The landscape of alternative polyadenylation in single cells of the developing mouse embryo

Vikram Agarwal[1,*,‡], Sereno Lopez-Darwin[2,*], David R. Kelley[1], Jay Shendure[2,3,4,‡]

[1]Calico Life Sciences, South San Francisco, CA 94080, USA

[2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

[3]Howard Hughes Medical Institute, Seattle, WA 98195, USA

[4]Brotman Baty Institute for Precision Medicine, University of Washington, Seattle, WA 98195, USA

[*]These authors contributed equally to this work

[‡]Correspondence to Vikram Agarwal (vagar@calicolabs.com) and Jay Shendure (shendure@uw.edu)

ABSTRACT

3′ untranslated regions (3′ UTRs) post-transcriptionally regulate mRNA stability, localization, and translation rate. While 3′-UTR isoforms have been globally quantified in limited cell types using bulk measurements, their differential usage among cell types during mammalian development remains poorly characterized. In this study, we examined a dataset comprising ~2 million cells spanning E9.5−E13.5 of mouse embryonic development to quantify transcriptome-wide changes in alternative polyadenylation (APA). We observe a global lengthening of 3′ UTRs across embryonic stages in all cell types, although we detect shorter 3′ UTRs in hematopoietic lineages and longer 3′ UTRs in neuronal cell types within each stage. While the majority of individual genes possess 3′ UTRs that lengthen with time, a subset appear to be spatiotemporally regulated through APA. By measuring 3′-UTR isoforms in an expansive single cell dataset, our work provides a transcriptome-wide and organism-wide map of the dynamic landscape of alternative polyadenylation during mammalian organogenesis.

INTRODUCTION

During transcriptional elongation, the cleavage and polyadenylation machinery governs the specification of the 3′ terminal end of an mRNA (Di Giammartino et al., 2011). This regulated process can generate a diversity of 3′-UTR isoforms for any given gene, dramatically altering the 3′-UTR length and sequence of the resulting mature transcripts (Elkon et al., 2013). This phenomenon, known as alternative polyadenylation (APA), has been observed in over 70% of mammalian genes (Derti et al., 2012; Tian and Manley, 2013). Alternative 3′-UTR isoforms bind to different sets of microRNAs and RNA-binding proteins, which collectively modulate a multitude of post-transcriptional gene regulatory mechanisms (Tian and Manley, 2017). These include changes in mRNA localization (Berkovits and Mayr, 2015), degradation rates (Agarwal et al., 2015; Lianoglou et al., 2013; Nam et al., 2014), and translational efficiency (Mayr, 2016). The differential abundance of a variety of nuclear factors serves to regulate APA in a cell-type-specific manner (Di Giammartino et al., 2011). Abnormal regulation of the cleavage and polyadenylation machinery has also been associated with hyperproliferative or disease states such as cancer (Fu et al., 2011; Mayr and Bartel, 2009; Sandberg et al., 2008).

Techniques to directly measure APA in the transcriptome largely rely upon the isolation of RNA from bulk tissue, resulting in an average readout of the landscape of 3′ ends in a heterogeneous population of cells. Existing 3′-end sequencing methods include 3′-seq/3SEQ (Beck et al., 2010; Elkon et al., 2012), 3P-seq (Jan et al., 2011; Nam et al., 2014), PAS-seq (Shepard et al., 2011), 3′READS (Hoque et al., 2013), PolyA-seq (Derti et al., 2012), and 2P-seq (Spies et al., 2013). The successful application of these methodologies in mammalian cells has led to the annotation

of hundreds of thousands of polyadenylation sites (PAS) in both human and mouse genomes (Herrmann et al., 2020; Wang et al., 2018). Bulk 3′-end sequencing and similar transcriptomic data have guided the observation that 3′ UTRs generally lengthen during mammalian embryogenesis (Ji et al., 2009), with proliferating cell types such as blood exhibiting shortened 3′ UTRs (Mayr and Bartel, 2009; Sandberg et al., 2008) and neuronal ones exhibiting lengthened 3′ UTRs (Miura et al., 2013; Shepard et al., 2011).

In contrast to bulk methods, single-cell RNA sequencing (scRNA-seq) protocols capture a rich diversity of individual cell types, with many protocols enriching for mRNA 3′ ends via poly(A) priming (Cao et al., 2017; Hashimshony et al., 2016; Jaitin et al., 2014; Macosko et al., 2015; Semrau et al., 2017; Zheng et al., 2017). Thus, these technologies inherently offer an unprecedented opportunity to observe APA events during the process of cellular differentiation. They also enable the decomposition of complex tissues into individual cell types, enabling the assessment of APA with greater cell type resolution. Although a proof-of-concept study has demonstrated the utility of scRNA-seq data in evaluating APA (Shulman and Elkon, 2019), such methods have not yet been applied to investigate more comprehensive scRNA-seq datasets such as those capturing dozens of cell types during a mammalian developmental time course (Cao et al., 2019; Pijuan-Sala et al., 2019). In this study, we examined APA using MOCA ("mammalian organogenesis cell atlas"), a dataset comprising single nucleus transcriptional profiling of ~2 million cells encompassing 38 major cell types across five stages (*i.e.*, E9.5, E10.5, E11.5, E12.5, and E13.5) of mouse embryonic development (Cao et al., 2019).

RESULTS

An integrated annotation set of 3′ UTRs and poly(A) sites to evaluate APA

Given the reliance of many scRNA-seq protocols on poly(A) priming, such methods enrich for both mRNA 3′ ends as well as internal A-rich stretches of homopolymers. Thus, internal priming artifacts obscure accurate quantitation of APA, even more so in datasets in which immature mRNAs (*i.e.*, without excised introns) are isolated from the nucleus, as is the case with the sci-RNA-seq3 protocol used in MOCA (Cao et al., 2019). We therefore sought to develop a simple computational method to deconvolve the data to specifically isolate and quantify mRNA 3′ ends. Towards this goal, we built integrated databases of poly(A) site (PAS) and 3′-UTR annotations to guide the interpretation of which subset of mapped reads were supported by orthogonal evidence to reflect authentic 3′ termini, as opposed to A-rich sites internal to a mature or nascent transcript. In doing so, our goal was to minimize the shortcomings of any individual database, each of which utilize different data sources and strategies for PAS and 3′-UTR annotation.

To generate a reliable PAS set, we considered three of the most comprehensive mouse PAS annotation databases available with respect to the mm10 mouse genome build: Gencode M25 (Frankish et al., 2019), which contains 56,592 PASs; PolyA_DB v3 (Wang et al., 2018), which contains 128,052 PASs; and PolyASite 2.0 (Herrmann et al., 2020), which contains 108,938 PASs. We intersected the PASs from each pair of these resources to evaluate the consistency among databases. While the majority of sites were present in at least two databases, 40.0%, 29.4%, and 30.4% were unique to PolyA_DB, PolyASite, and Gencode, respectively (**Figure**

1A). To verify the reliability of PASs present in only a single database (and therefore the most likely to contain false positives), we plotted the profile of nucleotide frequencies in the ±50nt region surrounding the annotated cleavage and polyadenylation sites (**Figure 1—figure supplement 1A**). The unique PASs of each resource exhibited profiles consistent with positionally-enriched mammalian motifs known to guide mRNA cleavage, including several U-rich motifs, the upstream AAUAAA motif, and the downstream GU-rich motif (Tian and Graber, 2012). Moreover, we detected a strong enrichment of reads mapping immediately upstream of this set of PASs, with the strongest enrichment spanning the -300nt to +20nt region around the PAS (**Figure 1—figure supplement 1B**). Given that each PAS database was enriched in known PAS motifs, associated with mapped reads, and held information complementary to the other databases, we carried forward an integrated PAS set derived from the union of the three databases. This integrated PAS set recapitulated these same characteristics, exhibiting both consistency with known PAS motifs and strong read enrichment upstream of the sites (**Figure 1B,C**). Finally, to link PASs to specific genes, we utilized our previous 3′-UTR annotation pipeline (**Methods**) (Agarwal et al., 2018, 2015) to establish an integrated set by carrying forward the longest 3′ UTRs from four resources: i) Gencode M25 (Frankish et al., 2019), ii) RefSeq (O'Leary et al., 2016), iii) 3′ UTRs with extreme lengthening (Miura et al., 2013), and iv) bulk 3P-seq-based annotations derived from ten mouse tissues and cell lines (Nam et al., 2014). This integrated 3′-UTR annotation set helped minimize the possibility that a PAS may be annotated outside of a known 3′ UTR and thus remain unlinked to a specific gene.

Using our integrated PAS and 3′-UTR databases (**Figure 1D**), we sequentially filtered our scRNA-seq reads from MOCA to focus on the subset mapping to 3′ UTRs within the -300 to +20 vicinity of a known PAS (**Figure 1C**). Due to the abundant mapping of reads to introns in the nucleus-derived MOCA dataset (**Figure 1—figure supplement 1C**), generally representing internal priming within unspliced transcripts, there was nearly a 10-fold loss in read counts after iterative steps of filtering; however, over 200 million reads were carried forward (**Figure 1—figure supplement 1D**). We then counted reads passing the filtering steps towards the single annotated PAS in its vicinity, enabling the tabulation of read counts associated with each PAS (**Figure 1D**). In ambiguous cases in which a read was located in the vicinity of multiple PASs, we greedily assigned the read to count towards the PAS harboring the most uniquely assignable reads. Finally, based upon the relative counts assigned to each PAS for a given gene, we visualized the "isoform inclusion rate" (IIR), reflecting the proportion of 3′-UTR isoforms which include a given nucleotide position (Agarwal et al., 2018, 2015; Nam et al., 2014) (**Figure 1D**).

To validate that these filtering and read-to-PAS assignment procedures led to reliable results, we performed two quality control (QC) comparisons. As a first QC, reasoning that the removal of internal priming artifacts should improve the quantitation of relative gene expression levels, we compared the relationship between PAS counts and median gene expression levels computed across a panel of 254 mouse RNA-seq samples (Agarwal and Shendure, 2020). While the traditional method of counting reads in the gene body displayed a strong correlation to median expression levels (Pearson r = 0.81, Spearman rho = 0.77), it displayed a clear bias in inflating estimates for a large cohort of genes (**Figure 1E**). Considering only our filtered PAS-assigned

reads ameliorated this bias, which led to a stronger correlation to relative mRNA expression levels (Pearson r = 0.88, Spearman rho = 0.85) (**Figure 1E**). We speculated that the bias in the gene body method relative to the PAS-assigned read counting method could be explained by the over-abundance of intron-mapping reads (**Figure 1—figure supplement 1C**) and enrichment of A-rich stretches that nucleate the production of internal priming artifacts. Indeed, a lasso regression model trained to predict the difference between the two strategies confirmed that intron length was strongly associated with inflated counts; moreover, "AAA" was the top-ranked of all 3-mers associated with inflated counts in the gene body (Pearson r = 0.56, Spearman rho = 0.58, **Figure 1—figure supplement 2**).

