Identifying genes with cell-type-specific alternative polyadenylation in multi-cluster single-cell transcriptomics data

Yulong Bai[1], Yidi Qin[1], Zhenjiang Fan[2], Robert M. Morrison[5,6,7], KyongNyon Nam[3], Hassane Mohamed Zarour[5,6], Radosveta Koldamova[3], Quasar Saleem Padiath[1,4], Soyeon Kim[7,8†], Hyun Jung Park[1†]

[1] Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, USA

[2] Department of Computer Science, School of Computing and Information, University of Pittsburgh, Pittsburgh, USA

[3]Department of Environmental and Occupational Health, Graduate school of Public Health, University of Pittsburgh, Pittsburgh, USA

[4]Department of Neurobiology, School of Medicine, University of Pittsburgh, Pittsburgh, USA

[5]Department of Medicine and Division of Hematology/Oncology, University of Pittsburgh, School of Medicine, Pittsburgh, USA

[6]Department of Immunology, University of Pittsburgh, School of Medicine, Pittsburgh, USA

[7]Department of Computational and Systems Biology, University of Pittsburgh Medical Center, Pittsburgh, USA

[8]Department of Pediatrics, University of Pittsburgh Medical Center, Pittsburgh, USA

[9]Division of Pulmonary Medicine, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, Pennsylvania, USA

[†]senior author; Correspondence: hyp15@pitt.edu (H.J.P.)

ABSTRACT

Alternative polyadenylation (APA) causes widespread shortening or lengthening of the 3′-untranslated region (3′-UTR) of genes across multiple cell types (dynamic APA). Bioinformatic tools have been developed to identify dynamic APA in single cell RNA-Seq (scRNA-Seq) data, but they suffer from low power and cannot identify APA genes specific to each cell type (cell-type-specific APA) when multiple cell types are analyzed. To address these limitations, we develop a model-based method, scMAPA. scMAPA quantifies 3′-UTR long and short isoforms without posing assumptions on the signal shape of input data, enabling a sensitive identification of APA genes. In human peripheral blood mono cellular data, this enhanced power identifies unique associations of dynamic APA with hematological processes, e.g. progenitor cell development. Further, scMAPA identifies APA genes specific to each cell type while controlling

confounders using a sophisticated statistical model. In mouse brain data, scMAPA identifies APA genes specific to each cell type and provides a novel implication of neuron-specific APA genes in the interaction between neurons and blood vessels. Altogether, scMAPA sheds novel insights into the function of cell-type-specific APA dynamics in complex tissues.

Keywords: post-transcriptional regulation, alternative polyadenylation, single-cell RNA

INTRODUCTION

The majority of mammalian messenger RNAs contain multiple polyadenylation (pA) sites, such as proximal and distal, in their 3′-untranslated region (3′-UTR) [1,2]. By transcribing with different pA sites, alternative polyadenylation (APA) produces distinct isoforms with different lengths of the 3′-UTRs (long and short 3′-UTR isoforms using distal and proximal pA sites, respectively). These APA events are involved in diverse physiological and pathological processes (reviewed in [3]). For example, global 3′-UTR shortening events promote tumorigenesis by removing microRNA binding sites in certain types of cancer[4–6]. Notably, these events occur in tissue-specific and cell-type-specific manners [1,7]. To identify tissue-specific and cell-type-specific APA genes, single-cell RNA sequencing (scRNA-Seq) data provide an excellent resource, since scRNA-Seq data allow us to collect transcriptome of the same type of cells.

In scRNA-Seq data, several tools have been developed to identify genes with APA events (APA genes), such as scDAPA[8], Sierra [9] and scAPA[10]. However, they show several limitations to identify APA genes specific to each cell type (cell-type-specific APA genes). First, these methods are based on assumptions on the signal shape in input RNA-Seq data. For scAPA and Sierra, the assumptions are to differentiate signals from noises in their peak calling approaches. Since several scRNA-Seq utilize 3′ selection and enrichment steps in library construction, accumulation of the reads that originate from a common pA site forms a peak. To quantify the reads from the peaks, they assume a particular shape and length of the signal distribution. Likewise, scDAPA assumes the size of the window it uses to split gene regions to compare the difference of read coverage near 3′ end. Due to the assumptions, these methods are not guaranteed to identify APA events for the genes that do not hold the assumptions. Second, scDAPA and Sierra identify APA genes mainly between two cell clusters and are not directly applicable for scRNA-Seq data typically of more than two clusters. While scAPA is the only method to identify APA genes in more than two clusters, it still shows several limitations. First, to identify APA genes, scAPA statistically tests if the APA usage (the ratio of long and short 3′-UTR isoforms) of each gene is similar across cell clusters. However, it does not estimate significance in which cell clusters the genes undergo APA events, which direction (3′-UTR shortening or lengthening), and to what degree the APA events occur for multiple cell types in a statistically consistent way. These identifications are essential if one is interested in identifying cell-type-specific functions of APA. Second, since scAPA employs a statistical test that explicitly stratifies the input samples (e.g. case vs. control), it cannot directly control confounding factors that would exist across the stratification. Confounding arises when cells are affected by factors that are not parts of the research hypothesis under investigation. Since

multiple factors affect the molecular dynamics of complex tissues, it is important to control confounders to study complex tissues. For example, since brain transcriptome is known to be specific to regions (e.g. cortex and dorsal midbrain) and cell types (e.g. neuron and astrocyte) [11–13], one may need to control brain region as confounders depending on how brain cells are clustered and the research question. To address these limitations and identify cell-type-specific APA events, we developed a statistical method, scMAPA.

RESULTS

Alternative Polyadenylation identification across multiple cell clusters of single-cell RNA-Seq data (scMAPA)

For accurate APA gene identification, scMAPA quantifies 3′UTR long and short isoforms without posing assumptions on the signal shape of input scRNA-Seq data. While all current methods operate based on assumptions on the RNA-Seq signal shape, these assumptions are not guaranteed to hold for all genes. For example, in the scRNA-Seq data on Peripheral Blood Monocellular Cells (PBMC) of a healthy donor (10k in https://www.10xgenomics.com/), FLT3 3′ tags form peaks with different shapes and lengths between pDC/HSPC and B/NK cells (**S. Fig. 1A**), complicating the quantification process. To quantify the 3′UTR isoforms without such assumptions, we reasoned that each 3′ biased read represents the 3′ end part of a transcript. With this reasoning, we will pad each read along the 3'UTR region up to where the read ends (step 1 in **Fig. 1**, see Methods). This padding transforms each 3′ biased read to represent the full-length 3′UTR of the transcript. Then, this transformation reveals different pA usage between pDC/HSPC and B/NK cells without an assumption on the signal shape (**S. Fig. 1B**). Further, this transformation allows us to use approaches developed for bulk RNA-Seq data, since most bulk RNA-Seq data represent the full-length 3′UTRs. Among multiple methods for bulk RNA-Seq data, we employ the approaches of DaPars[14] due to the following reasons. First, DaPars maintains highest sensitivity and specificity compared to other methods in benchmark tests on biological and simulation data [15]. Second, DaPars identifies APA genes without any assumption on the RNA-Seq signal density by determining 3'UTR isoforms such that the difference between the sum of the isoforms and the input RNA-Seq signal density is minimized (step 2 in **Fig. 1**).

