell resolution. *eLife*. 2021;10:.
54. Black S, Phillips D, Hickey JW, et al. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nat. Protoc.* 2021;16(8):3802–3835.
55. Keren L, Bosse M, Thompson S, et al. MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Sci. Adv.* 2019;5(10):eaax5851.
56. Chaung K, Baharav T, Zheludev I, Salzman J. A statistical, reference-free algorithm subsumes myriad problems in genome science and enables novel discovery. *BioRxiv*. 2022;
57. Emanuel G, Seichhorn, Babcock H, Leonardosepulveda, Timblosser. ZhuangLab/MERlin: MERlin v0.1.6. *Zenodo*. 2020;
58. Moffitt JR, Bambah-Mukku D, Eichhorn SW, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*. 2018;362(6416):.
59. Zhang M, Eichhorn SW, Zingg B, et al. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics. *BioRxiv*. 2020;
60. Liu J, Tran V, Pranathi Vemuri VN, et al. Comparative analysis of MERFISH spatial transcriptomics with bulk and single-cell RNA sequencing. *BioRxiv*. 2022;
61. toblerity.org SG. Shapely: manipulation and analysis of geometric objects. 2007.
62. Billingsley P. PROBABILITY AND MEASURE, 3RD EDITION (WILEY SERIES IN PROBABILITY AND MATHEMATICAL STATISTICS). WI; 1995.
63. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1995;57(1):289–300.
64. Kang J, Tang Q, He J, et al. RNAInter v4.0: RNA interactome repository with redefined confidence scoring system and improved accessibility. *Nucleic Acids Res.* 2022;50(D1):D326–D332.
65. McGeary SE, Lin KS, Shi CY, et al. The biochemical basis of microRNA targeting efficacy. *Science*. 2019;366(6472):.
66. Sticht C, De La Torre C, Parveen A, Gretz N. miRWalk: An online resource for prediction of microRNA binding sites. *PLoS ONE*. 2018;13(10):e0206239.
67. Savojardo C, Martelli PL, Fariselli P, Casadio R. DeepSig: deep learning improves signal peptide detection in proteins. *Bioinformatics*. 2018;34(10):1690–1696.

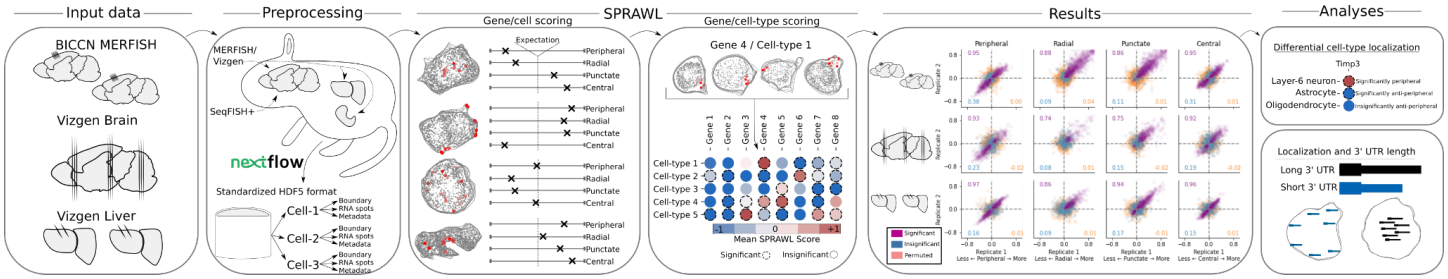


Figure 1: SPRAWL scores are highly consistent across biological replicate correlations. a) SPRAWL can process any dataset that includes RNA spot locations, cell boundaries, and cell-type or cell cluster information. MERFISH/Vizgen and SeqFISH+ datasets from mouse brain, liver, and kidney were each transformed into a standardized HDF5 format using custom nextflow pipelines. SPRAWL assigns scores between -1 and 1 for each gene in each cell across all z-planes based on simple and statistically tractable metrics (Methods). The mean gene ranks are used to compute the SPRAWL score per gene/cell-type and assigned p-values. Gene/cell-type scores are then used in a variety of downstream analyses. SPRAWL gene/cell-type scores are highly correlated between biological replicates in the BICCN MERFISH (Results: top), Vizgen Brain (Results: middle), and Vizgen Liver (Results: bottom) datasets analyzed in the study. Pearson correlation coefficients (purple for significant gene/cell-types called marginally, blue for insignificant gene/cell-types, and orange for the permuted null datasets). SPRAWL cell-type specific scores are highly reproducible as shown by strong correlation between independent biological replicates.