As a second QC, we evaluated the similarity between our IIR profiles to those derived from bulk 3P-seq data (Nam et al., 2014). We considered the latter as a gold standard in accurately quantifying PAS abundances due to the involvement of a splint-ligation step in the 3P-seq protocol, which specifically removes internal priming artifacts (Jan et al., 2011). We found that the IIR profiles for our PAS-assigned reads more strongly mirrored those of bulk data for two representative genes (**Figure 1F**). Quantifying the deviation from bulk as the Mean Absolute Deviation (MAD) (**Figure 1—figure supplement 3A**) allowed us to measure the deviations between our pre- and post-processed data to bulk 3P-seq measurements. Applying this metric globally to all genes uncovered that 78% of genes exhibited improved similarity to bulk after the reads were assigned to PASs; moreover, 47% of genes achieved MAD <= 0.1 after read-to-PAS assignment, relative to only 7% of genes beforehand (**Figure 1—figure supplement 3B**). Inspection of IIR profiles for nine representative genes further confirmed the general

improvement in consistency with bulk 3P-seq data (**Figure 1—figure supplement 3C**).

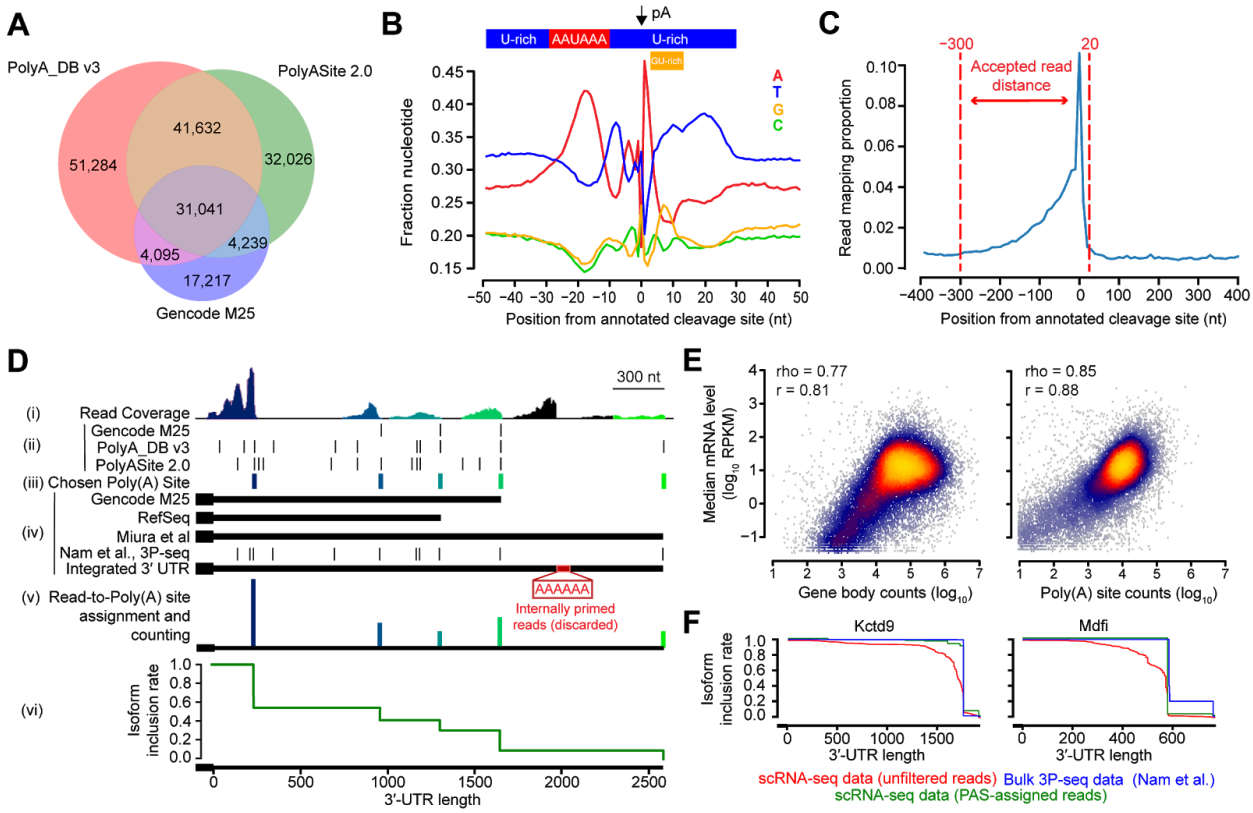

Figure 1. A computational pipeline to accurately quantify 3′-UTR isoform abundances from scRNA-seq data. (A) Venn diagram of a set of three PAS annotation resources and their degree of intersection. A PAS intersecting within ±20nt from another was considered an intersecting hit to account for the heterogeneity of the cleavage and polyadenylation machinery (Jan et al., 2011). **(B)** Profile of nucleotide frequencies in the ±50nt vicinity of the annotated cleavage site position, derived from the union of the three databases. Shown above the plot are the known positionally-enriched mammalian motifs known to guide mRNA cleavage (Tian and Graber, 2012). **(C)** Distribution of scRNA-seq reads mapping within the ±400nt vicinity of the annotated cleavage site position, derived from the union of the three databases. To avoid an ambiguous signal, the analysis was restricted to PASs not within the same ±400nt window as another PAS. Data is binned at 5nt resolution. Shown within the dotted red lines are the acceptable distance thresholds to associate a read to an annotated PAS. See also **Figure 1—figure supplement 1** for comparisons of (B-C) for each individual PAS database. **(D)** Schematic depicting the association of each scRNA-seq read to a PAS in order to quantify relative PAS abundances for a gene. Shown from top to bottom are: (i) The read coverage of scRNA-seq reads mapped to the gene. (ii) The three PAS annotation resources considered, showing the location of each PAS along the 3′ UTR. (iii) The subset of chosen PASs to which reads were greedily assigned, colored from blue to green to indicate which reads from the coverage plot were assigned to them. (iv) The three gene annotation databases integrated with bulk 3P-seq data from ten tissues and cell lines (Nam et al., 2014) to identify the longest known 3′ UTR. This integrated 3′ UTR was used to associate PASs to the gene. (v) A visualization of relative 3′-UTR isoform abundances after read-to-PAS assignment, with vertical lines at each chosen PAS

proportional to the assigned number of read counts. Reads not overlapping within the -300 to +20 vicinity of a known PAS were treated as likely internal priming artifacts and discarded. (vi) The resulting isoform inclusion rate (IIR) plot to quantify the cumulative proportion of 3′-UTR isoforms remaining along the length of a 3′ UTR. See also **Figure 1—Source Data 1** for the integrated 3′-UTR database and gene annotations. **(E)** Scatter plots comparing gene expression levels estimated using scRNA-seq read abundances mapping to the full gene body (left panel) or the sum of reads mapping to PASs (right panel), relative to median gene expression levels from bulk RNA-seq data (Agarwal and Shendure, 2020) (n = 19,517 protein-coding genes). Regions are colored according to the density of data from light blue (low density) to yellow (high density). Shown are the corresponding Pearson (r) and Spearman (rho) correlations for each comparison. See also **Figure 1—figure supplement 2** for sequence features explaining biased estimates in the gene body approach. **(F)** Shown are IIR plots for two genes, comparing the profiles for the raw scRNA-seq data and post-processed data after read-to-PAS assignment with respect to the profile for bulk 3P-seq data (Nam et al., 2014) as a gold standard. Slight vertical jitter was added for enhanced line visibility. See also **Figure 1—figure supplement 3** for a global comparison among all genes.

Global differences in 3′-UTR length across mouse cell types and developmental time

Having assigned reads to PASs and linked them to genes, we next sought to evaluate global properties of 3′-UTR shortening and lengthening (*i.e.*, as quantified by differential PAS usage) across cell types and developmental time. Towards this goal, we computed a "gene by cell" sparse matrix of the mean length among all 3′-UTR isoforms, weighted by their respective counts. For each gene, we then computed each cell's deviation from the mean of 3′-UTR lengths across cells, considering only non-missing values. Finally, for each cell, we computed the mean of these deviations across genes as a measure of the global behavior of the transcriptome through the perspective of APA. We projected these measurements onto a global map of 38 t-SNE clusters representing all major mouse cell types (Cao et al., 2019). Highlighting the top ten ranked t-SNE clusters with the largest differences, we discovered the greatest average 3′-UTR lengths among stromal cells and three neuronal cell types; in contrast, the shortest lengths occurred in three blood cell types, hepatocytes, chondrocytes, and osteoblasts (**Figure 2A**). Sub-clustering each of the 38 t-SNE clusters reinforced these findings, but revealed additional

heterogeneity within each cell type (**Figure 2—figure supplement 1**). Segregating our dataset by the five sampled timepoints, we observed an apparent global 3′-UTR lengthening across developmental time (**Figure 2B**). Finally, to quantify the joint impact of cell type and developmental stage, we computed the average behavior among genes and cells associated with each of 38 t-SNE clusters and 5 developmental stages. Partitioning the data in this manner reinforced our observation that average 3′-UTR length increased in nearly every cell type as developmental time progressed (**Figure 2C**). Neuronal cell types clustered together and exhibited the greatest 3′-UTR lengthening relative to other clusters at E13.5; in contrast, blood cell types exhibited highly shortened 3′ UTRs at E9.5 and grew until E13.5 to mean lengths similar to those of other cell types at E9.5 (**Figure 2C**).