To call cell-type-specific APA genes based on the quantification, scMAPA builds a statistical model (step 3 in **Fig. 1**). This statistical model brings up three advantages over current methods. First, scMAPA identifies APA genes across multiple cell clusters (multi-cluster) of scRNA-Seq data. Between scMAPA and scAPA, the only methods for multi-cluster setting, our simulation experiments and biological data analyses show that scMAPA successfully transfers the sensitivity of DaPars approach validated in bulk RNA-Seq data [15]. Second, it identifies cell clusters in which the APA genes show distinct APA patterns with statistical significance (step 4 in **Fig. 1**). Although this is critical to study the impact of APA events for each cell type, no current methods provide a direct and solid statistical technique. Third, it provides flexibility to consider not only cell types but also potential confounders, such as tissue type, age, or sex.

Although considering confounders is critical in identifying APA genes from complex tissues, no current methods provide this function.

scMAPA identifies true APA events with an enhanced statistical power

To compare statistical power of scMAPA with scAPA, the only other method developed for multiple cell clusters, we developed a novel simulation platform. This platform simulates 3′-UTR long and short isoforms of APA and non-APA genes by generating the isoform proportions and the gene expressions in reference to a biological data. In our simulation, we set as our reference data the mouse brain scRNA-Seq data consisting of five main cell types (neurons, astrocytes, immune cells, oligodendrocytes and vascular) that are sampled from mouse cortex and dorsal midbrain regions [16]. On the data, we ran both scMAPA and scAPA and identified APA and non-APA genes by intersecting their results (step 1 in **Fig. 2A**, see Methods). In the APA and non-APA genes, we evaluated the proportion of the long and short isoforms (step 2 in **Fig. 2A**) and calculated the standard deviation (SD) of the proportions across the five cell clusters. We will refer to this SD as $SD_{isoprop}$ (step 3 in **Fig. 2A**). We found that the APA genes have significantly higher $SD_{isoprop}$ (p-value < 2.2×10^{-16}) than the non-APA genes (0.127 vs. 0.009 on average, **S. Fig. 2A**), indicating that the APA genes have wider distributions of the 3′-UTR isoform proportions across the clusters than non-APA genes. In reference to these $SD_{isoprop}$ values, we generated the isoform proportions for 500 APA and 4,500 non-APA genes across 5 clusters, each of 600 cells. For the APA genes, we randomly generated the 3′-UTR isoform proportions across the 5 clusters based on a single $SD_{isoprop}$ value ranging from 0.06 to 0.18 (step 4 in **Fig. 2A**). For non-APA genes, we fixed $SD_{isoprop}$ to be 0.009 in generating the 3′-UTR isoform proportions. Equally for APA and non-APA genes, we simulated the gene expression in Splatter [17] by determining the parameters in reference to the mouse brain data (step 5 in **Fig. 2A**). The gene expressions are then divided into 3′-UTR long and short isoform abundances based on the 3′-UTR isoform proportions simulated (step 6 in **Fig. 2A**). On these simulated isoform abundances of APA and non-APA genes, we ran the statistical component of scAPA and scMAPA, which is Pearson's χ^2 and Regression + LRT (likelihood ratio test), respectively. Across all $SD_{isoprop}$ values simulated for APA genes (ranging from 0.06 to 0.18), the statistical component of scMAPA consistently identifies more true APA genes than that of scAPA (**Fig. 2B**). Since both methods perform equally good at identifying true non-APA genes (**Fig. 2C**), scMAPA outperforms scAPA overall.

While the above simulation fixed the number of APA and non-APA genes and the cell cluster sizes, we then ran other simulations by varying the number of APA and non-APA genes and the cell cluster size while fixing $SD_{isoprop}$ values for APA and non-APA genes (to 0.127 and 0.09, respectively). With 500 (10% of the total genes) true APA genes, the statistical component of scMAPA consistently outperforms that of scAPA in terms of sensitivity (**Fig. 2D**) in all three cluster size distributions with a slight loss of specificity (**Fig. 2E**). This trend holds true with 250 and 1,000 true APA genes simulated (**S. Fig. 2 B, C, D, E**).

scMAPA identification is accurate while consistent with other methods

Further, we evaluated scMAPA's accuracy using biological data in comparison to other APA detection methods, scDAPA, scAPA and Sierra. In the three PBMC data sets with different numbers of cells (also known as 5k and 10k data representing the number of cells), we defined different numbers of cell clusters (8 and 13 clusters respectively) based on Seurat's graph-based clustering [18] and annotated their cell types based on established marker genes [19] (see Methods, **S. Table 1**). First, scMAPA identifies the highest proportion of the pA sites in proximity to the known pA sites annotated in PolyASite 2.0[20] across different degrees of proximity (**Fig. 3A** for 10k and **S. Fig. 3A** for 5k). scDAPA was not included in this comparison, because it does not return results that are compatible for the comparison, such as pA peaks, sites, or intervals. Second, scMAPA results substantially overlap with the results of the other methods. To assess the overlap, we identified significant APA genes across the cell clusters in scMAPA and scAPA. Also, since scDAPA and Sierra identify APA genes only between cell cluster pairs, we combined the pairwise significant APA genes in each method separately. After controlling FDR on the combined APA genes, we called APA genes if they are significant in any of the pairwise identifications. While scMAPA identifies an intermediate number of APA genes between scDAPA and Sierra/scAPA (10k in **Fig. 3B** and 5k in **S. Fig. 3B**), more than half of the scMAPA's findings are found in other methods (59.9% for 10k and 51.9% for 5k). While scMAPA is the only method of 'change-point' approach based on the padding of 3′ biased reads (step 1 in **Fig. 1C**), the number of APA genes identified by scMAPA and its high overlaps with other methods validate the use of scMAPA. Further, we focused our comparison on scMAPA and scAPA. In the 10k PBMC data, scMAPA and scAPA identifies 3,465 and 325 significant APA genes respectively, with 109 found in common. To test whether 3,356 APA genes unique to scMAPA are due to its high statistical power or high false positive, we inspected $SD_{isoprop}$ value of the APA genes identified by scMAPA and scAPA, separately. Their identification results differ largely in the range of $SD_{isoprop}$ values (**Fig. 3D**) where scMAPA is more sensitive in our simulation study (**Fig. 2A**), suggesting scMAPA's high statistical power. Altogether, the results suggest that scMAPA identification is accurate and sensitive while consistent with other methods, although it employs a novel approach than other methods.

scMAPA identification is robust, facilitating the understanding of APA dynamics

We further studied robustness of scMAPA in comparison to the other methods. First, we checked the overlap of APA genes across different cell numbers of the PBMC data (10k, 5k, and 1k, **Fig. 3C** and **S. Fig. 3C**). To test the overlap with more variability, we added the 1k data in this analysis. scMAPA identifies 1,651 APA genes in all three data sets, which is 40.1% of the total number of APA genes identified in any of the data sets (4,059). Since the cell types are similar across the data sets (**S. Table 1**) and represent healthy adults, the APA genes are expected to overlap across the data. On the other hand, scAPA, scDAPA, and Sierra identify less than 20% of the total APA genes in all three data (18.9% (82/435), 11.6% (719/6,192), 16.9% (668/3,953), **S. Fig. 3C**), suggesting that scMAPA identification is most robust to the data size. Second, we sampled different numbers of cell clusters from the 13 clusters of the 10k data and evaluated how

many APA genes identified in the 13 clusters scMAPA can recover with different numbers of clusters. Based on 20 random combinations of each cluster size (5, 7, 9, and 11), scMAPA performance does not depend much on the number of clusters. For example, when 5 clusters were sampled, 70.4% of all the APA genes were identified (**S. Fig. 3D**). Further, scMAPA retrieves more of the APA genes identified in the 13 clusters as more clusters are sampled.