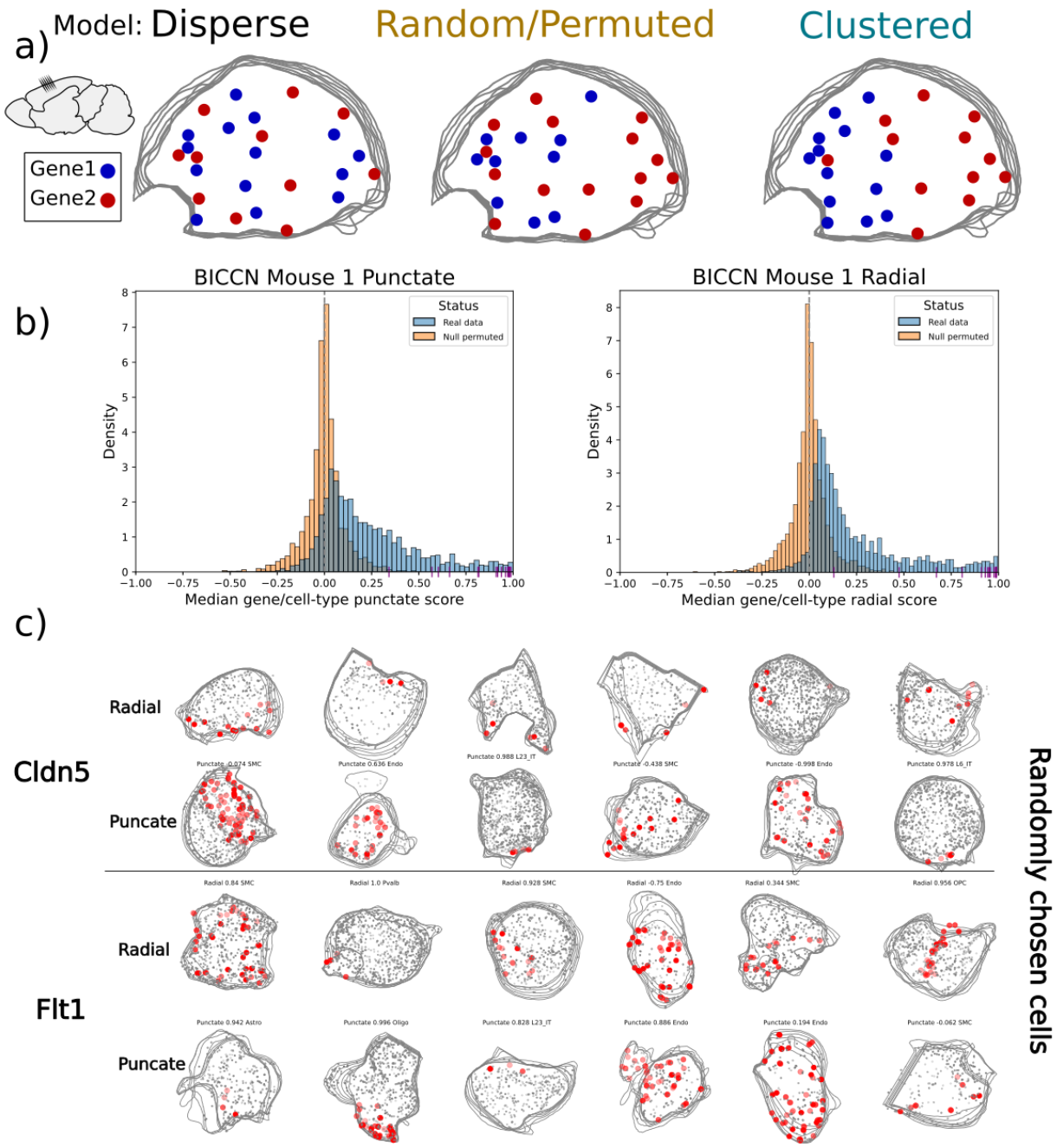


Figure 2: RNA spots of the same gene are enriched for radial and in punctate clustering.
a) Models of RNA radial and punctate distribution for two genes of interest in a cell ranging from high dispersion (left), random dispersion (center) to high clustering (right). The clustered model depicts a situation where both genes are assigned large positive SPRAWL punctate and radial scores. b) The distribution of radial and punctate scores of the null-permuted dataset (orange) is consistent with the expected random/permuted model, while the SPRAWL scores of real-data (blue) show an enrichment to positive scores suggesting a clustered regime. Purple ticks on the x-axis show the extremely punctate and radial median scores of gene Cldn5 for all cell-types. c) Visualizing randomly selected Cldn5-expressing cells, from any cell-type, depicts highly radial and punctate distributions.