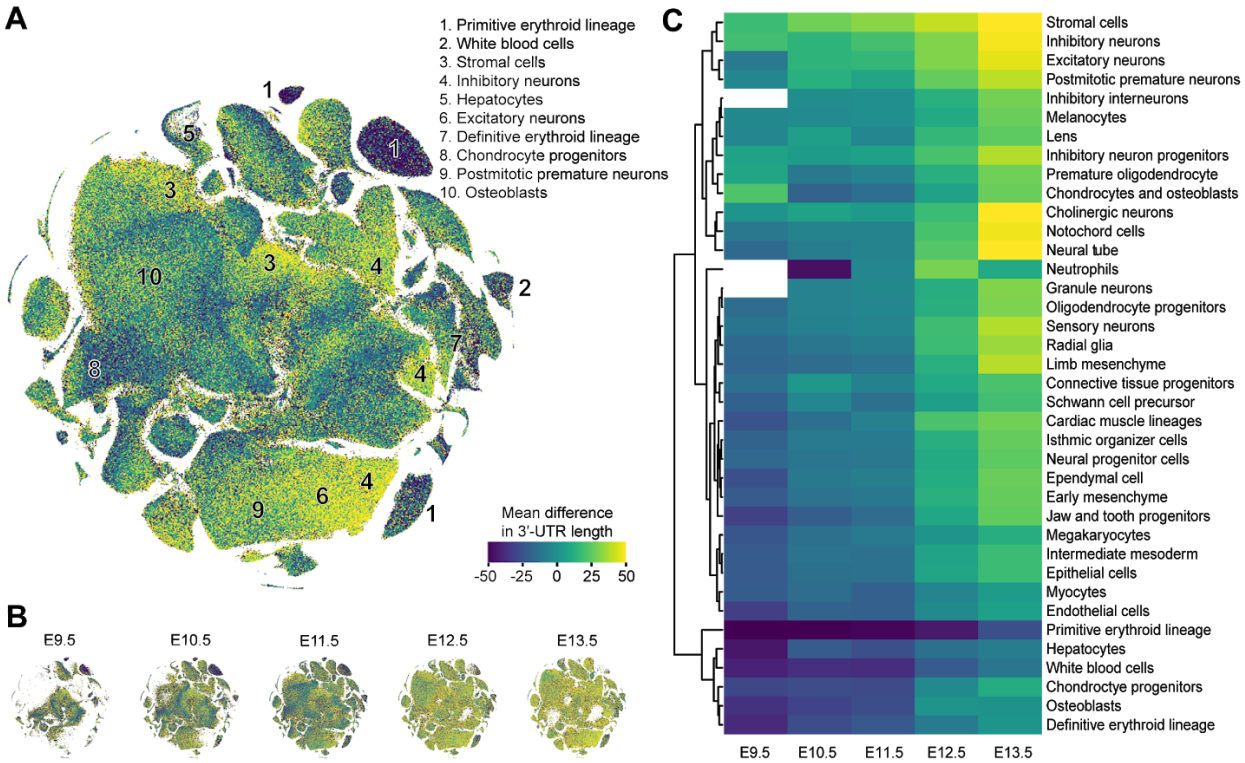

Figure 2. Differential 3′-UTR lengthening among diverse cell types and developmental stages. (A) t-SNE embedding of all cells from all developmental stages sampled (Cao et al., 2019), with each cell colored according to the mean difference in 3′-UTR lengths across all genes. The top ten ranked clusters

with the greatest differences are annotated according to their corresponding cell type. **(B)** Shown are the same embeddings and color scales as those in panel (A), except after partitioning the dataset into its five composite developmental stages (spanning E9.5-E13.5). **(C)** Heatmap of the mean difference in 3′-UTR length after aggregating cells from each developmental stage and cell type, derived from each of 38 t-SNE clusters. Color scales are the same as those shown in panel (A). Missing values (shown in white) correspond to instances with too few (<20) cells to accurately estimate. Heatmap is clustered by Euclidean distance as a distance metric. See also **Figure 2—figure supplement 1** for comparisons among t-SNE subclusters.

Next, we evaluated differences in global 3′-UTR length with respect to developmental trajectories computed using UMAP, an embedding that more faithfully recapitulates cell–cell relationships and intermediate states of differentiation relative to t-SNE. Evaluating ten developmental UMAP trajectories (Cao et al., 2019), we again observed a global lengthening in 3′ UTRs in nearly every trajectory (**Figure 3A**). Mirroring our previous findings, the neural tube/notochord and the neural crest trajectories (capturing neurons of the peripheral nervous system) showed the greatest lengths relative to other cell types at E13.5, while the hematopoiesis trajectory displayed the shortest lengths relative to other cell types at E9.5 (**Figure 3A**). A visual comparison of these three trajectories with respect to changes in both developmental time and 3′-UTR length showed that the process of 3′-UTR lengthening occurred contemporaneously with cellular differentiation, with gradients of lengthening emerging in intermediate cellular states (**Figure 3B-D**). Notably, in the hematopoiesis trajectory, a major difference in 3′-UTR length could be explained by the switch from primitive to definitive erythropoiesis, rather than gradual lengthening within either lineage (**Figure 3D**).

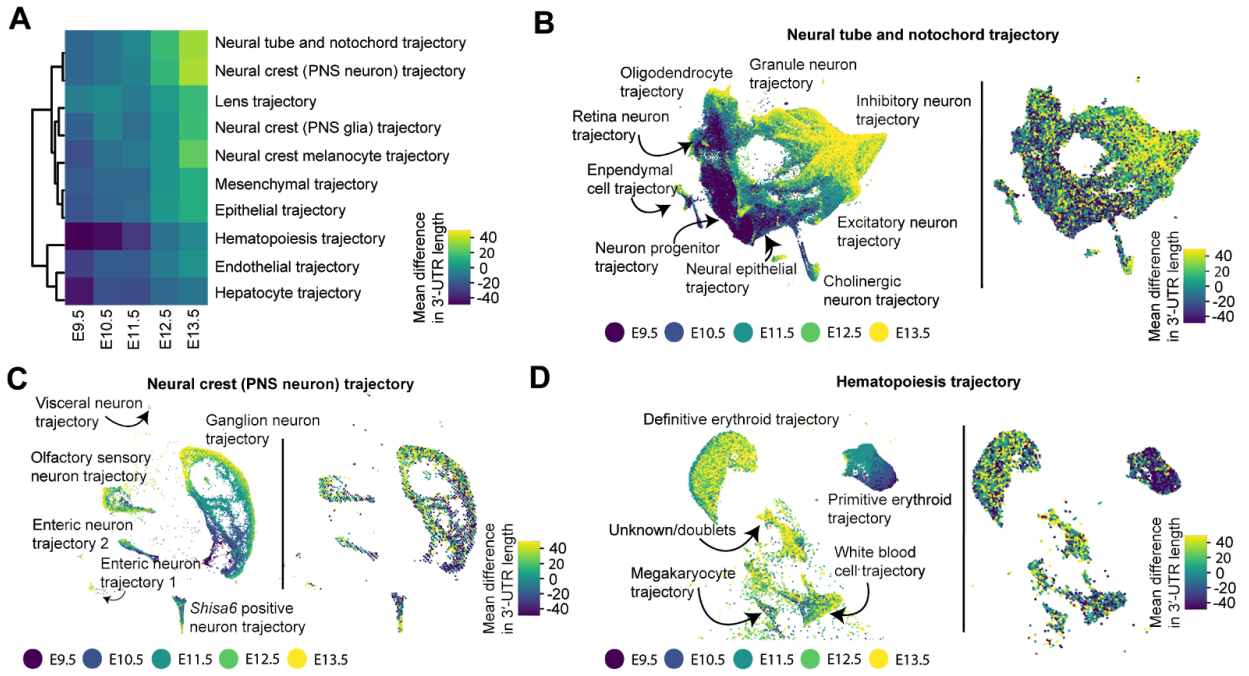

Figure 3. Differential 3′-UTR lengthening among diverse developmental trajectories. (A) Heatmap of the mean difference in 3′-UTR length after aggregating cells from each developmental stage and one of ten developmental trajectories computed using UMAP (Cao et al., 2019). Heatmap is clustered by Euclidean distance as a distance metric. **(B-D)** UMAP embeddings of the neural tube and notochord (B), neural crest of the peripheral nervous system (C), and hematopoiesis trajectory (D). Cells from each plot are colored by developmental stage (left panel) and mean difference in 3′-UTR length across genes (right panel). Mean differences are calculated with respect to the cells shown in the UMAP rather than all cells.

Dynamic gene-specific patterns of alternative polyadenylation across early development

While our previous analyses revealed transcriptome-wide trends, it remained unclear how specifically changes in APA manifested at the resolution of individual genes. To investigate this, we tabulated read counts assigned to each gene (*i.e.*, for each PAS and developmental stage, aggregating information across cell types), and used a χ^2 test (Shulman and Elkon, 2019) to evaluate statistically significant differences in APA for 8,653 genes with at least 100 reads in each of the five developmental stages (**Supplementary File 1**). This procedure identified 5,169 genes surpassing a False Discovery Rate (FDR) corrected p-value threshold of 0.05. Evaluating the dynamics of the mean 3′-UTR length for this cohort of significant genes at each stage, we

discovered that 62% of genes fell into a large cluster that exhibited consistent lengthening over time (**Figure 4A**). While the majority of these genes showed the greatest increase in lengthening from E11.5 to E12.5, a minority lengthened the most strongly from E10.5 to E11.5 (**Figure 4A**). About 38% of the significant genes did not simply lengthen across developmental stages, with about half of these progressively shortening over time (**Figure 4A**). As an alternative method to evaluate transcriptome-wide changes, we computed the entropy across PASs for each gene and developmental stage. This alternative visualization scheme uncovered that nearly 75% of genes obey a progressive decrease in entropy, indicating that as developmental time progresses, a few PASs become increasingly dominant for the vast majority of genes; in contrast, about 15% of genes exhibited the opposite pattern of increased entropy across time, with the remaining displaying heterogeneous patterns (**Figure 4—figure supplement 1A-B**).

We extended our previous gene-centric IIR plotting scheme (**Figure 1D**) to visualize the landscape of APA across the five developmental stages assayed, this time using a χ^2 test to highlight individual PASs which were significantly different in at least one stage (**Figure 4B**). Using this scheme, we visualized an assortment of genes from different clusters to dissect the nature of the isoform switching events contributing to changes in 3′-UTR lengths (**Figure 4C**). Many of these genes contained dozens of PASs whose relative proportions significantly changed across time. For genes belonging to the dominant cluster (Tmem33, Lrtm1, Dcp1b, and Add2 in **Figure 4A**), later developmental stages led to the progressive selection of distal isoforms, leading to progressive 3′-UTR lengthening (**Figure 4C**). The opposite pattern was observed for a gene belonging to a smaller cluster (Srl in **Figure 4A**), whereby the proximal isoform was

selected more frequently over the distal as time progressed, leading to progressive 3′-UTR shortening (**Figure 4C**). For yet other genes, the choice of distal isoforms was highly time-dependent. For example, Mtap displayed a near-complete proximal-to-distal isoform switching event in E11.5, subsequently lengthening beyond baseline levels in later developmental stages; in contrast, Mrpl22 exhibited progressive shortening, with a dominant distal-to-proximal isoform switching event occurring in E10.5 (**Figure 4C**).