To demonstrate biological implications brought by the accurate and robust scMAPA identification, we ran Ingenuity Pathway Analysis (IPA) on 1,432 APA genes identified only by scMAPA that are not identified by any other methods (**S. Table 2, Fig. 3B**). Manual inspection of the genes demonstrated a different usage of pA sites across the clusters. For example, as FLT3 clearly showed a different usage of pA sites across the clusters (**S. Fig. 1 A, B**), it is included in the 1,432 scMAPA-unique APA genes. Further, GATA2 also showed different pA usages across the clusters and is included in the 1,432 scMAPA APA genes (**S. Fig. 1 C, D**). Interestingly, GATA2 was polyadenylated in the scRNA-Seq data of bone marrow mononuclear cell from acute myeloid leukemia patients[21]. Due to the developmental relationship between bone marrow and peripheral blood, GATA2 can undergo APA events also in the PBMC using similar molecular mechanisms. Collectively, the 1,432 APA genes are significantly enriched (B-H p-value < 0.05) for multiple IPA Disease & Function terms with implication for hematology developmental processes, including 9 with keyword "hemato" or "blood" (**Fig. 3E**). As "hemato" terms refer to diverse developmental processes of hematopoietic progenitor cells, previous reports on the role of APA in the hematopoietic stem cell differentiation[22] supports the use of scMAPA. Altogether, scMAPA enables accurate and robust identification of dynamic APA in complex tissues.

scMAPA identifies APA genes specific to cell types

To demonstrate how scMAPA further identifies APA genes specific to each cell type (cell-type-specific APA genes) in complex tissues, we analyzed the mouse brain scRNA-Seq data [16] that defined five cell types with large sample size: neurons, astrocytes, immune cells, oligodendrocytes and vascular (see Methods). Across the five cell types, scMAPA identified 3,223 significant APA genes (**S. Table 3**) which do not overlap much (1,048) with 2,494 genes scAPA identified as APA genes (**Fig. 4A**). The IPA analysis shows that the APA genes identified by both scMAPA and scAPA and uniquely by scMAPA are enriched for the IPA terms with keyword "neurology", while scAPA-unique APA genes are not enriched for any of the terms (**Fig. 4B, S. Table 4)**. Expression analysis further validates the functional implication of scMAPA APA genes. Globally, when 3,223 differentially expressed genes are calculated based on Seurat package, 1,018 of them are the scMAPA APA genes, showing a significant overlap (p-value < $2.2e^{-16}$ by hypergeometric test). This result is consistent with the potential effect of APA on differential expression that previous studies discussed[14,23]. Previous studies discussed that 3′-UTR shortening removes microRNA (miRNA) binding sites on the 3′-UTR and thus evade miRNA-mediated repression and 3′-UTR lengthening adds miRNA binding sites and thus enhance miRNA-mediated repression. We checked that our overlap is not because the scMAPA identification is biased toward highly expressed genes, as the p-values of the scMAPA APA

genes are not strongly correlated with their average CPM values (R^2=0.08). Thus, scMAPA identifies APA genes that are potentially functional not biased by high expression. Altogether, our analysis reaffirms that scMAPA identifies functional APA events in the scRNA-Seq data.

On the 3,223 scMAPA APA genes with functional implications, we further investigated which cell clusters the genes undergo APA events in which direction and to which degree. For each APA gene, this investigation estimates the coefficients representing the degree and the direction of APA events for each cluster (see Methods). Running hierarchical clustering on the scMAPA coefficients (**Fig. 4C**), we found that immune cells and neuron cells are most distinguished from the other cell types. This result is in line with the previous finding of scAPA that they are most different in the APA pattern [24]. Further, scMAPA revealed a large number of genes characterized with cell-type-specific 3′-UTR shortening and lengthening (**Fig. 4D**). Neuron cells are characterized with 3′-UTR lengthening, which is consistent with previous findings of the dominance of 3′-UTR lengthening in neuron cells [25–28]. By running IPA on them, we found that significantly enriched terms (B-H p-value < 0.05, **S. Table 5)** are related to the cell-type-specific biological function. For example, 438 neuron-specific APA genes are enriched for 11 IPA Disease and Bio Functions terms with keyword “blood” and “blood vessel”, such as Proliferation/Survival of blood cells and Area/Size of blood vessel, while other cells do not show as strong enrichment (**Fig. 4E**) with the terms. Neuron and blood cells interact to allow ready exchange of nutrients and waste products, enabling the high metabolic activity of the brain despite its limited intrinsic energy storage[29]. Although this interaction is believed to play critical function in maintaining the operational condition of brains, little is known as to how this highly dynamic process is tightly regulated. With the neuron-specific APA genes, APA is expected to contribute to the dynamic and tight regulation. Together with a general independence between the expression level and the scMAPA coefficients across all clusters (-0.003 in Spearman’s ρ on average across the five cell types. **S. Fig. 4**), scMAPA’s gene-cluster-level identification suggests dynamic APA as an additional layer for complex gene regulation mechanism.

scMAPA controls confounding factors

To demonstrate how scMAPA controls confounding factors and why it is critical in the APA analysis, we first split the mouse brain scRNA-Seq data by both cell type (neurons, immune cells, astrocytes, oligos, and vascular cells) and brain region (cortex and midbrain dorsal) information. Since the cell types and the brain regions are not perfectly matched (**S. Fig. 5A, B**), we quantified 3′-UTR long and short isoforms in each combination using scMAPA. With the quantified isoforms for 10 scRNA-Seq data from the combinations (5 cell types ×2 brain regions), we set scMAPA in two different runs. In the first run, we set only the cell type information as covariates, identifying 2,793 APA genes (**Fig. 5A, S. Table 6)**. In the second run, we set the cell type information as covariates and brain region information as confounders, identifying 2,715 APA genes (**S. Table 6**). Since the second run with the confounder would not identify genes whose APA usage is associated with brain regions, 113 genes included only in the first set would be associated with brain-region-specific function. To test if the 113 brain-region-specific APA genes indeed play roles specific to brain regions, we first mapped their homologs