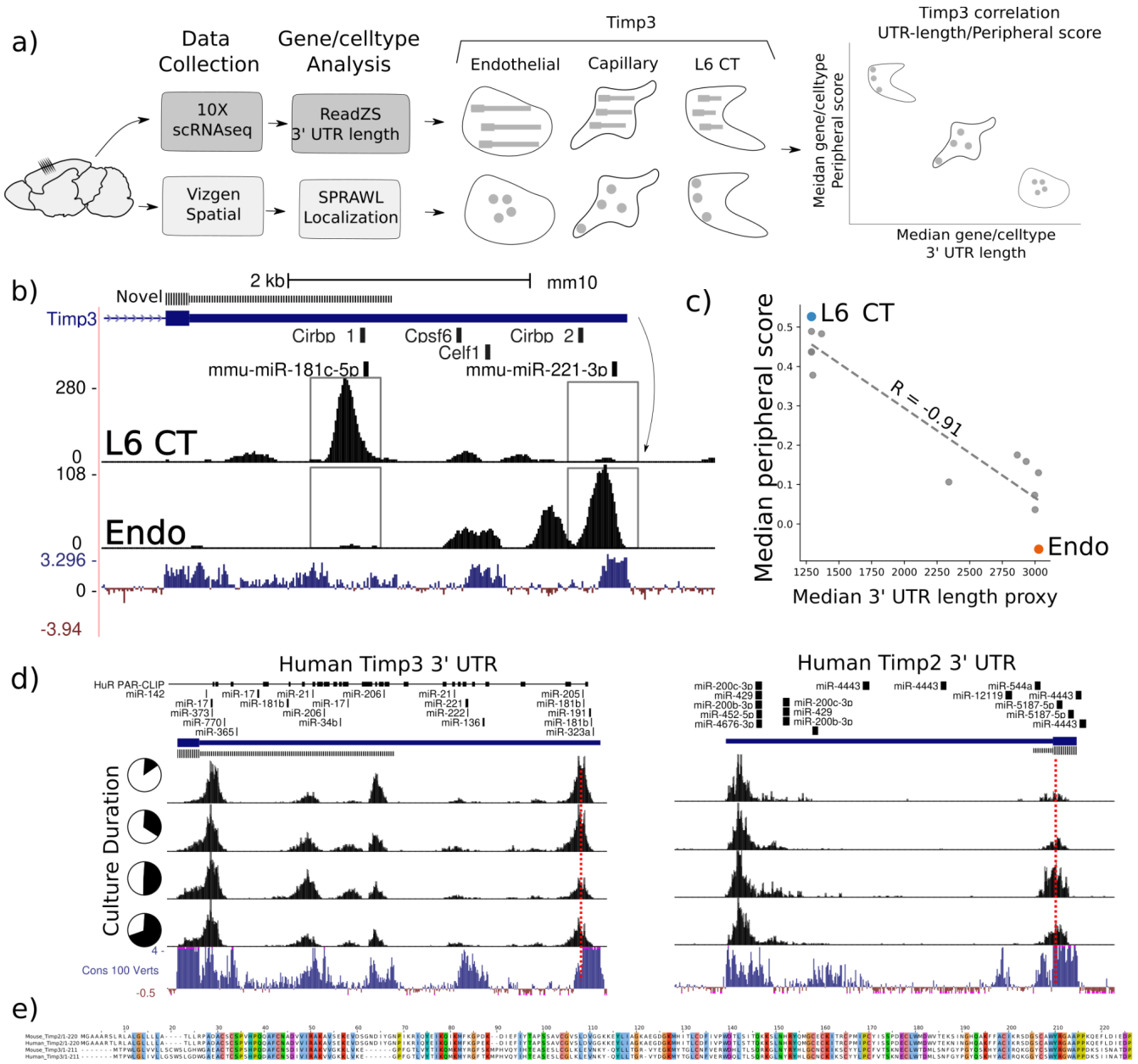


Figure 3: Timp3 alternative peripheral localization across MOp cell-types is statistically correlated with ReadZs differences in 3' UTR length.

a) Workflow to calculate median 3' UTR length and peripheral score per gene/cell-type. b) ReadZs detects two major alternative 3' UTRs in mouse Timp3 from 10X scRNAseq which correspond to miR-181c-5p and miR-221-3p binding sites. Reads from L6 CT cells predominantly map to a novel upstream shortened 3' UTR while endothelial cells primarily express the longer annotated 3' UTR. The UCSC genome browser placental animal sequence conservation shows highly conserved regions in blue. A Fisher's exact test was highly significant between the two peaks denoted by the dotted lines between the two cell-types. c) Timp3 mean periphery score is significantly correlated with Timp3 median ReadZs score across MOp cell-types with Pearson correlation coefficient of -0.91 and $p << 0.05$. d) Timp3 and Timp2 RNA 3' UTR length decreases during SARS-CoV2 infection and incubation duration in human lung tissue cultures. e) Timp3 and Timp2 mouse and human protein sequence alignment shows high sequence similarity and conservation.