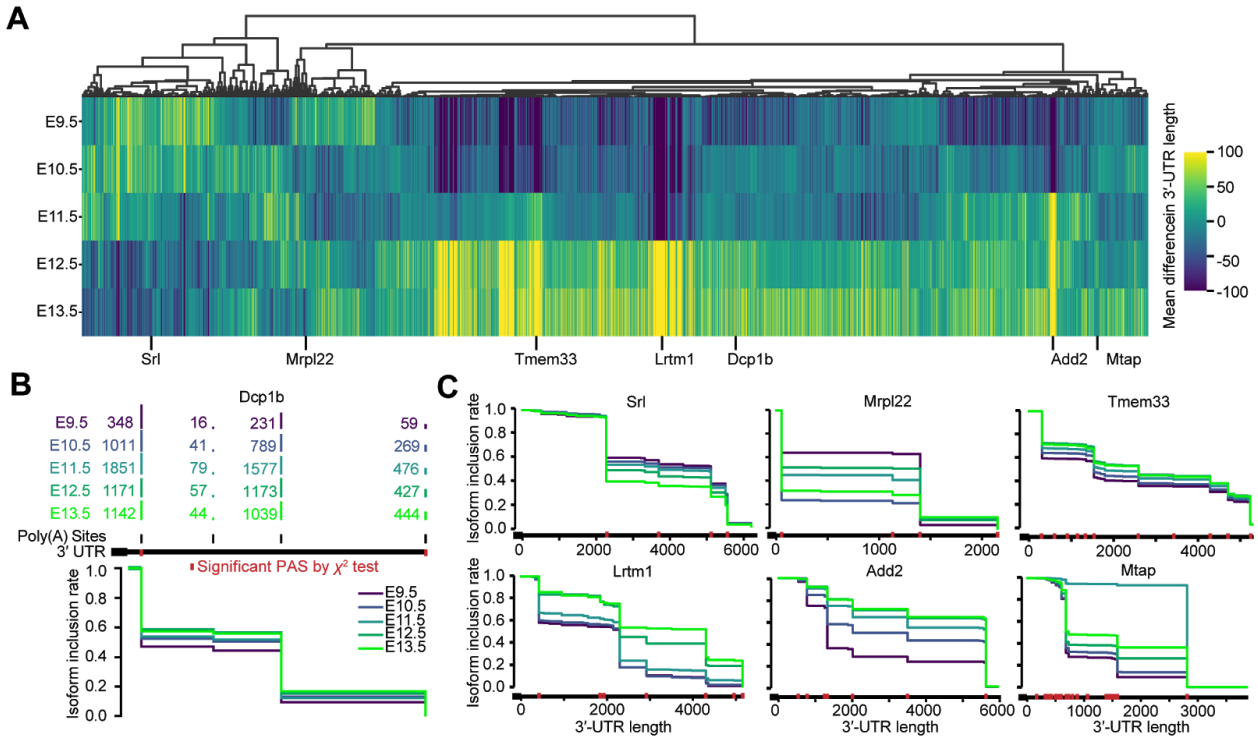

Figure 4. Identification of distinct gene lengthening patterns and genes responsible for overall data trends across embryonic ages. (A) Heatmap of mean differences in 3′-UTR lengths for 5,169 genes with significant differences in PAS usage across embryonic stages. Heatmap is column-centered and clustered by Pearson correlation as a distance metric. **(B)** Schematic of IIR plot visualization using PAS counts for each of five embryonic stages. Vertical red lines along the 3′ UTR indicate PASs that are significantly different between stages by the χ^2 test ($p<0.05$). **(C)** IIR plots for six genes among representative clusters shown in panel (A), and colored by embryonic stage. See also **Figure 4—figure supplement 1** for clustering according to differences in entropy. See also **Figure 4—Source Data 1** for a table of read counts associated with each PAS for each gene and embryonic stage.

Finally, we performed a similar gene-centric analysis, this time evaluating differences among individual cell types (*i.e.*, aggregating information across the five developmental stages). Among 1,491 genes with at least 20 reads in each of the 38 t-SNE clusters, we identified 1,078 genes surpassing an FDR-corrected p-value threshold of 0.05, as evaluated by the χ^2 test (**Supplementary File 2**). This subset of significant genes largely clustered into four cell type groups (C1-C4, **Figure 5A**) when evaluating differences in mean 3′-UTR length, with cell types within each group displaying strongly correlated patterns across all of the genes. C1, which consisted primarily of neuronal cell types, was unique in that the vast majority of genes displayed global lengthening; conversely, the primitive erythroid lineage was dominated by genes experiencing 3′-UTR shortening (**Figure 5A**). In special cases, we detected a highly gene-specific and cell-type-specific pattern, as in the case of Bclaf1 showing 3′-UTR lengthening within a t-SNE cluster annotated as lens cells (**Figure 5B**). However, for most genes, all of the cell types within each cluster displayed a concerted shift towards either 3′-UTR lengthening (*e.g.*, cluster C1 in Gnb1, C1 and C4 in Samm50) or 3′-UTR shortening (*e.g.*, cluster C2 in Polr3k, C1 in Hoxd4) (**Figure 5B**).

When visualizing PAS usage with respect to entropy, several cell types emerged as displaying interesting patterns: the primitive erythroid lineage showed heightened entropy across most genes, whereas neutrophils—and to a smaller degree, the lens—showed decreased entropy (**Figure 5—figure supplement 1A-B**). This observation suggests a potential for cell-type-specific regulatory mechanisms that guide a more stochastic or more defined choice of PASs, respectively. A smaller subset of genes, such as Sec11a, notably displayed higher entropy

among neuronal cell types, consistent with an active mechanism governing a switch towards longer 3′-UTR isoforms (**Figure 5—figure supplement 1A-B**).

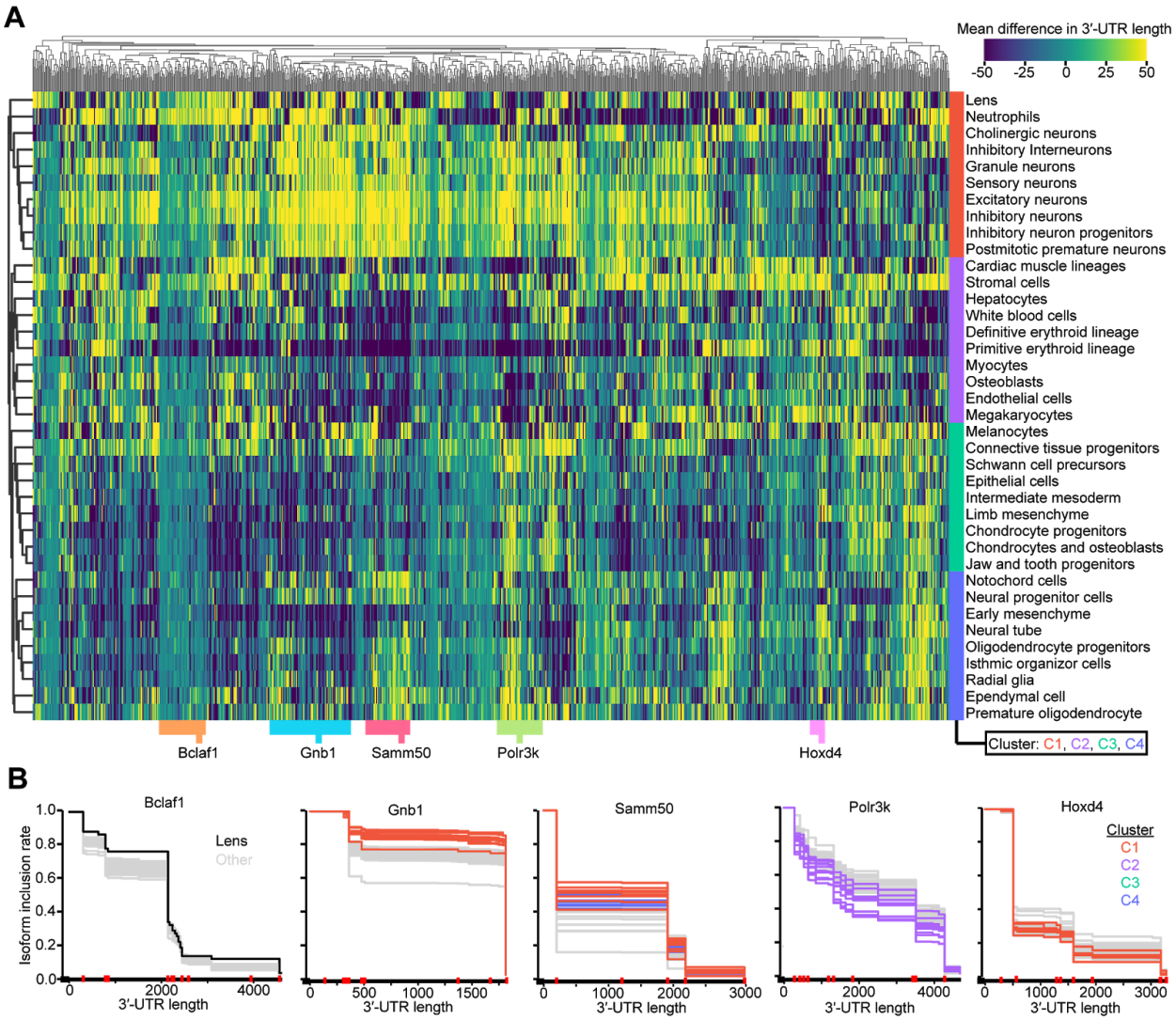

Figure 5. Identification of distinct gene lengthening patterns and genes responsible for overall data trends across embryonic ages. (A) Heatmap of mean differences in 3′-UTR lengths for 1,078 genes with significant differences in PAS usage across cell types derived from 38 t-SNE clusters. Heatmap is column-centered and clustered in both rows and columns by Pearson correlation as a distance metric. Row colors indicate the four dominant cell-type clusters. Column colors indicate five representative clusters chosen from the dendrogram above. **(B)** IIR plots for five genes among the representative clusters shown in panel (A), and colored by either individual cell types or cell-type clusters shown in panel (A). Grey lines allude to all other cell types. Vertical red lines along the 3′ UTR indicate PASs that are significantly different between cell types by the χ^2 test ($p<0.05$). See also **Figure 5—figure supplement 1** for clustering according to differences in entropy. See also **Figure 5—Source Data 1** for a table of read counts associated with each PAS for each gene and cell type.

DISCUSSION

Despite the rapid growth of single cell RNA sequencing data in recent years, the vast majority of analyses routinely overlook the phenomenon of alternative polyadenylation. Although scRNA-seq was initially developed to measure gene expression levels, multiple orthogonal forms of information are also effectively captured. For example, RNA velocity analysis, which estimates future transcriptome state by modeling intron/exon ratios, illustrates the ability to extract dynamical information about cellular differentiation (La Manno et al., 2018). In this work, we further develop a computational pipeline to quantify 3′ ends in scRNA-seq data by cross-referencing a novel integrated annotation set of 3′ UTRs and polyadenylation sites, thereby enabling a more spatiotemporally resolved understanding of APA. This pipeline closely recapitulates prior bulk measurements, yet further enables a closer spatiotemporal dissection of APA. Although the utility of scRNA-seq to give insight into APA has been recognized recently (Shulman and Elkon, 2019), we extend this line of work to an expansive atlas of cell types in a developmental time course spanning multiple stages of embryonic development (Cao et al., 2019).