in human and checked the Genotype-Tissue Expression (GTEx)[30] human samples that upregulated the homologs compared to other human samples (see Methods). Since the upregulated genes represent those playing roles in the samples, we hypothesized that the homologs are upregulated mainly in brain samples, especially brain cortex samples. To test our hypothesis, we ranked GTEx samples based on the overlap between the upregulated genes and the homolog genes using a database that curates the up- and down-regulated genes for each GTEx sample, Enrichr [31]. When Enrichr evaluates the overlap by combining p-value and odds ratio (Combined Score in Enrichr), the homologs are more up-regulated in GTEx brain samples than in non-brain (T-test statistic=22.7, p-value=$6.57e^{-105}$, **Fig. 5B)**. Further, they are more up-regulated in GTEx brain cortex samples than brain non-cortex samples (T-test statistic=5.08, p-value=$5.86e^{-7}$, **Fig. 5C**), demonstrating their brain region specificity. It is important to note that the homologs are not as significantly down-regulated both in brain vs. non-brain (T-test statistic=-6.5, p-value=$8.4e^{-11}$, **S. Fig. 5A)** and in brain cortex vs. brain non-cortex samples (T-test statistic=-1.0, p-value=0.29, **S. Fig. 5B**). To further analyze the implication in brain-region-specific functions, we ran Ingenuity Pathway Analysis (IPA) upstream regulator analysis separately on the 113 genes and on 2,715 other genes that are identified by the second run (2,680 found by both and 35 by only the second run, **S. Table 7)**. The result shows that the common upstream regulators for the 113 genes are indeed implied for brain-region-specific functions. For example, IL1B, LRRC4, and TREX1 are three of the most significant upstream regulators of the 113 genes. All three genes are expressed specifically in the cortex region[32–34]. Also, they are known to be heavily involved in brain development where their abnormal regulations are associated to neurological diseases [32–34]. Since this mouse brain data was collected from brain regions including cortex, APA events on the 113 genes would help understand how those upstream genes affect the APA events of the 113 genes for brain development. Altogether, scMAPA facilitates unbiased analyses by explicitly addressing confounders in the model.

DISCUSSION

In this work, we developed scMAPA that identifies APA genes in two novel ways. First, while all current methods operate with assumptions on the shape of input scRNA-Seq data, scMAPA quantifies 3′-UTR long and short isoforms without posing such assumptions, enabling an accurate and robust identification of APA genes. scMAPA outperforms existing methods in identifying APA genes in various simulation (**Fig. 2**) and the PBMC data (**Fig. 3)**. Especially, it is important to note that scMAPA makes point estimates for the pA sites. Although point estimations are more directly relevant than interval estimations for further analyses, e.g. conducting omics analyses and designing experiments, point estimation methods are generally disadvantageous in checking the distance with the annotated pA sites (**Fig. 3A** and **S. Fig. 3A**), because point estimation returns a single point to calculate the proximity while interval estimation returns two points (start and end of the interval). Still, scMAPA outperforms the interval estimation results of Sierra and scAPA, while the interval estimation results are better than point estimation results of Sierra and scAPA. Further, 1,432 APA genes identified in the PBMC data only by scMAPA, not by scAPA, scDAPA, or Sierra, suggest that the intricate

developmental process of hematopoietic progenitor cells may involve APA events in line with previous reports on the role of APA in the hematopoietic stem cell differentiation[22]. As the second novelty, it identifies APA genes specific to each cell type while controlling confounders using a sophisticated statistical model, enhancing interpretability of the APA genes. This novel analytical layer further elucidates cell-type-specific function of APA in the mouse brain data (**Fig. 4 and Fig. 5**). For example, 438 APA genes unique to neuron cells suggest its potential role for the interaction between neurons and blood vessels, which is critical to maintain the operational condition of brains.

With the improved accuracy/robustness and enhanced interpretability mentioned above, scMAPA is extendible in the following directions. First, scMAPA assumes two types of 3′-UTR isoforms, short and long that use proximal and distal pA sites respectively. It also assumes no pA sites on the introns. With recent works reporting genes with more than two 3′-UTR isoforms[20] and pA sites on the introns[3], we will extend scMAPA to model such cases. Second, since we transform 3′-biased reads to represent full-length 3′UTR of the transcripts, this transformation allows to use other established methods developed for bulk RNA-Seq data that work for full-length transcripts, such as APATrap[35], TAPAS[36], and DaPars[14]. Third, in the same sense, scMAPA is directly amenable for other scRNA-Seq data that are not 3′tag-based (e.g. Smart-seq2[37]). scMAPA is also applicable for bulk RNA-Seq data sets that are collected from multiple biological conditions.

Altogether, we developed a statistical method to identify APA genes in the multi-cluster setting. With high sensitivity and interpretability, scMAPA allows to understand cell-type-specific function of APA events, which is essential to shed novel insights into the function of dynamic APA in complex tissues.

METHODS

Processing data sets

PBMC data. Aligned BAM file were downloaded from the 10X genomics repository (https://support.10xgenomics.com/single-cell-gene-expression/datasets). According to the data description of 10X, 1K and 10K data were generated from same materials. 5K data was generated from different cells. PCR duplicates were removed using UMI-tools 1.0.0 with "--method=unique --extract-umi-method=tag --umi-tag=UB --cell-tag=CB". Cell clustering was performed using R package Seurat 3.1.4[18]. We filtered to keep cells with more than 1000 UMI counts and 500 genes expressed. Cells with more than 15% UMI counts from mitochondrial genes were filtered out. Then, raw data was normalized by regressing against UMI count, mitochondrial mapping percentage, and ribosome genes mapping percentage using SCTransform function. We ran PCA analysis and took top 20 principle components as input to FindNeighbors function. Finally, FindClusters function was run with resolution set to 0.2 to identify cell communities. Cell types were annotated by matching the expression pattern of well-known marker genes for PBMC [19].

Mouse brain data. Aligned BAM file and clustering result of cortex and midbrain dorsal from two donors were downloaded from [16]. PCR duplicates were removed using UMI-tools same parameters used for PBMC data. To keep consistent with analysis performed by scAPA, we included only neurons, immune cells, astrocytes, oligos, and vascular cells in our analysis. Differential expression analysis was performed by FindAllMarkers function of Seurat package with min.pct set to 0.25 and all other parameters as default.

Mouse human homology data. It is downloaded from Vertebrate homology database in the Mouse Genome Informatics (MGI) (http://www.informatics.jax.org/homology.shtml).

scMAPA algorithm

Filtering polyadenylation (pA) sites.

To make sure only genes with strong APA signal among multiple cell types are identified, we first filter out genes in which only 1 PA site is detected in less than 3 cell types. Then, for each gene, we calculate the CPM for long and short isoforms separately and average over all cell types. Only genes with average CPM larger than 10 for both long and short isoforms are kept. In addition to gene-wise filtering, we also apply cell-wise filtering on passed genes to let only cell types with at least 20 raw counts enter the model fitting step. This ensures the coefficients estimation would not be biased by cell types with extreme low counts.

Transformation of 3′ biased scRNA-Seq Data.

For optimization-free APA identification, we transform scRNA-Seq data that utilize 3′ selection and/or enrichment techniques in library construction (e.g. Drop-Seq, CEL-Seq, and 10x Genomics). Due to the 3′ selection/enrichment, the reads are distributed toward 3′ parts of the transcripts. To make the data compatible for the methods developed for bulk RNA-Seq data where the read coverage is distributed across the whole 3′UTRs, we pad each read along the 3'UTR region until the end based on the 3′UTR definition estimated by DaPars2.