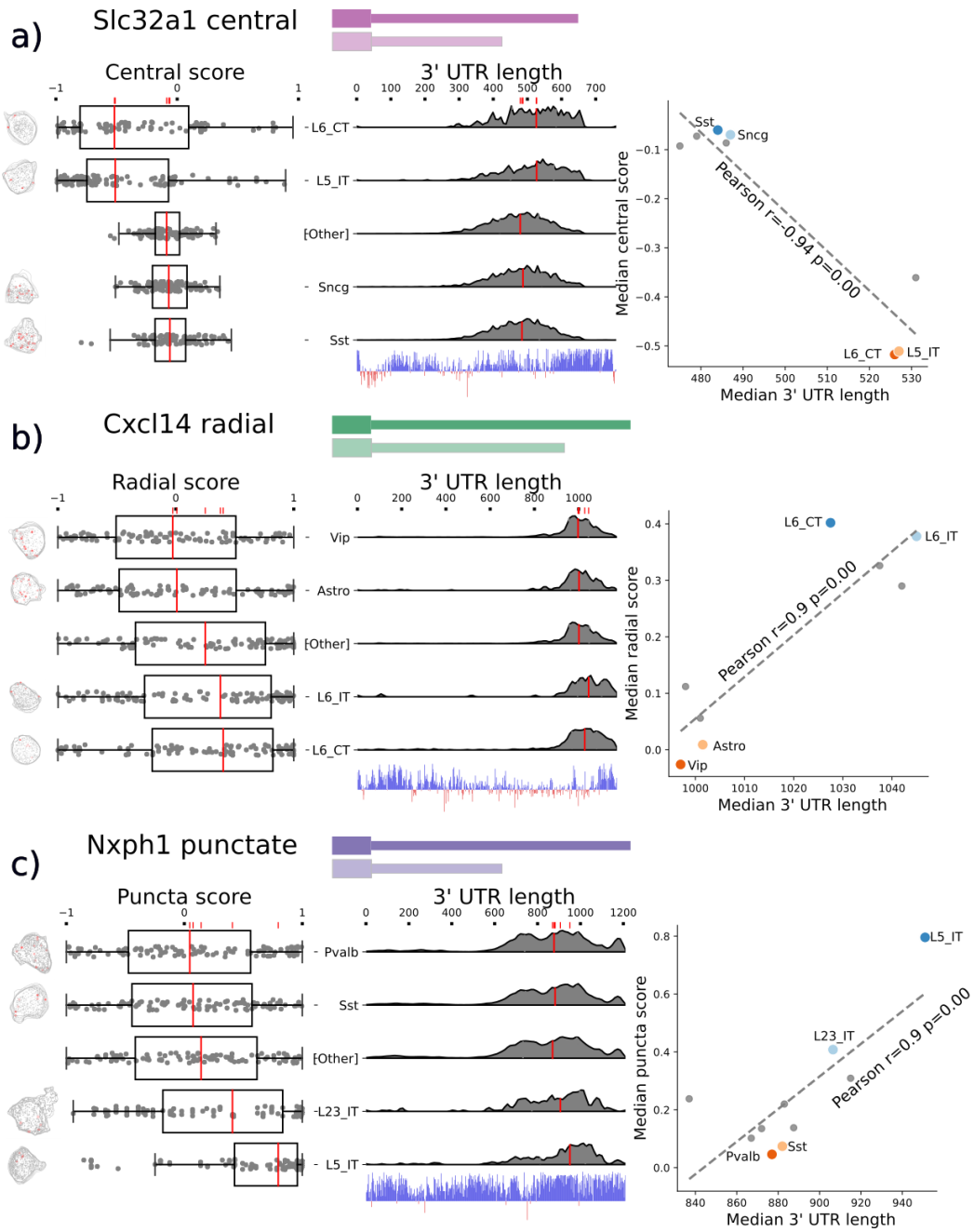
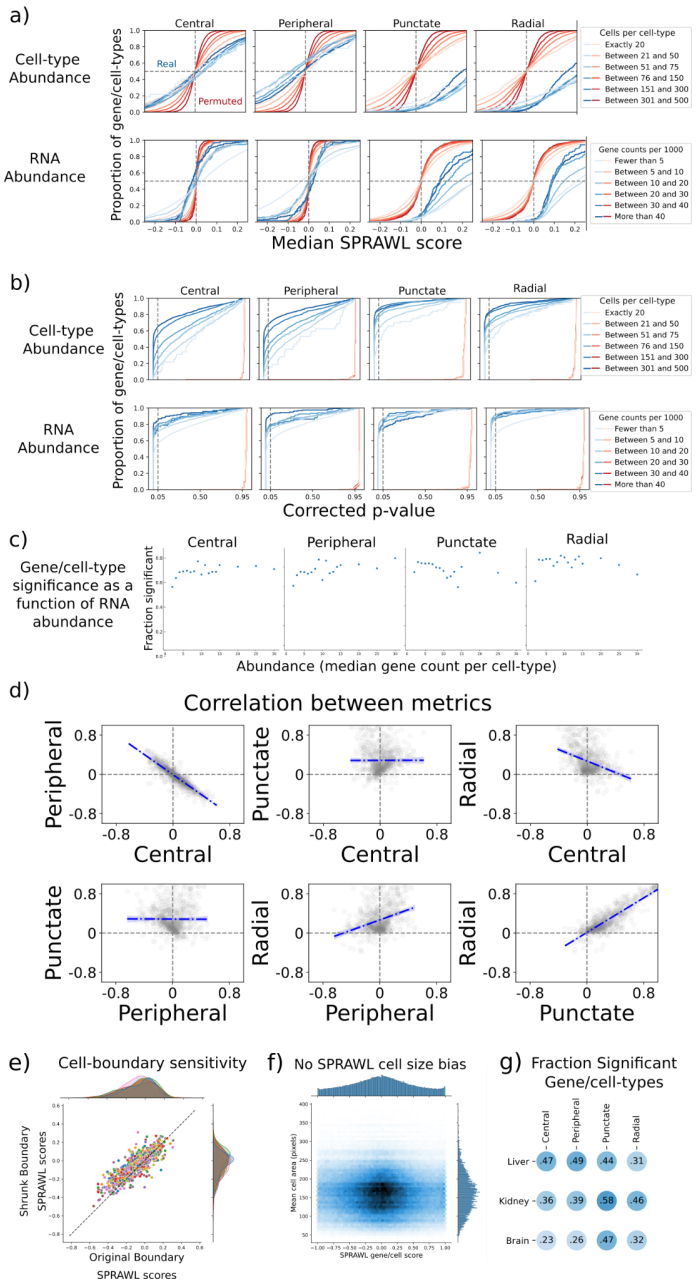