Our findings reinforce the principle that the most proliferative cell types such as blood maintain shorter 3′ UTRs (Mayr and Bartel, 2009; Sandberg et al., 2008), on average, while lowly proliferative ones such as neurons maintain lengthened 3′ UTRs (Miura et al., 2013; Shepard et al., 2011). As differentiation progresses, cells of all types naturally become less proliferative, leading to an observed global lengthening of 3′ UTRs in all cell types (Ji et al., 2009; Ulitsky et al., 2012). A major functional consequence of this is that the global shortening of 3′ UTRs could

lead to the evasion of microRNA-mediated repression, resulting in greater mRNA stabilities across the transcriptome and enhanced protein synthesis rates in proliferative cells (Mayr and Bartel, 2009). In contrast to previous work, which often binarized the landscape of 3′ termini into proximal and distal isoforms due to a limited PAS annotation set (Ji et al., 2009; Miura et al., 2013), we develop more general metrics (*e.g.*, changes in mean length and entropy) that consider the relative proportions of the many PASs within each gene. While most genes obey a canonical pattern of 3′-UTR lengthening over time, a small subset of genes deviate from this trend. Most cell types can be grouped into one of four clusters that obey similar trends across genes. These observations are consistent with the evolution of regulatory mechanisms that act in a gene-specific and tissue-dependent manner (Di Giammartino et al., 2011; Tian and Manley, 2017).

Given the exaggerated 3′-UTR lengthening we observed in neuronal subtypes, it is interesting to consider how these observations might give insight into regulatory function in neurons. It has been previously observed that APA guides differential mRNA localization (Ciolli Mattioli et al., 2019; Taliaferro et al., 2016), and that APA itself is directly regulated by neural activity (Tushev et al., 2018) such as long-term potentiation (Fontes et al., 2017). These findings open the possibility that APA might serve as an important process in guiding mRNAs to axons and dendrites, thereby modulating synaptic potential. One promising direction for this work is to use our APA atlas, and those derived from other single cell datasets (Zeisel et al., 2018), to further dissect how differential mRNA localization across neuronal subtypes might contribute to their functional specialization. More generally, we anticipate that our characterization of APA across

genes, cell types, and developmental stages of a mammalian organism will serve as a resource to further guide the discovery of new regulatory mechanisms that control APA. It will also help to dissect how these changes impact the function of mRNA with respect to its cellular localization, half-life, and translation in cell types throughout the body.

METHODS

An integrated set of mouse 3′ UTRs

We established an integrated set of mouse 3′-UTR annotations for protein-coding genes in which each unique stop codon was associated with a representative transcript with the longest annotated 3′ UTR (Agarwal et al., 2015), using the Gencode M25 "comprehensive" set (Frankish et al., 2019) as our initial annotations (**Figure 1—Source Data 1**). For each unique stop codon, we selected the longest 3′ UTR from three additional resources: i) RefSeq (March 2020 release) (O'Leary et al., 2016), ii) 3′ UTRs with extreme lengthening (Miura et al., 2013), using liftOver (Navarro Gonzalez et al., 2020) to remap the coordinates from mm9 to mm10, and iii) bulk 3P-seq-based annotations derived from mouse muscle, heart, liver, lung, kidney, brain, testes, and white adipose tissues as well as NIH 3T3 and mESC cell lines (Nam et al., 2014). The choose_all_genes_for_TargetScan.pl Perl script in the TargetScanTools Github (Agarwal et al., 2018) was used to integrate these databases, allowing a 3P-seq read to exist up to 5,400nt (*i.e.*, the 99th percentile of annotated 3′-UTR lengths) downstream of a stop codon.

In certain scenarios, such as in the case of alternative splicing of the terminal exon, a gene is potentially associated with many unique stop codons, each of which is associated with its own

3′-UTR annotation. We therefore sought to avoid a bias in which genes with many such transcript isoforms would be overrepresented in the downstream results, and to avoid the possibility that PASs would be counted redundantly in cases in which multiple different 3′ UTRs overlapped the same genomic coordinates. We therefore carried forward a single transcript isoform with the greatest number of 3′-UTR mapping reads (or a random top-ranked one in the case of a tie) to represent each gene. To perform this counting, scRNA-seq reads were intersected with our 3′-UTR annotation set using bedtools intersect (-wa -wb -s) (Quinlan and Hall, 2010).

An integrated set of mouse poly(A) sites

To generate our union PAS set, we integrated three PAS annotation databases: Gencode M25 (Frankish et al., 2019), PolyA_DB v3 (Wang et al., 2018), and PolyASite 2.0 (Herrmann et al., 2020). First, PASs within ±10nt of another PAS within the same database were collapsed by selecting the most downstream PAS. Next, the following procedure was implemented to reduce redundancy between databases: i) we collected PASs from PolyASite 2.0, ii) we added PASs from PolyA_DB v3 not within ±10nt of the current PAS set, and iii) we added PASs from Gencode M25 not within ±10nt of the current PAS set. This method of sequential addition led to a total of 164,772 PASs in our union set; we provide the genomic coordinates and corresponding read counts associated with this set (**Figure 4—Source Data 1**, **Figure 5—Source Data 1**).

Calculation of 3′-UTR lengths, relative length differences, and corresponding visualizations

3′-UTR length corresponding to a given read was computed as the distance from the stop codon to the read's assigned PAS, minus the length of any intervening intron(s). These 3′-UTR lengths were used to compute a "gene by cell" sparse matrix of the mean length among all 3′-UTR isoforms, weighted by their respective counts. For each gene, we then computed each cell's deviation from the mean of 3′-UTR lengths across cells, considering only non-missing values. For heatmaps, these deviation values were then averaged according to the labels assigned to each cell (*i.e.*, with respect to t-SNE cluster, UMAP trajectory, and/or developmental stage). Cell labels were based upon those previously assigned (Cao et al., 2019). When indicated in the legend, in some instances the heatmaps were further centered by subtracting the mean of the row or column. t-SNE plots were visualized using the hexbin (gridsize = 500, vmin = -50, vmax = 50) function from pyplot, which averages values from cells captured in local bins. UMAP plots were binned by splitting each of the x, y, and z coordinates into 150 equally sized bins.

Gene-level isoform inclusion rate plots and corresponding statistics

For each gene, we counted reads assigned to each PAS to build contingency tables of counts for either (PAS by developmental stage) (**Figure 4—Source Data 1**) or (PAS by cell type) (**Figure 5—Source Data 1**). We then computed statistical significance using the χ^2 test as computed by the chisquare function in scipy, either with respect to the entire gene (**Supplementary File 1, Supplementary File 2**) or with respect to each PAS (axis = None or axis = 1, respectively). In both cases, we provided a matrix of expected counts, based on the joint probability of each cell multiplied with the total counts in the matrix. For the gene-level χ^2 test, we further derived a Benjamini-Hochberg (BH) based q-value to account for the FDR. Considering the read counts

associated with each PAS position, isoform inclusion rates were visualized in the same manner as previous work, which allude to this plotting style as the Affected Isoform Ratio (AIR) plot (Agarwal et al., 2018, 2015; Nam et al., 2014). Much like a survival curve, the IIR quantifies the proportion of 3′-UTR isoforms that include a given nucleotide position.

AUTHOR CONTRIBUTIONS

V.A. conceived of the study and designed analyses. S.L-D. performed the computational analyses and generated tables. V.A. and S.L-D. generated figures and wrote the paper with feedback from D.R.K. and J.S.

ACKNOWLEDGEMENTS

We thank Junyue Cao and other members of the Shendure Lab for discussions surrounding the single cell data as well as Nimrod Rubenstein for critical commentary. This material is based upon work supported under an NRSA NIH fellowship 5T32HL007093 (to V.A.). J.S. is an investigator of the Howard Hughes Medical Institute.

COMPETING INTERESTS STATEMENT

V.A. and D.R.K. are employees of Calico Life Sciences.

SUPPLEMENTAL TABLES

Figure 1—Source Data 1. GTF file of the gene models associated with the "Integrated 3′ UTR" set, as described in **Figure 1D**, along with the source database of the 3′-UTR annotation. Genomic coordinates are provided with respect to the mm10 genome.

Figure 4—Source Data 1. Table of read counts associated with each PAS for each gene and developmental stage. Genomic coordinates are provided for every PAS with respect to the mm10 genome, along with the annotation source of the PAS.

Figure 5—Source Data 1. Table of read counts associated with each PAS for each gene and cell type. Genomic coordinates are provided for every PAS with respect to the mm10 genome, along with the annotation source of the PAS.

Supplementary File 1. Table of FDR-corrected q-values for each transcript/gene tested for differential PAS usage across the five developmental stages. Genes that did not pass the threshold of having 100 reads at each age are listed as "Not Tested".

Supplementary File 2. Table of FDR-corrected q-values for each transcript/gene tested for differential PAS usage across the 38 cell types. Genes that did not pass the threshold of having 20 reads in each cell type are listed as "Not Tested".