Estimation of PA sites and abundance of long/short isoforms.

For this step, we redesigned multiple modules of DaPars2[38], a widely used method estimating significance of dynamic APA events in the bulk-RNA Seq data. Since it was originally designed to compare between two conditions, such as case and control [14], we extended this module to solve the following optimization problem as follows.

$$(w_{kL}^*, w_{kS}^*, P_k^*) = \underset{w_{kL}^*, w_{kS}^* \geq 0, 1 < P_k < L}{\operatorname{argmin}} || R_{ki} - (w_{kL} I_{kL} + w_{kS} I_{kP}) ||_2^2$$

where w_{kL} and w_{kS} are the transcript abundances of long and short 3′-UTR isoforms for cell cluster k, respectively. $R_{ki} = [R_{ki1}, \dots, R_{kij}, \dots, R_{kiL}]^T$ is the corresponding read coverage at single-nucleotide resolution normalized by total sequencing depth. L is the length of the longest 3′-UTR length from annotation, P_k is the length of alternative proximal 3′-UTR to be estimated, I_{kL} is an indicator function with L times of 1, andI_{kP} has P_k times of 1 and $L - P_k$ times of 0. We solve this equation using quadratic programming [39] as was done in DaPars.

Statistically detecting APA events.

In order to model the relationship between long/short isoform identified above and the given cell types, we build logistic regression for each gene with log-odds of the event that transcript uses distal polyA site (having long isoform) as the outcome and cell types as predictors using weighted effect coding scheme. When scRNA-Seq data were collected from multiple samples or individuals, scMAPA can be easily extended to control the effect of unmatched confounding factors by adding them into the regression model:

$$\ell = \ln\frac{p}{1-p} = \beta_0 + \sum_{i}^{n-1}\beta_i * C_i + \sum_{j}^{m}\beta_j * V_j$$

where $\frac{p}{1-p}$ is the odds of transcript having long isoform. β_i and C_i denote the coefficients and the binary indicator of each cell type, respectively. n is the number of cell types. Since one cell type needs to be chosen as reference for model fitting, scMAPA fits the model twice to get the estimates of coefficients for all cell types. V_j and β_j denote the sample-specific binary confounding variables (e.g. clinical variable) and their coefficients, respectively. m is the number of confounding factors.

When there is no confounding factor, the likelihood ratio test (LRT) between cell type only model and null model is conducted to test the unadjusted effect of cell type, which is equivalent to the likelihood ratio chi-squared test of independence between long/short isoforms and cell types. With the existence of confounding variables, LRT between full model and confounding variables only model is conducted to test the adjusted effect of cell type. P-values from all tests are further adjusted by the Benjamini–Hochberg procedure to control the false-discovery rate (FDR) at 5%. In addition, to ensure there is a significant change in effect size, odds ratio of each cell type against grand mean of all included cell types is calculated. There should be at least one cell type whose odds ratio is greater than 0.25 for a gene to be called as APA gene.

Currently, scMAPA inherits DaPars' focus to identify up to 2 peaks in the 3′-UTRs. However, our logistic model for step 2 can be easily extended to detect >2 peaks if employing other quantifiers that can consider >2 pA sites. For example, when only 2 peaks are detected for a gene, a binary logistic regression model would be fitted. However, when more than 2 peaks are detected for a gene, a multinomial logistic regression model would be fitted. To the best of our knowledge, since the only current tool that detects >2 peaks is scAPA, multinomial logistic regression mode is only compatible with the peak detection result of scAPA. LRT test is used to estimate the significance of APA among multiple peaks and cell types similarly.

Identification of cluster-specific 3′-UTR dynamics.

For the genes where significant APA dynamics is detected, scMAPA further analyses which cell type significantly contributes to the APA in which direction within each gene. By using weighted effect coding scheme, each coefficient in the logistic regression can be interpreted as a measurement of deviation from the grand mean of all cells. This grand mean is not the mean of all cell type means, rather it is the estimate of the proportion of long isoforms of all cells for each

gene. So, the unbalanced cell population sizes, which are common in scRNA-Seq would not affect the accuracy of estimation.

We use the following two criteria to determine the cluster-specific significant 3′-UTR dynamics:

First, given coefficients estimated from logistic regression, we use the Wald test to determine the p-value of each coefficient. P-values among all genes with significant APA of the same cell type are further adjusted by FDR. Then, the absolute coefficient must be greater than $\ln(2)$, corresponding to a 2-fold change in odds ratio. $coefficient \geq \ln(2)$ would be considered as 3′-UTR lengthening and $coefficient \leq -\ln(2)$ would be considered as 3′-UTR shortening. However, users can define a different cutoff value than $\ln(2)$ for $coefficient$ with respect to the stringency they want to set on the identification.

Simulation

First, we used Splatter, a widely known scRNA-Seq simulator, to simulate the cell-level count matrix, which acts as the base of synthetic data. Splatter was trained by unfiltered mouse brain data and set to generate count matrices containing 5000 genes and 3000 cells. The matrix then collapsed to 5 columns, representing the total count of 5 cell groups. We call this 5000 × 5 matrix as cluster-level count matrix.

From the analyses of PBMC and mouse brain data, we found that the standard deviation of PDUI (percentage of distal polyA site usage, which is equivalent to the proportion of long isoforms) of each gene could act as a classifier of APA gene and non-APA gene. Based on that, the standard deviation of PDUI for APA genes in synthetic data is estimated by calculating the mean of standard deviations of PDUI from APA genes detected by both scMAPA and scAPA from mouse brain data. Similarly, the standard deviation of PDUI for non-APA genes was estimated by calculating the mean of standard deviations of PDUI from genes identified as non-APA by both scMAPA and scAPA. With the estimated standard deviations, a PDUI matrix with the same size (5000 × 5) as the cluster-level count matrices was generated. Each row of the PDUI matrix has a standard deviation equal to either estimated standard deviation for the APA gene or non-APA gene. This is achieved by centering 5 randomly selected numbers from standard normal distribution to 0. Then multiply the desired standard deviation to these centered numbers and add them to the desired mean. The mean of each row was randomly picked from 0.05 to 0.95. Since the estimated $SD_{isoprop}$ values are averaged to 0.127 and 0.009 for the APA and the non-APA genes respectively, we generated simulation data with $SD_{isoprop}$ for APA genes in a range centered on 0.13 while fixing that for non-APAs at 0.009. The rows representing true APA genes were randomly selected. Then, each number in the cluster-level count matrix is divided into the count of long isoforms and the count of short isoforms by multiplying and PDUI matrix or (1-PDUI matrix), respectively. Finally, Pearson's chi-squared test (scAPA), logistic regression model + LRT (scMAPA) could be applied to assess the performance of these three methods. For each repeat of simulation, PDUI matrix is regenerated but cluster-level count matrix keeps same for the sake of computational burden. Every simulation design was repeated 100 times to derive summarized statistics.