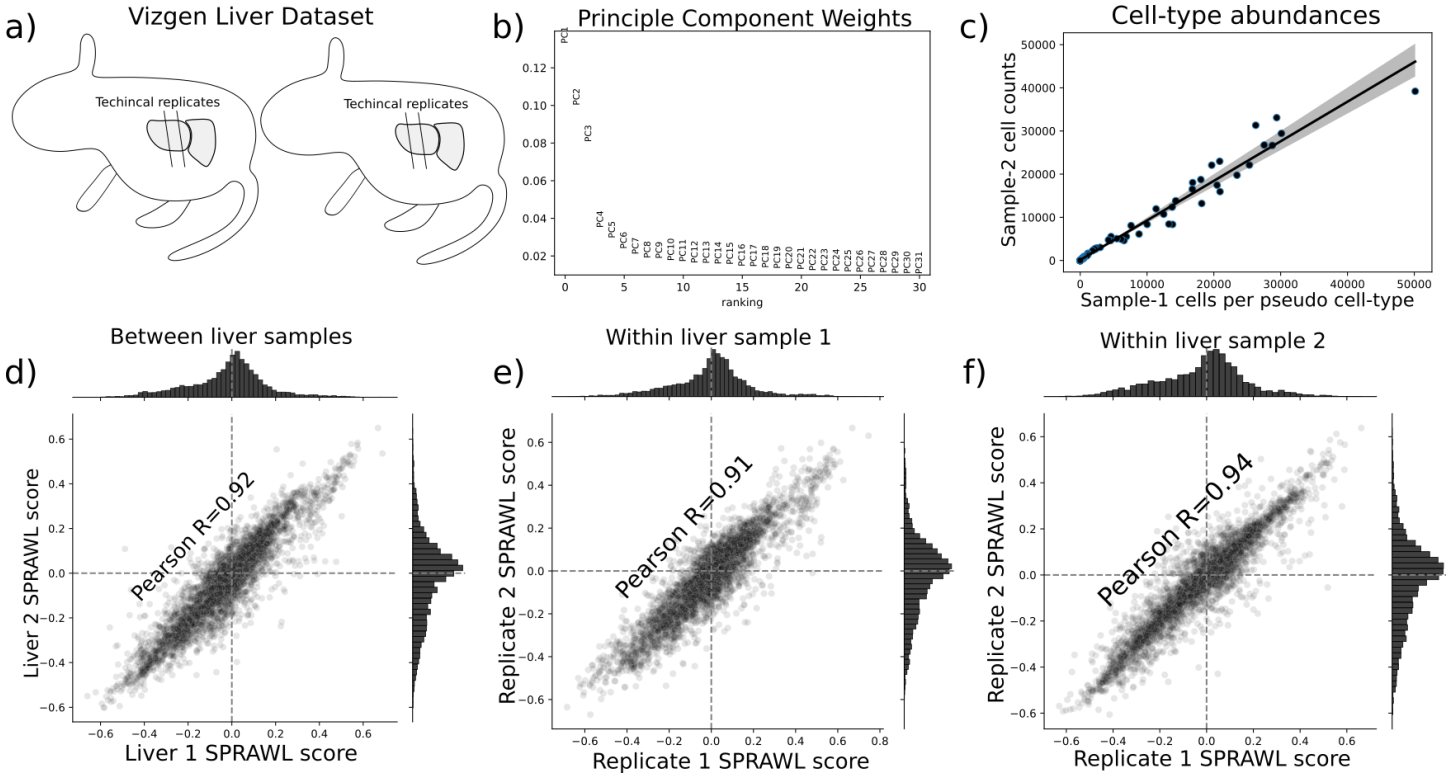
Figure 4: Slc32a1, Cxcl14, and Nxph1 have significant correlation between median SPRAWL central, radial, and punctate score and 3' UTR length respectively.
a) Slc32a1 median centrality, b) Cxcl14 radial, and c) Nxph1 punctate SPRAWL scores from the BICCN MERFISH dataset correlate significantly with 3' UTR length determined from 10X scRNAseq data by ReadZS. The left-column boxplots show individual SPRAWL cell scores as overlaid dots. The cell-types are sorted by increasing median score marked in red. The two cell-types with the highest and lowest median SPRAWL scores are plotted individually while the remaining cell-types are collapsed into the "Other" category. Gene/cell examples are shown to the left the boxplots for each extreme cell-type group. The density plots in the middle column show estimated 3' UTR lengths for each read mapping within the annotated 3' UTR, stratified by cell-type. Lengths were approximated as the distance between the annotated start of the 3' UTR and the median read-mapping position. Each density plot is normalized by cell-type to show relative shifts in 3' UTR length with median lengths depicted with red lines. The scatterplots show the significant correlations between median SPRAWL score and median 3' UTR length. The two cell-types with the highest, and the two with the lowest SPRAWL median scores are highlighted.