SUPPLEMENTAL FIGURES

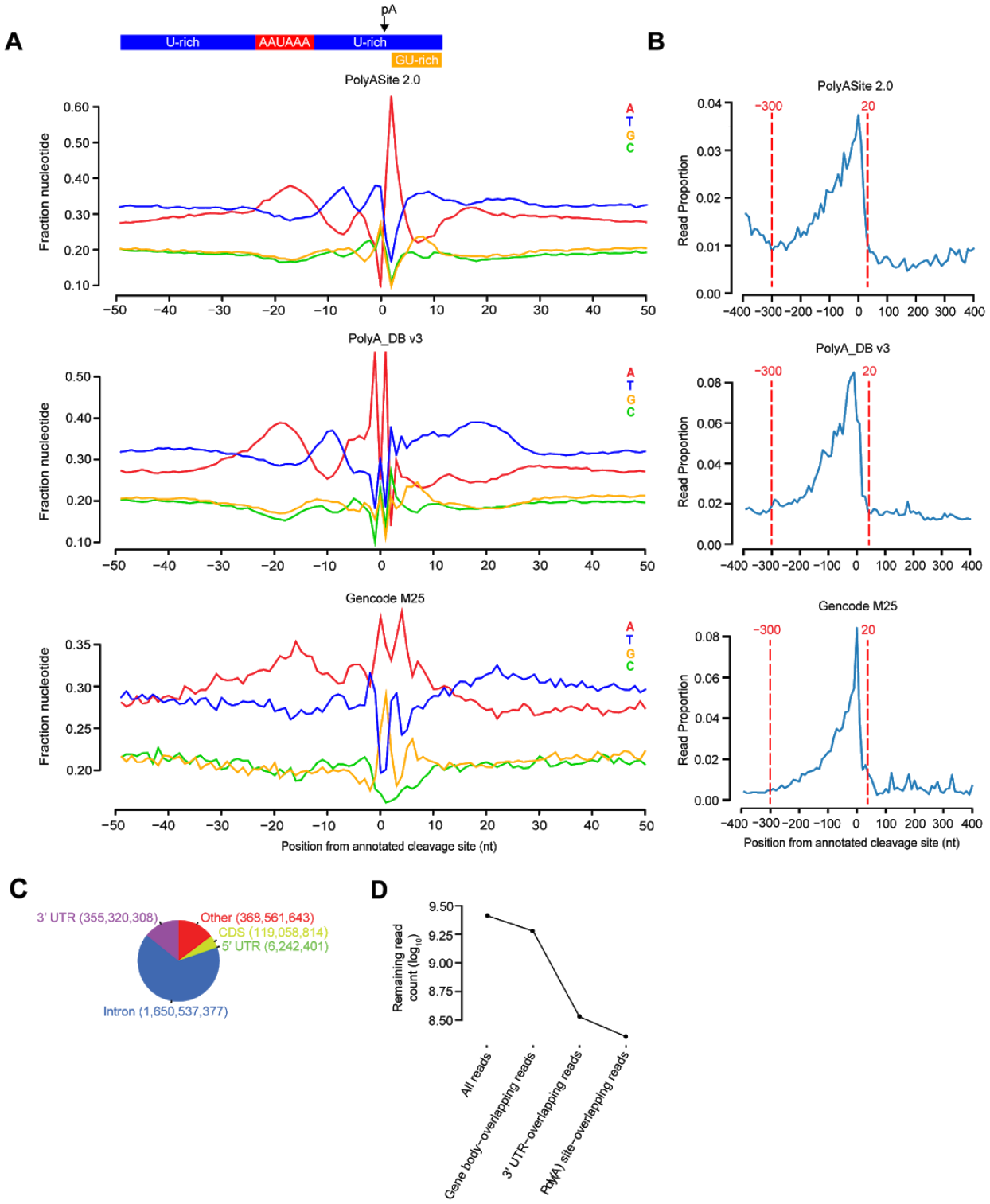

Figure 1—figure supplement 1. Characterization of PAS annotation databases and properties of scRNA-seq data read filtering. (A) This panel is the same as that shown in **Figure 1B**, except displays information for cleavage sites anchored on PASs unique to each of the three databases (*i.e.*,

non-intersecting with a PAS from any other database). **(B)** This panel is the same as that shown in **Figure 1C**, except displays information for cleavage sites anchored on PASs unique to each of the three databases (*i.e.*, non-intersecting with a PAS from any other database). **(C)** Pie chart showing the relative proportions of scRNA-seq reads that map to each functional region within the genome. **(D)** Plot of the decay in the numbers of reads remaining after each sequential filtering step that was required to isolate the subset of PAS-mapping reads.

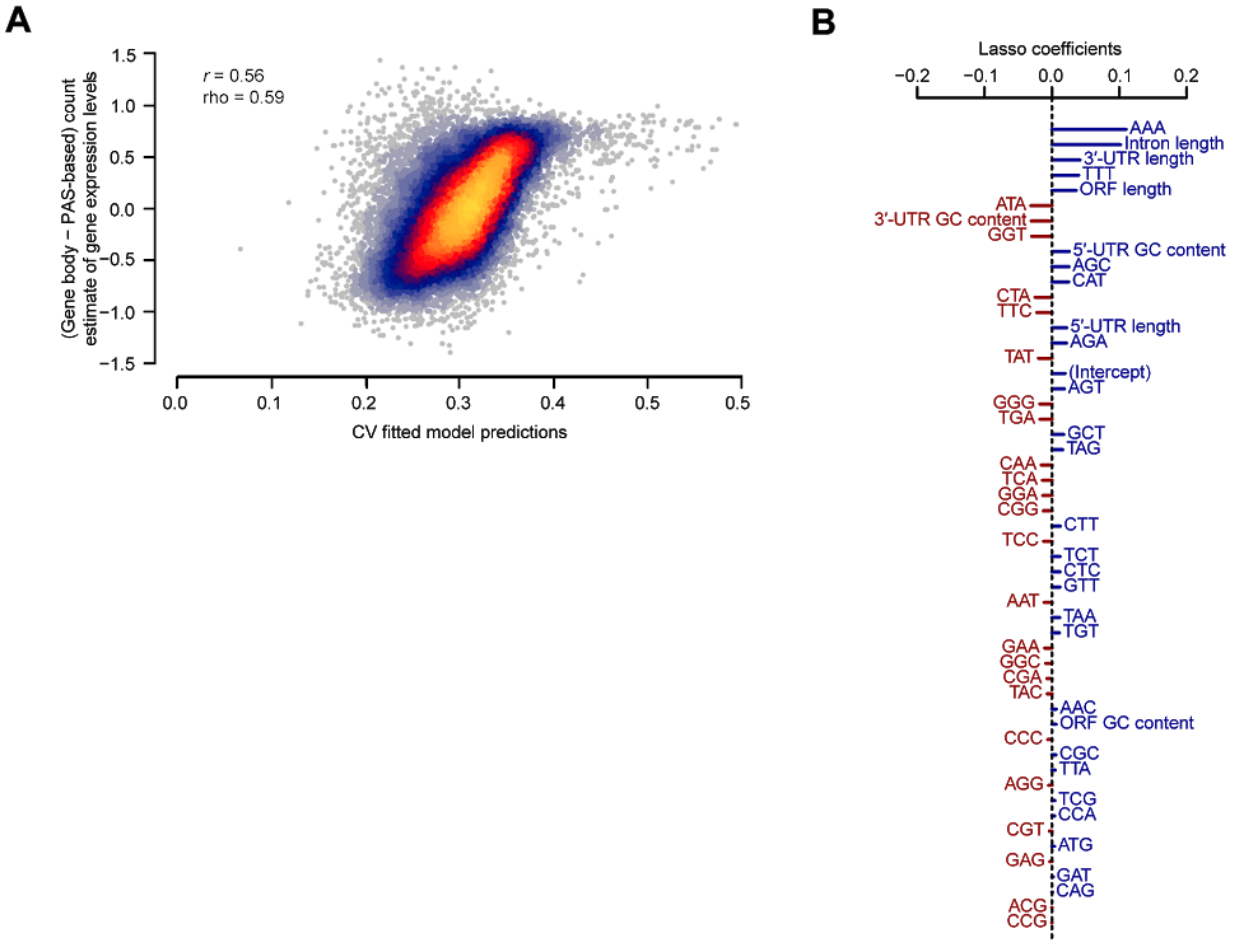

Figure 1—figure supplement 2. Sequence-based features partially explain biases in the gene-body-based method of estimating relative gene expression levels. We developed a regression model to predict the difference between the gene body and PAS counting methods to estimate gene expression levels, as shown on the x-axes of the left and right panels, respectively, of **Figure 1E**. The features considered in the model include the length of the 5′ UTR, ORF, introns, and 3′ UTR; the GC content of the 5′ UTR, ORF, and 3′ UTR; and the proportions of each of 64 possible 3-mers within the entire gene body. All features were z-score transformed to enable comparisons between regression coefficients. Following our previous work (Klein et al., 2020), we trained a lasso regression model using these features. The strength of the regularization was controlled by a single λ parameter, which was optimized using 10-fold cross-validation for each training set using the *cv.glmnet* function of the *glmnet* library in R. **A)** Scatter plot displaying the relationship between the 10-fold cross-validated predictions derived from the lasso regression model and the observed difference between gene body and PAS-based counting methods of estimating gene expression level. Also indicated are the Pearson (*r*) and Spearman (rho) correlation values. **B)** The ranked coefficients derived from a lasso regression model trained on the full dataset. Positive coefficients (blue) are associated with inflated gene body read counts, while negative coefficients (red) are associated with underrepresented gene body read counts.

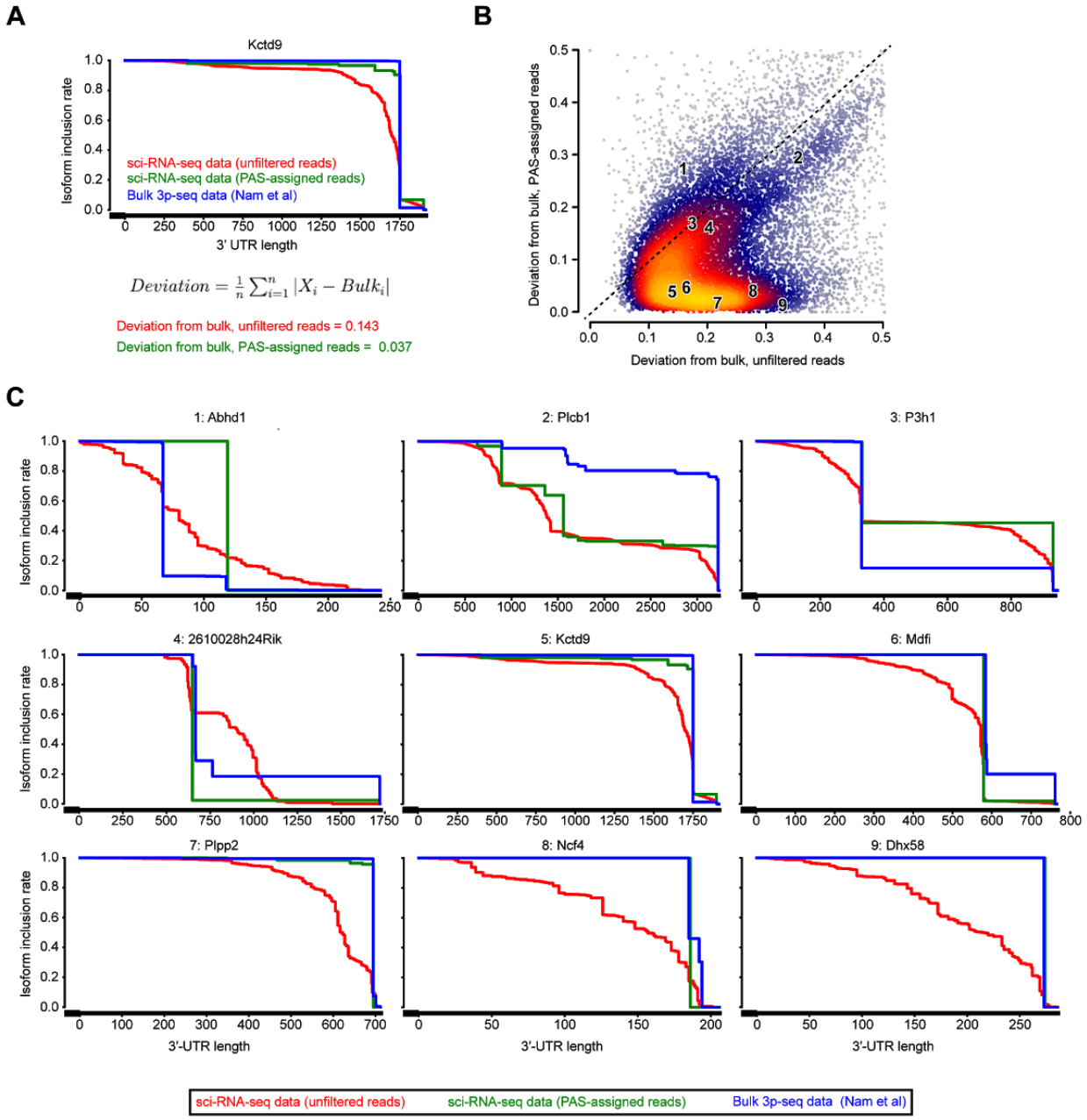

Figure 1—figure supplement 3. Improvement in the quantification of 3′-UTR isoform abundances after read-to-PAS assignment. (A) Calculation of the Mean Absolute Deviation (MAD) metric on an example gene to quantify the divergence between IIR profiles derived from either raw scRNA-seq data or post-processed data after read-to-PAS assignment, relative to the profile for bulk 3P-seq data (Nam et al., 2014) as a gold standard. Larger numbers indicate poorer agreement. **(B)** Scatter plot of MAD values for either raw scRNA-seq data (x-axis) or post-processed data after read-to-PAS assignment (y-axis) (n = 16,334 protein-coding genes). Regions are colored according to the density of data from light blue (low density) to yellow (high density). 78% of genes exist below the diagonal dotted line, indicating an improved similarity to bulk measurements after post-processing. Nine genes from different representative regions of the plot are indicated. **(C)** IIR plots for the nine representative genes numbered in panel (B). The majority of genes show strongly improved agreement to bulk 3P-seq data (genes 5-9).