To examine the impact of experimental design on statistical power to detect significant APA genes, we assess the performance of scMAPA and scAPA in the following aspects: 1) To test the impact of unbalanced cell populations, the proportion of 5 cell types in the synthetic cell-level count matrices were set to three scenarios with different distribution of cell type populations: (20%, 20%, 20%, 20%, 20%), (30%, 17.5%, 17.5%, 17.5%, 17.5%), and (50%, 12.5%, 12.5%, 12.5%, 12.5%). 2) To test the impact of the proportion of true APA genes, we set three levels of true APA proportions, 5%, 10%, and 20%. 3) To test the impact of the extent of APA dynamics, instead of using mean of standard deviations, we set the standard deviations of true APA genes in the simulated PDUI matrix to 15 equally spaced sequence of numbers between the first quartile and the third quartile of standard deviations estimated from APA genes in mouse brain data. In total, there were 9 scenarios, corresponding to 9 combinations of factors 1) and 2). When testing factor 3), we chose balanced cell type proportion (0.2, 0.2, 0.2, 0.2, 0.2) and 10% true APA genes.

Assessing accuracy of PA site estimation

To assess the PA site/ peak interval prediction accuracy, we used peak lists or PA site list from scMAPA, scAPA, and Sierra on PBMC data. The estimation accuracy is measured by the percentage of the predicted peaks or PA sites overlapped with PA sites annotated in PolyASite 2.0. Since it is meaningless to find the overlap between two point estimates, we expanded the point position from annotation database to an interval by manually adding a distance ranging from 10 bp to 150 bp in a 10 bp increment to both sides of the annotated PA sites. scMAPA gives point estimate of PA site as predicted proximal PA site and Sierra gives two point estimates as fit max position and max position. To make the comparison more comprehensive, we calculated the midpoint of peak interval as the pseudo point estimate of scAPA. The point estimates from these methods are considered as supported by annotation database if the point position falls in the annotated interval (annotated PA site ± distance). For peak intervals estimated by scAPA and Sierra, as long as there is 1 bp overlap between the estimated interval and the annotated interval (either start or end of estimated interval falls in annotated PA site ± distance), the estimate would be considered as supported by annotation database. Then, the percentage supported by annotation is calculated as number of PA sites or peak intervals supported by annotation database divided by total peaks detected for each method.

Running scDAPA, scAPA and Sierra

Sierra and scDAPA were run with default parameters. scAPA was ran with default parameters and intronic regions omitted. The genes with CPM less than 10 were filtered out. We want to point out that scAPA employs chisq.test function in R to estimate the significance of dynamic PA sites usage among multiple clusters. This potentially makes the identification of scAPA much conservative than other tools in the multi-cluster setting since it does not allow any cell type to have 0 count, as R's chisq.test would return NA as p-value if there is 0 presented in the count table. However, it is common to observe that a few cell types would not express certain genes in

scRNA-Seq, especially when the whole cell population is split to more than 5 clusters (cell types), which is typical for complex biological systems.

REFERENCES

1. Derti A, Garrett-Engele P, MacIsaac KD, et al. A quantitative atlas of polyadenylation in five mammals. *Genome Res*. 2012;22(6):1173-1183. doi:10.1101/gr.132563.111

2. Masamha CP, Xia Z, Yang J, et al. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature*. 2014;510(7505):412-416. doi:10.1038/nature13261

3. Chen W, Jia Q, Song Y, Fu H, Wei G, Ni T. Alternative Polyadenylation: Methods, Findings, and Impacts. *Genomics Proteomics Bioinformatics*. 2017;15(5):287-300. doi:10.1016/J.GPB.2017.06.001

4. Park HJ, Ji P, Kim S, et al. 3′ UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk. *Nat Genet*. 2018;50:783-789. doi:10.1038/s41588-018-0118-8

5. Fan Z, Kim S, Bai Y, Diergaarde B, Oesterreich S, Park HJ. 3′-UTR shortening contributes to subtype-specific cancer growth by breaking stable ceRNA crosstalk of housekeeping genes. *Front Bioeng Biotechnol*. 2020;to appear:601526. doi:10.1101/601526

6. Kim S, Bai Y, Diergaarde B, Tseng GC, Park HJ. Alternative Polyadenylation Modifies Target Sites of MicroRNAs with Clinical Potential for Breast Cancer Progression. *bioRxiv*. Published online January 1, 2019:601518. doi:10.1101/601518

7. Zhang H, Lee JY, Tian B. Biased alternative polyadenylation in human tissues. *Genome Biol*. 2005;6(12):R100. doi:10.1186/gb-2005-6-12-r100

8. Ye C, Zhou Q, Wu X, et al. scDAPA: detection and visualization of dynamic alternative polyadenylation from single cell RNA-seq data. *Bioinformatics*. 2020;36(4):1262-1264. doi:10.1093/bioinformatics/btz701

9. Patrick R, Humphreys DT, Janbandhu V, et al. Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome Biol*. 2020;21(1):167. doi:10.1186/s13059-020-02071-7

10. Shulman ED, Elkon R. Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic Acids Res*. 2019;47(19):10027-10039. doi:10.1093/nar/gkz781

11. Nott A, Holtman IR, Coufal NG, et al. Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science (80-)*. 2019;366(6469):1134 LP - 1139. doi:10.1126/science.aay0793

12. Doorn KJ, Brevé JJP, Drukarch B, et al. Brain region-specific gene expression profiles in freshly isolated rat microglia. *Front Cell Neurosci*. 2015;9:84.

doi:10.3389/fncel.2015.00084

13. McKenzie AT, Wang M, Hauberg ME, et al. Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Sci Rep*. 2018;8(1):8868. doi:10.1038/s41598-018-27293-5

14. Xia Z, Donehower LA, Cooper TA, et al. Dynamic Analyses of Alternative Polyadenylation from RNA- Seq Reveal Landscape of 3 ' UTR Usage Across 7 Tumor Types. *Nat Commun*. Published online 2014:1-38.

15. Chen M, Ji G, Fu H, et al. A survey on identification and quantification of alternative polyadenylation sites from RNA-seq data. *Brief Bioinform*. 2019;21(4):1261–1276. doi:10.1093/bib/bbz068

16. Zeisel A, Hochgerner H, Lönnerberg P, et al. Molecular Architecture of the Mouse Nervous System. *Cell*. 2018;174(4):999-1014.e22. doi:10.1016/j.cell.2018.06.021

17. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol*. 2017;18(1):174. doi:10.1186/s13059-017-1305-0

18. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031

19. Zhang X, Lan Y, Xu J, et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res*. 2019;47(D1):D721-D728. doi:10.1093/nar/gky900

20. Herrmann CJ, Schmidt R, Kanitz A, Artimo P, Gruber AJ, Zavolan M. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Res*. 2020;48(D1):D174-D179. doi:10.1093/nar/gkz918

21. Ye C, Zhou Q, Hong Y, Li QQ. Role of alternative polyadenylation dynamics in acute myeloid leukaemia at single-cell resolution. *RNA Biol*. 2019;16(6):785-797. doi:10.1080/15476286.2019.1586139