Experiment	Metric	Unique genes	Significant genes	Bidirectional genes
Vizgen Brainmap	Peripheral	589	224	3
Vizgen Brainmap	Central	589	208	2
Vizgen Brainmap	Radial	370	77	0
Vizgen Brainmap	Punctate	370	138	3
Vizgen Liver	Peripheral	385	380	215
Vizgen Liver	Central	385	379	202
Vizgen Liver	Punctate	385	323	64
BICCN MERFISH	Peripheral	252	251	92
BICCN MERFISH	Central	252	249	96
BICCN MERFISH	Radial	251	216	2
BICCN MERFISH	Punctate	251	236	10
SeqFISH+	Peripheral	9805	155	0
SeqFISH+	Central	9805	113	1
SeqFISH+	Radial	2423	907	131
SeqFISH+	Punctate	2423	815	83
CZB Kidney	Peripheral	307	273	44
CZB Kidney	Central	307	270	58
CZB Kidney	Radial	272	250	83
CZB Kidney	Punctate	272	250	68
CZB Liver	Peripheral	305	249	44
CZB Liver	Central	305	240	43
CZB Liver	Radial	162	136	67
CZB Liver	Punctate	162	136	28

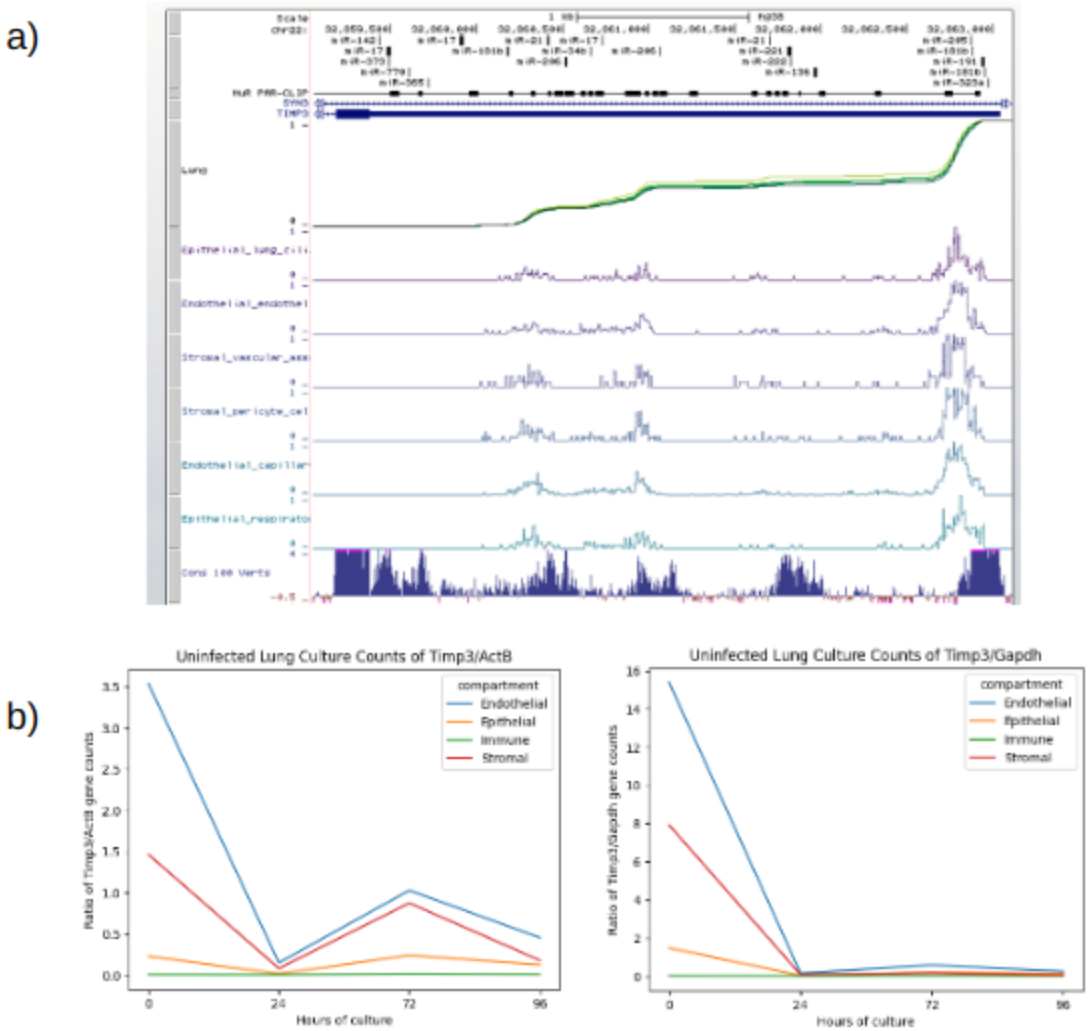
Supplemental Table 1: Counts of unique, significant, and opposite-effect genes in each experiment/metric combination. Genes are defined as significant if they are observed to be significant in at least cell-type in any replicate. Opposite-effect genes are those observed to have at least one significantly positive SPRAWL gene/cell-type score, and one significantly negative SPRAWL gene/cell-type score.