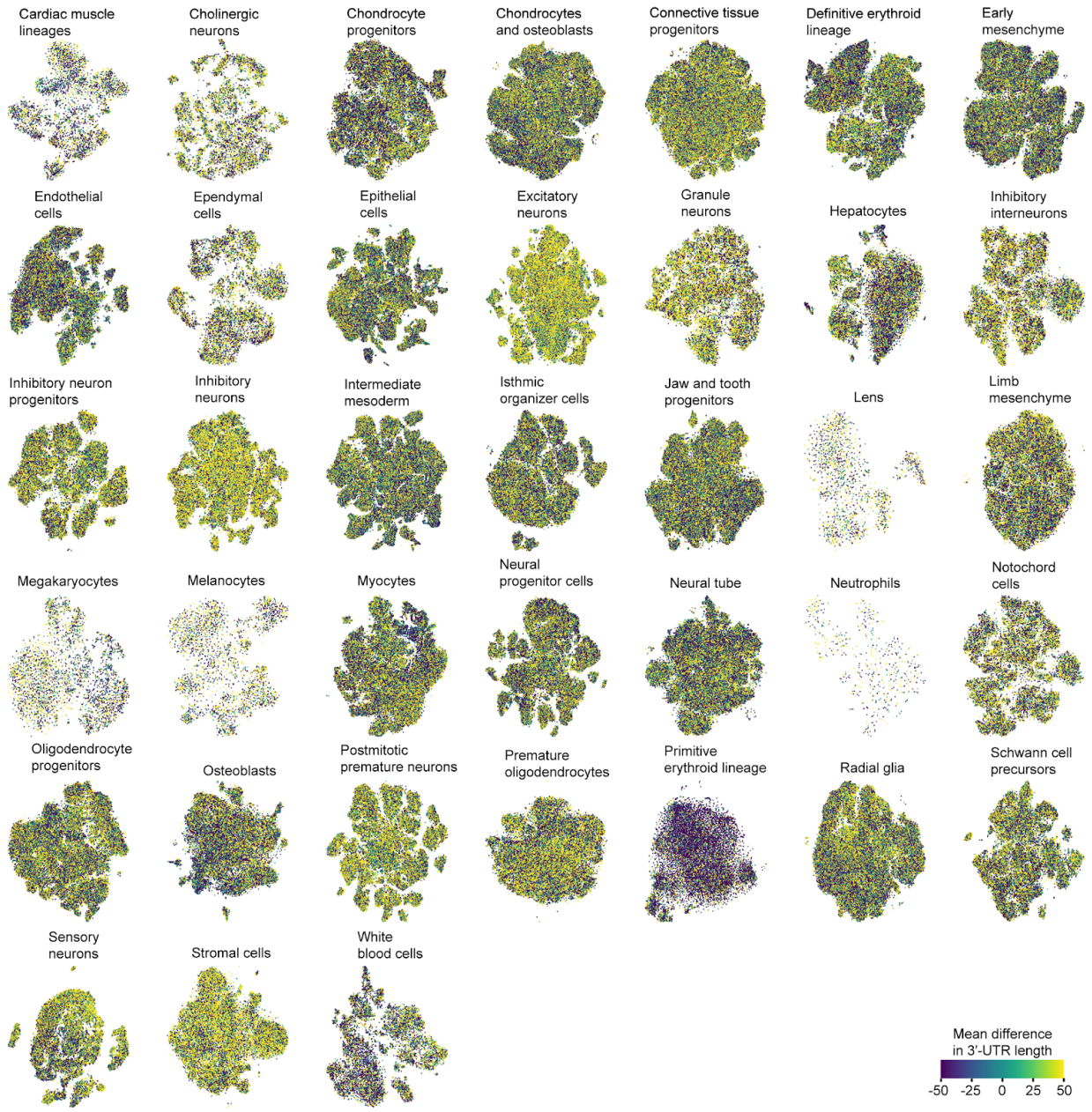

Figure 2—figure supplement 1. Analysis of differential 3′-UTR lengthening among cellular subtypes. t-SNE embeddings were generated to identify cellular subtypes for cell types derived from each of 38 t-SNE clusters (Cao et al., 2019). Each cell was colored according to the mean difference in 3′-UTR lengths across all genes. Local subcluster heterogeneity can be observed (*e.g.*, cellular subtypes in hepatocytes and osteoblasts), along with global differences between clusters (*e.g.*, the primitive erythroid lineage relative to neuronal cells and stromal cells).

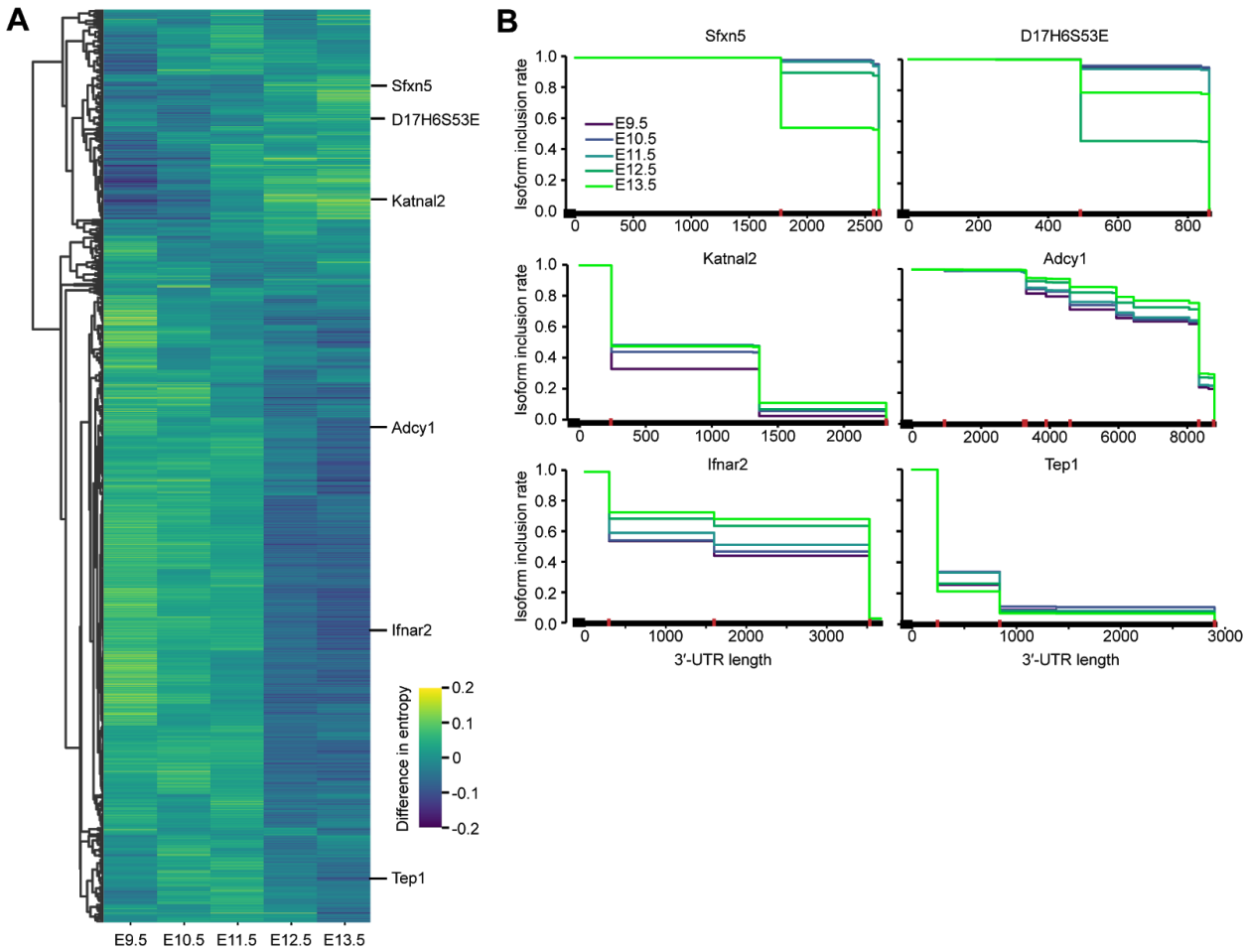

Figure 4—figure supplement 1. Evaluation of differential PAS usage and diversity across developmental time. (A) Heatmap of differences in entropy for the statistically significant genes from panel (A). Entropy for a given gene and developmental stage was calculated as $H(x) = \left(- \sum_{x \in PAS} P(x)\, log_2\, P(x) \right)$ using relative read proportions assigned to each PAS for the set of all PASs associated with the gene. Heatmap is row-centered and clustered by Pearson correlation as a distance metric. Higher values of entropy represent a greater degree of randomness and uniformity in selection among multiple PASs; conversely, lower values represent greater fidelity in selection amongst fewer PASs. **(B)** IIR plots for six genes among representative clusters shown in panel (B), and colored by developmental stage. Vertical red lines along the 3′ UTR indicate PASs that are significantly different between stages by the χ^2 test ($p<0.05$).