22. Sommerkamp P, Altamura S, Ladel L, et al. ALTERNATIVE POLYADENYLATION REGULATES HEMATOPOIETIC STEM CELL METABOLISM. *Exp Hematol*. 2019;76:S86. doi:https://doi.org/10.1016/j.exphem.2019.06.436

23. Xiang Y, Ye Y, Lou Y, et al. Comprehensive Characterization of Alternative Polyadenylation in Human Cancer. 2018;110(November 2017):1-11. doi:10.1093/jnci/djx223

24. Hilgers V, Lemke SB, Levine M. ELAV mediates 3' UTR extension in the Drosophila nervous system. *Genes Dev*. 2012;26(20):2259-2264. doi:10.1101/gad.199653.112

25. Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*. 2005;123(6):1133-1146. doi:10.1016/j.cell.2005.11.023

26. Shepard PJ, Choi E-A, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*. 2011;17(4):761-772. doi:10.1261/rna.2581711

27. Hilgers V, Perry MW, Hendrix D, Stark A, Levine M, Haley B. Neural-specific elongation of 3' UTRs during Drosophila development. *Proc Natl Acad Sci U S A*. 2011;108(38):15864-15869. doi:10.1073/pnas.1112672108

28. Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456(7221):470-476. doi:10.1038/nature07509

29. Tsai PS, Kaufhold JP, Blinder P, et al. Correlations of Neuronal and Microvascular Densities in Murine Cortex Revealed by Direct Counting and Colocalization of Nuclei and Vessels. *J Neurosci*. 2009;29(46):14553 LP - 14570. doi:10.1523/JNEUROSCI.3287-09.2009

30. Feiglin A, Allen BK, Kohane IS, Kong SW. Comprehensive Analysis of Tissue-wide Gene Expression and Phenotype Data Reveals Tissues Affected in Rare Genetic Disorders. *Cell Syst*. 2017;5(2):140-148.e2. doi:10.1016/j.cels.2017.06.016

31. Kuleshov M V., Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44(W1):W90-W97. doi:10.1093/nar/gkw377

32. Silverman HA, Dancho M, Regnier-Golanov A, et al. Brain region-specific alterations in the gene expression of cytokines, immune cell markers and cholinergic system components during peripheral endotoxin-induced inflammation. *Mol Med*. 2014;20(16):601-611. doi:10.2119/molmed.2014.00147

33. Li P, Xu G, Li G, Wu M. Function and mechanism of tumor suppressor gene LRRC4/NGL-2. *Mol Cancer*. 2014;13:266. doi:10.1186/1476-4598-13-266

34. Kothari PH, Kolar GR, Jen JC, et al. TREX1 is expressed by microglia in normal human brain and increases in regions affected by ischemia. *Brain Pathol*. 2018;28(6):806-821. doi:10.1111/bpa.12626

35. Ye C, Long Y, Ji G, Li QQ, Wu X. APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. doi:10.1093/bioinformatics/bty029

36. Arefeen A, Liu J, Xiao X, Jiang T. TAPAS : tool for alternative polyadenylation site analysis. *Bioinformatics*. 2018;34(February):2521-2529. doi:10.1093/bioinformatics/bty110

37. Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc*. 2014;9(1):171-181. doi:10.1038/nprot.2014.006

38. Li L, Gao Y, Peng F, Wagner EJ, Li W. Genetic Basis of Alternative Polyadenylation is an Emerging Molecular Phenotype for Human Traits and Diseases. *BioRxiv*. Published online 2019. doi:10.1101/570176

39. Bohnert R, Rätsch G. rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res*. 2010;38(Web Server issue):W348-51. doi:10.1093/nar/gkq448

Acknowledgements We thank Daniel Weeks, Ph.D., Professor, Department of Human Genetics, University of Pittsburgh for valuable discussion. This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. We also acknowledge the authors of scAPA for their generous provision of their data.

Author Contributions H.J.P and Y.B. conceived the project, designed the experiments. Y.B. and Z.F. implemented the software. Y.B., Y.Q., R.M. performed the analysis. S.K., K.N., H.M.Z., R.K., Q.P. interpreted the results statistically and/or biologically.

Competing interests The authors declares no competing financial interests.

Availability of data and materials The open source scMAPA program (version 0.9.1) is freely available at https://github.com/ybai3/scMAPA with necessary example data for this analysis.

Funding This work was supported partly by the Joan Gollin Gaines Cancer Research Fund at the University of Pittsburgh to H.J.P.. This project used the UPMC Hillman Cancer Center Biostatistics Shared Resource that is supported in part by award P30CA047904.

Figure 1. Schematic illustration of scMAPA. Bars represent the estimated abundance of 3′-UTR shortening (left) and lengthening (right) isoforms in each cluster-bulk data. The black bars on the bottom represent the grand mean of all long/short isoforms across the clusters.

Figure 2. Performance assessment using simulated data. (A). Illustration of the simulation process. With fixed number of true APA events (500 out of 5000) and uniform distribution of cell cluster size (600 cells in each cell type), (B) sensitivity and (C) specificity were plotted against varying degree of standard deviation (SD) of PDUI values across clusters ($SD_{isoprop}$) for true APA genes. With fixed number of true APA events (500) and SD values (0.127 for true APA genes and 0.009 for non-APA genes), (D) sensitivity and (E) specificity in scenarios with different distributions of cell cluster size: (20%, 20%, 20%, 20%, 20%) for scenario a, (30%, 17.5%, 17.5%, 17.5%, 17.5%) for b, and (50%, 12.5%, 12.5%, 12.5%, 12.5%) for c.

Figure 3. Performance assessment of scMAPA and scAPA using PBMC data. (A) Percentage of pA sites each method identified in the 10k data that are in proximity to known pA sites annotated in PolyASite 2.0 by the distance defining the proximity. (B) Upset plot showing diverse overlaps among APA genes in the 10k data identified by four methods, scAPA, Sierra, scMAPA and scDAPA. Barplot on top shows the number of genes corresponding to the set combination indicated below. Black bars correspond to the sets involving scMAPA results. Colored horizontal bars represent (C). Venn diagram of significant APA genes detected by scMAPA across 10k, 5k, and 1k data. (D) Frequency polygon plot shows the distribution of standard deviations (SD) of PDUI values across clusters ($SD_{isoprop}$) for significant APA genes. (E) Bar plot shows that the significant APA genes identified by scMAPA is significantly enriched with “hemato”-related “Disease & Function” IPA terms. The blue bar represents the -log10(B-H p values) from the enrichment tests.

Figure 4. Gene-level and gene-cluster-level identification using mouse brain data. (A) Venn diagram of significant APA genes detected by scMAPA and scAPA. (B) Enrichment analysis of the APA genes uniquely detected by scMAPA (green) and scAPA (red), and those commonly detected by both (yellow) (C) Heatmap of coefficients of cell type-specific APA genes. Coefficients were estimated in logistic regression model. (D) Bar plot shows the number of 3′-UTR lengthening and shortening detected in each cell type. (E) Bar plot shows the enrichment (-log10(B-H p-value)) of brain cell-type-specific APA genes (blue for astrocyte, orange for immune, green for oligos, red for vascular, and violet for neurons).