Supplementary Figure 1: SPRAWL metrics have high specificity and lack bias. a) SPRAWL scores for permuted null datasets, reds, have expected mean values of zero regardless of either the number of cells per cell-type or the gene abundance. The permuted datasets have expectedly lower variance for higher cells per cell-type and gene abundance. The real data, blue, shows expected means near 0 for the central and peripheral metrics, but higher scores for the punctate and radial metrics. b) Under null simulations, reds, no gene/cell-type pairs are incorrectly called significant at an alpha level of 0.05 (vertical dashed line) in any of the four metrics. In the real data, blues, more gene/cell-type pairs are significant, after Benjamini-Hochberg correction, with higher cell-type and RNA abundance. c) The fraction of significant gene/cell-type pairs in the BICCN samples are consistent across abundance levels measured as gene/cell-type median spot counts. d) Peripheral and central scores are strongly anti-correlated for gene/cell-type scores while the radial and punctate scores are positively correlated. e) To test whether peripheral localization patterns were driven artifactually by incorrect cell boundary calling, the cell boundary locations were computationally shrunk by a factor of 0.8 in the x and y direction, discarding spots that fell outside the new boundaries. In both the BICCN MOp and Vizgen Brainmap datasets, a Pearson correlation coefficient of greater than 0.85 was observed between the shrunk and original median gene/cell-type periphery scores. f) SPRAWL scores are not conflated with cell size g) Similar fractions of gene/cell-types are significant between the different datasets and metrics.

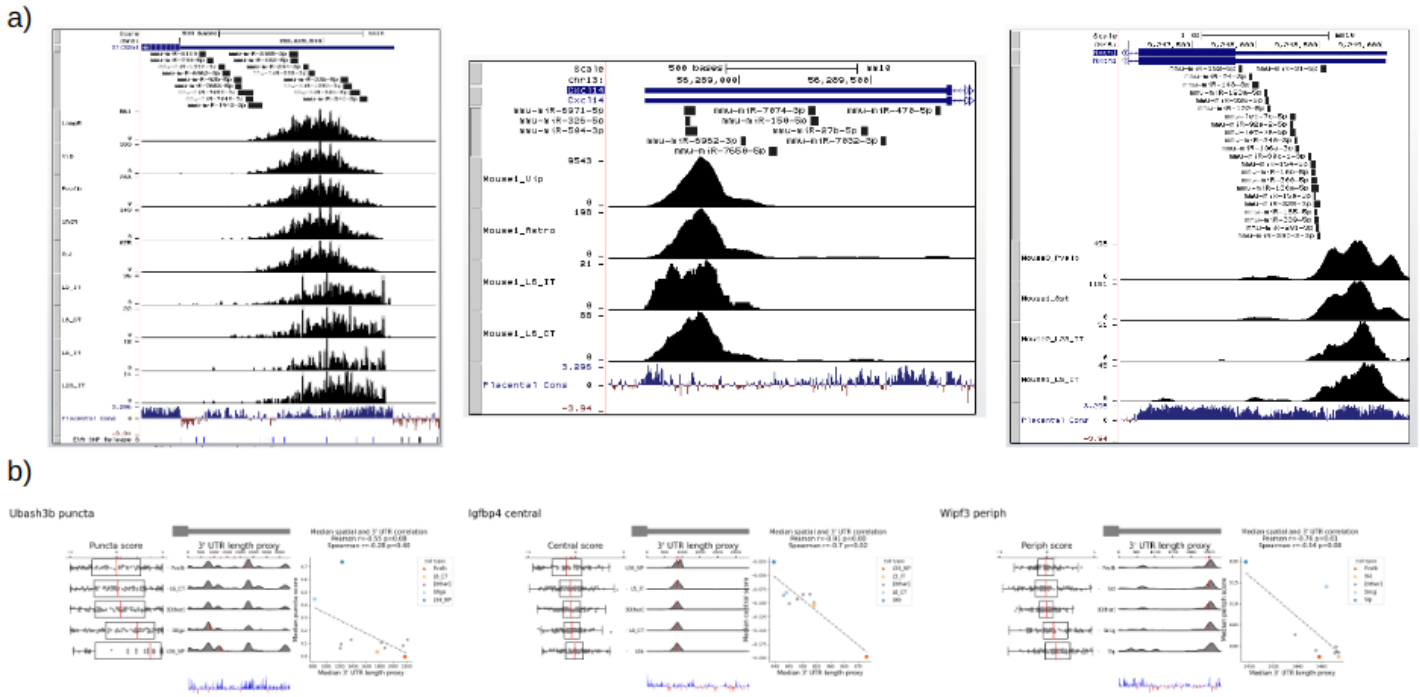


Supplemental figure 2: Vizgen Liver Showcase scores are highly correlated between replicates.
a) The Vizgen Liver showcase dataset provides spatial information for 2 mouse livers with 2 slices each. Cell annotation data was not provided in the Vizgen Liver Showcase, instead clusters produced from off-the-shelf Leiden clustering (python scanpy package) were used as pseudo cell-types. All four datasets were combined without reference to biological or technical replicate by first normalizing the read counts per cell, identifying highly variable genes, reducing to the first 10 principle components (b) and then computing a neighbor graph with n = 40 which resulted in 100 clusters which had similar number of cells from each animal c). As well as having high Pearson correlation coefficient between mice (d), the technical replicates were highly correlated within both Liver 1 (e) and Liver 2 (f).