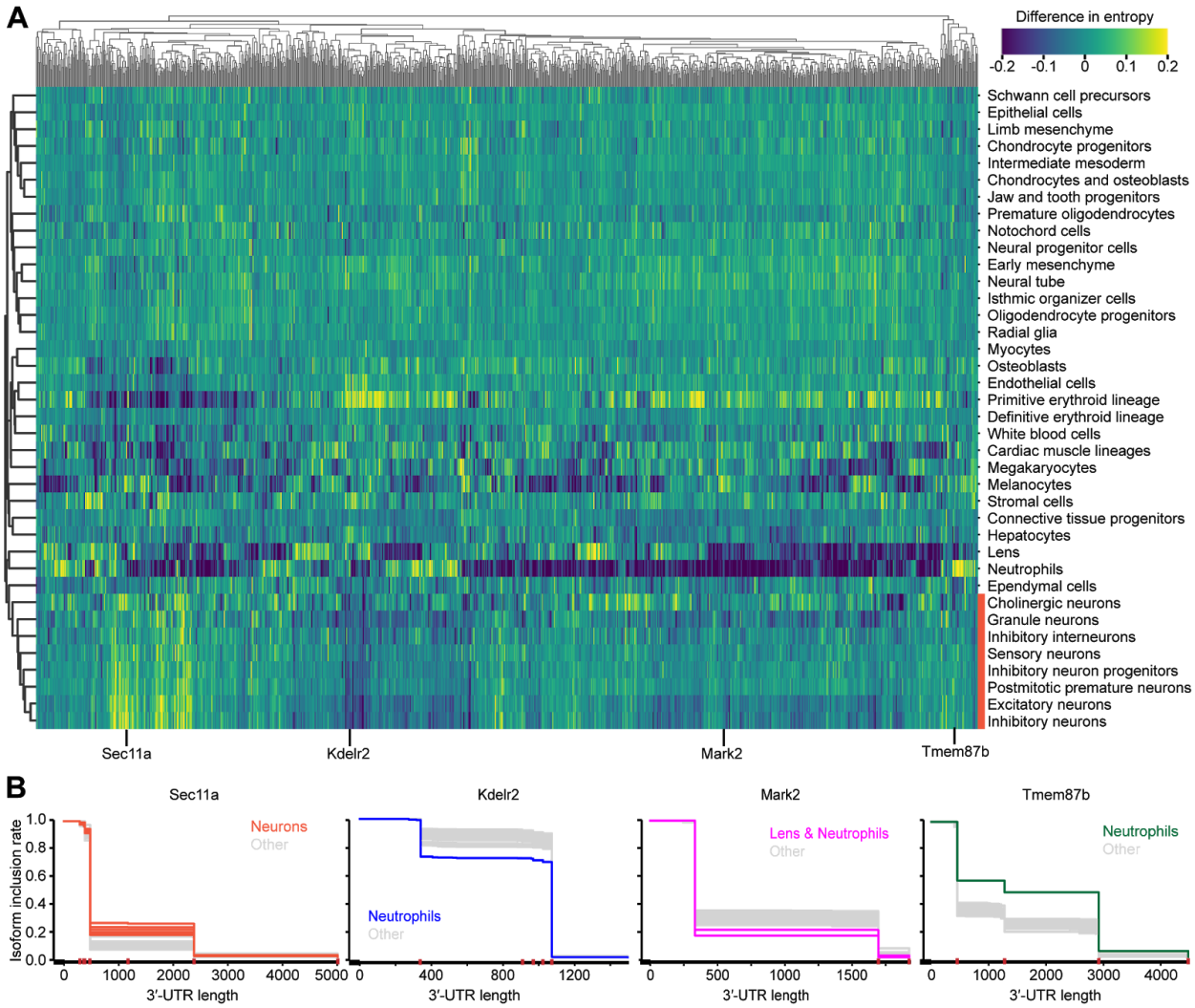

Figure 5—figure supplement 1. Evaluation of differential PAS usage and diversity across cell types. **(A)** Heatmap of differences in entropy for the statistically significant genes from panel (A). Entropy for a given gene and cell type (derived from each of 38 t-SNE clusters) was calculated as $H(x) = \left(- \sum_{x \in PAS} P(x) \, log_2 \, P(x) \right)$ using relative read proportions assigned to each PAS for the set of all PASs associated with the gene. Heatmap is column-centered and clustered in both rows and columns by Pearson correlation as a distance metric. Higher values of entropy represent a greater degree of randomness and uniformity in selection among multiple PASs; conversely, lower values represent greater fidelity in selection amongst fewer PASs. **(B)** IIR plots for five genes among representative clusters shown in panel (B), and colored either by the indicated cell types or the cluster of neuronal cell types shown in panel (B). Grey lines allude to all other cell types. Vertical red lines along the 3′ UTR indicate PASs that are significantly different between cell types by the χ^2 test ($p<0.05$).

REFERENCES

Agarwal V, Bell GW, Nam J-W, Bartel DP. 2015. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**. doi:10.7554/eLife.05005

Agarwal V, Shendure J. 2020. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep* **31**:107663.

Agarwal V, Subtelny AO, Thiru P, Ulitsky I, Bartel DP. 2018. Predicting microRNA targeting efficacy in Drosophila. *Genome Biology*. doi:10.1186/s13059-018-1504-3

Beck AH, Weng Z, Witten DM, Zhu S, Foley JW, Lacroute P, Smith CL, Tibshirani R, van de Rijn M, Sidow A, West RB. 2010. 3'-end sequencing for expression quantification (3SEQ) from archival tumor samples. *PLoS One* **5**:e8768.

Berkovits BD, Mayr C. 2015. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**:363–367.

Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, Adey A, Waterston RH, Trapnell C, Shendure J. 2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. doi:10.1126/science.aam8940

Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, Trapnell C, Shendure J. 2019. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**:496–502.

Ciolli Mattioli C, Rom A, Franke V, Imami K, Arrey G, Terne M, Woehler A, Akalin A, Ulitsky I, Chekulaeva M. 2019. Alternative 3' UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. *Nucleic Acids Res* **47**:2560–2573.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**:1173–1183.

Di Giammartino DC, Nishida K, Manley JL. 2011. Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**:853–866.

Elkon R, Drost J, van Haaften G, Jenal M, Schrier M, Oude Vrielink JAF, Agami R. 2012. E2F mediates enhanced alternative polyadenylation in proliferation. *Genome Biol* **13**:R59.

Elkon R, Ugalde AP, Agami R. 2013. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat Rev Genet* **14**:496–506.

Fontes MM, Guvenek A, Kawaguchi R, Zheng D, Huang A, Ho VM, Chen PB, Liu X, O'Dell TJ, Coppola G, Tian B, Martin KC. 2017. Activity-Dependent Regulation of Alternative Cleavage and Polyadenylation During Hippocampal Long-Term Potentiation. *Sci Rep* **7**:17377.

Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Carbonell Sala S, Chrast J, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Hunt T, Izuogu OG, Lagarde J, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Ruffier M, Schmitt BM, Stapleton E, Suner M-M, Sycheva I, Uszczynska-Ratajczak B, Xu J, Yates A, Zerbino D, Zhang Y, Aken B, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B,

Reymond A, Tress ML, Flicek P. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**:D766–D773.
Fu Y, Sun Y, Li Y, Li J, Rao X, Chen C, Xu A. 2011. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res* **21**:741–747.
Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, Gennert D, Li S, Livak KJ, Rozenblatt-Rosen O, Dor Y, Regev A, Yanai I. 2016. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* **17**:77.
Herrmann CJ, Schmidt R, Kanitz A, Artimo P, Gruber AJ, Zavolan M. 2020. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing. *Nucleic Acids Res* **48**:D174–D179.
Hoque M, Ji Z, Zheng D, Luo W, Li W, You B, Park JY, Yehia G, Tian B. 2013. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat Methods* **10**:133–139.
Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I. 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**:776–779.
Jan CH, Friedman RC, Graham Ruby J, Bartel DP. 2011. Formation, regulation and evolution of Caenorhabditis elegans 3′UTRs. *Nature*. doi:10.1038/nature09616
Ji Z, Lee JY, Pan Z, Jiang B, Tian B. 2009. Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci U S A* **106**:7028–7033.
Klein JC, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, Ahituv N, Shendure J. 2020. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat Methods* **17**:1083–1091.
La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, Fan J, Borm LE, Liu Z, van Bruggen D, Guo J, He X, Barker R, Sundström E, Castelo-Branco G, Cramer P, Adameyko I, Linnarsson S, Kharchenko PV. 2018. RNA velocity of single cells. *Nature* **560**:494–498.
Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27**:2380–2396.
Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**:1202–1214.
Mayr C. 2016. Evolution and Biological Roles of Alternative 3'UTRs. *Trends Cell Biol* **26**:227–237.
Mayr C, Bartel DP. 2009. Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**:673–684.
Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC. 2013. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Res* **23**:812–825.
Nam J-W, Rissland OS, Koppstein D, Abreu-Goodger C, Jan CH, Agarwal V, Yildirim MA, Rodriguez A, Bartel DP. 2014. Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol Cell* **53**:1031–1043.
Navarro Gonzalez J, Gonzalez JN, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney

BJ, Powell CC, Nassar LR, Maulding ND, Lee CM, Lee BT, Hinrichs AS, Fyfe AC, Fernandes JD, Diekhans M, Clawson H, Casper J, Benet-Pagès A, Barber GP, Haussler D, Kuhn RM, Haeussler M, James Kent W. 2020. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Research*. doi:10.1093/nar/gkaa1070

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**:D733–45.

Pijuan-Sala B, Griffiths JA, Guibentif C, Hiscock TW, Jawaid W, Calero-Nieto FJ, Mulas C, Ibarra-Soria X, Tyser RCV, Ho DLL, Reik W, Srinivas S, Simons BD, Nichols J, Marioni JC, Göttgens B. 2019. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*. doi:10.1038/s41586-019-0933-9

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842.

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. *Science* **320**:1643–1647.

Semrau S, Goldmann JE, Soumillon M, Mikkelsen TS, Jaenisch R, van Oudenaarden A. 2017. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat Commun* **8**:1096.

Shepard PJ, Choi E-A, Lu J, Flanagan LA, Hertel KJ, Shi Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**:761–772.

Shulman ED, Elkon R. 2019. Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic Acids Res*. doi:10.1093/nar/gkz781

Spies N, Burge CB, Bartel DP. 2013. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res* **23**:2078–2090.

Taliaferro JM, Vidaki M, Oliveira R, Olson S, Zhan L, Saxena T, Wang ET, Graveley BR, Gertler FB, Swanson MS, Burge CB. 2016. Distal Alternative Last Exons Localize mRNAs to Neural Projections. *Mol Cell* **61**:821–833.

Tian B, Graber JH. 2012. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdisciplinary Reviews: RNA*. doi:10.1002/wrna.116

Tian B, Manley JL. 2017. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**:18–30.

Tian B, Manley JL. 2013. Alternative cleavage and polyadenylation: the long and short of it. *Trends Biochem Sci* **38**:312–320.

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

Ulitsky I, Shkumatava A, Jan CH, Subtelny AO, Koppstein D, Bell GW, Sive H, Bartel DP.

2012. Extensive alternative polyadenylation during zebrafish development. *Genome Res* **22**:2054–2066.

Wang R, Nambiar R, Zheng D, Tian B. 2018. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res* **46**:D315–D319.

Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, Häring M, Braun E, Borm LE, La Manno G, Codeluppi S, Furlan A, Lee K, Skene N, Harris KD, Hjerling-Leffler J, Arenas E, Ernfors P, Marklund U, Linnarsson S. 2018. Molecular Architecture of the Mouse Nervous System. *Cell* **174**:999–1014.e22.

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH. 2017. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**:14049.