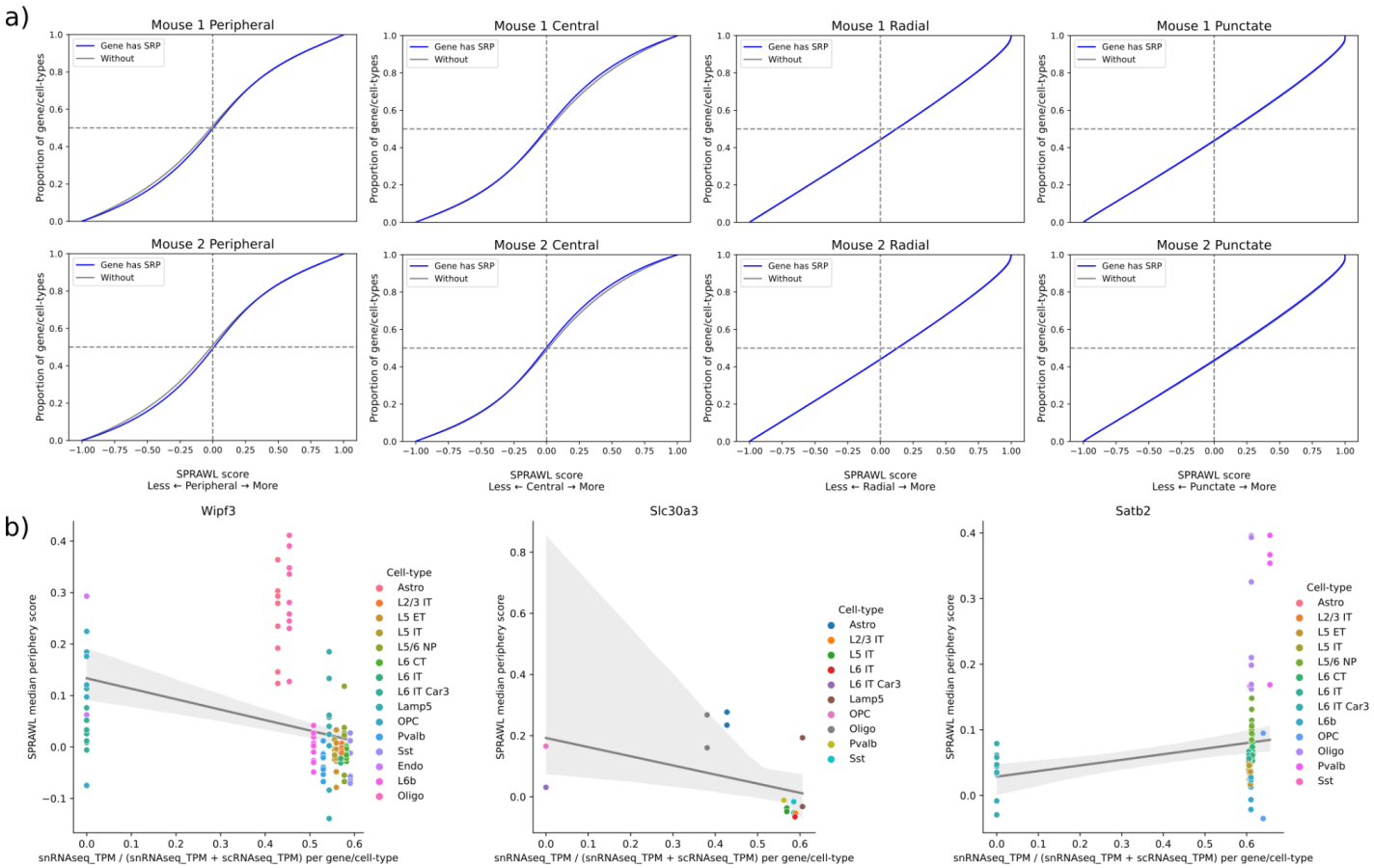
Supplemental Figure 3: ReadZS detects Tabula Sapiens Lung differential 3' UTR length Timp3 and decrease in Timp3 expression throughout culture

a) ReadZS detects statistically significant 3' UTR length differences in the human Timp3 3' UTR across endothelial cell-types from the Tabula Sapiens consortium datasets. The eCDF plot below the gene annotation shows slight variation in read buildup over all cell-types below which individually show read-density. HuR binding sites from PAR-CLIP are shown above the Timp3 gene structure diagram. The last track shows high vertebrate sequence conservation throughout the UTR. b) Normalized expression of Timp3 against Actin and Gapdh show decreasing expression of Timp3 throughout increasing culture direction in all tissue compartments.



Supplemental Figure 4: Computationally predicted miRNA binding sites in the 3' UTRs of Slc32a1, Cxcl4, and Nxph1 and additional 3' UTR correlated genes

a) Subset of computationally predicted 3' UTRs from the miRWalk database for Asic4, Slc32a1, and Nr2f2 indicate a potential mechanism of regulation for 3' UTRs of different lengths. b) Three genes Ubash3b, Igfbp4, and Wipf3 show significant negative correlation between various SPRAWL metrics and estimated 3' UTR length.



Supplemental figure 5 SPRAWL scores do not correlate with presence of signal recognition peptide, but do correlate with nuclear enrichment:
a) Genes encoding signal recognition peptides do not have significantly differential SPRAWL scores while b) genes such as *Wipf3* and *Slc30a3* have significantly lower peripheral scores in cell-types with higher nuclear expression. *Satb2* shows the opposite unexpected correlation.