Figure 5. (A) Venn diagrams show the significant APA identification by scMAPA before and after adjusting for the brain region. Pink and violet diagrams show the APA genes identified with and without the adjustment, respectively. (B) Significance of overlap between the 113 brain-region-specific APA genes and the up-regulated genes in GTEx samples whether they are from brain (red) or not (green). A higher overlap significance indicates a more significant overlap, calculated by Enrichr. (C) Significance of overlap between the 113 genes and the up-regulated genes in GTEx brain samples whether they are from cortex (red) or not (green).

Supplemental Figure 1. Signal density of 10k PBMC scRNA-Seq reads mapped onto 3′-UTR of GATA2 (A) and FLT3 (B) in terms of original 3′ tag-based (top panel) or of padded reads (C)

and (D) respectively, for selected clusters for presentation purpose. While the genomic coordinates are shared between A and C, B and D, the blue arrows indicate the polyA site annotated in polyASite database (v. 2.0). Red arrow indicates the proximal polyA site predicted by DaPars. (E) Algorithm overview of bioinformatic tools and statistical methods to identify dynamic APAs in scRNA-Seq data

Supplemental Figure 2. Performance assessment of significance estimation methods using simulated data. (A) shows the frequency of standard deviations (SD) of PDUI values across clusters from mouse brain data. Genes identified as significant APA genes by both scMAPA and scAPA were considered as APA genes. Genes identified as non-significant APA genes by both methods were considered as non-APA genes. (B) to (E) show the performance assessment using simulated data. With fixed number of true APA events (250) and SD values (0.1268 for true APA genes and 0.009190 for non-APA genes), box plots in (B) and (C) show the sensitivity and specificity in scenarios with different distributions of cell type populations: (20%, 20%, 20%, 20%, 20%) for scenario a, (30%, 17.5%, 17.5%, 17.5%, 17.5%) for b, and (50%, 12.5%, 12.5%, 12.5%, 12.5%) for c. Box plots in (D) and (E) show the sensitivity and specificity with the number of true APA events set to 1,000 and all other factors remain same.

Supplemental Figure 3. (A) Percentage of pA sites each method identified in the 5k data that are in proximity to known pA sites annotated in PolyASite 2.0 by the distance defining the proximity. (B) Upset plot showing diverse overlaps among APA genes in the 5k data identified by four methods, scAPA, Sierra, scMAPA and scDAPA. Barplot on top shows the number of genes corresponding to the set combination indicated below. Black bars correspond to the sets involving scMAPA results. Colored horizontal bars represent the total number of APA genes identified by each method. (C). Venn diagram of significant APA genes across 10k, 5k, and 1k data detected by scAPA, scDAPA, and Sierra, respectively. (D) Boxplot representing the number of APA genes identified by scMAPA by the number of clusters sampled from the 13 clusters of the 10k data.

Supplemental Figure 4. (A) Heatmaps of log(CPM+1) of all cell type-specific APA genes shown in Fig 4.D. (B)-(F) Scatter plots show the correlation pattern between APA dynamic and expression of genes in Fig 4.C by cell type. X-axis represents coefficients shown in Fig 4.C, Y-axis represents log(CPM+1) shown in Fig S3.A. (B) shows the pattern for Astrocytes, (C) for Immune, (D) for Neurons, (E) for Oligos, (F) for Vascular cells.

Supplemental Figure 5. tSNE plot showing the cell type (A) and brain region (B) of the mouse brain scRNA-Seq data. (C) Significance of overlap between the 2,575 APA genes that are not specific to brain regions and the up-regulated genes in GTEx samples whether they are from brain (red) or not (green). A higher overlap significance indicates a more significant overlap, calculated by Enrichr. (D) Significance of overlap between the 2,575 genes and the up-regulated genes in GTEx brain samples whether they are from cortex (red) or not (green).

S. Table 1. Cell type annotation based on marker genes curated in CellMarker[19] for 10k, 5k, and 1k in the PBMC data.

S. Table 2. Detailed information of APA genes detected by scMAPA, scAPA, scDAPA, and Sierra on the PBMC data including Ingenuity Pathway Analysis (IPA) analysis result.

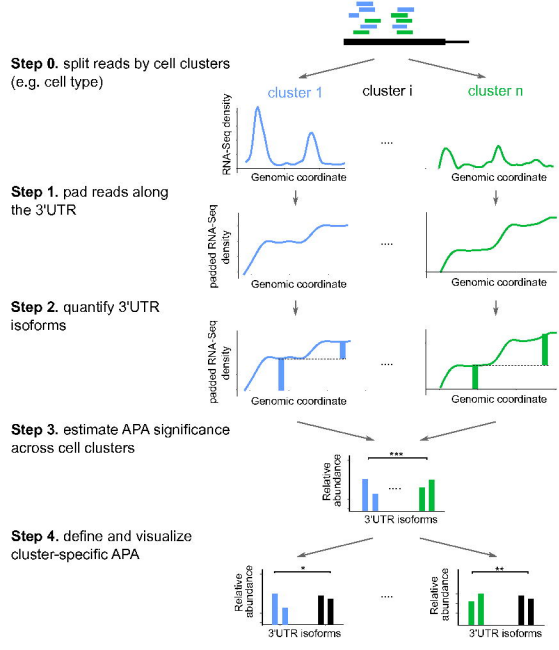
S. Table 3. scMAPA estimation result for cell-type-specific APA genes on the mouse brain data.

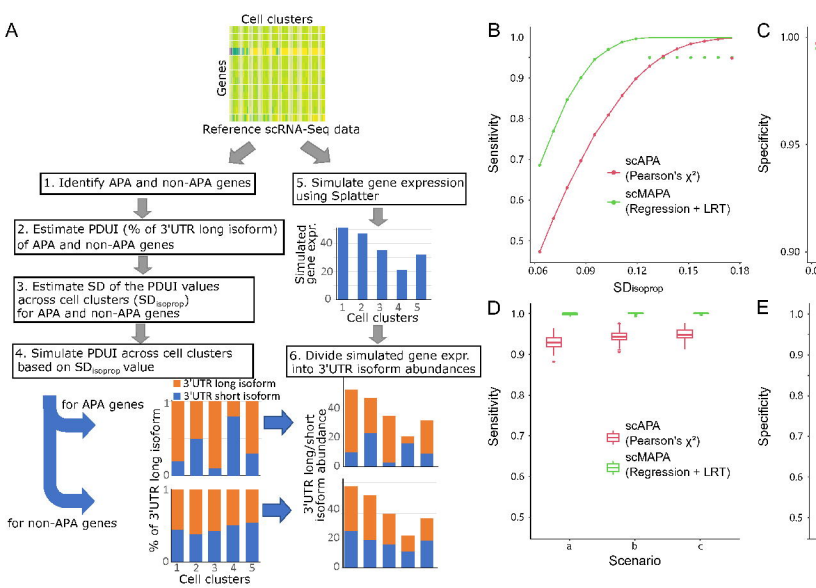
S. Table 4. Result of IPA comparison analysis on the "Disease & Function" terms enriched for APA genes identified uniquely by scAPA, scMAPA and commonly by both on the mouse brain data (1,446, 2,175, and 1,048 respectively).

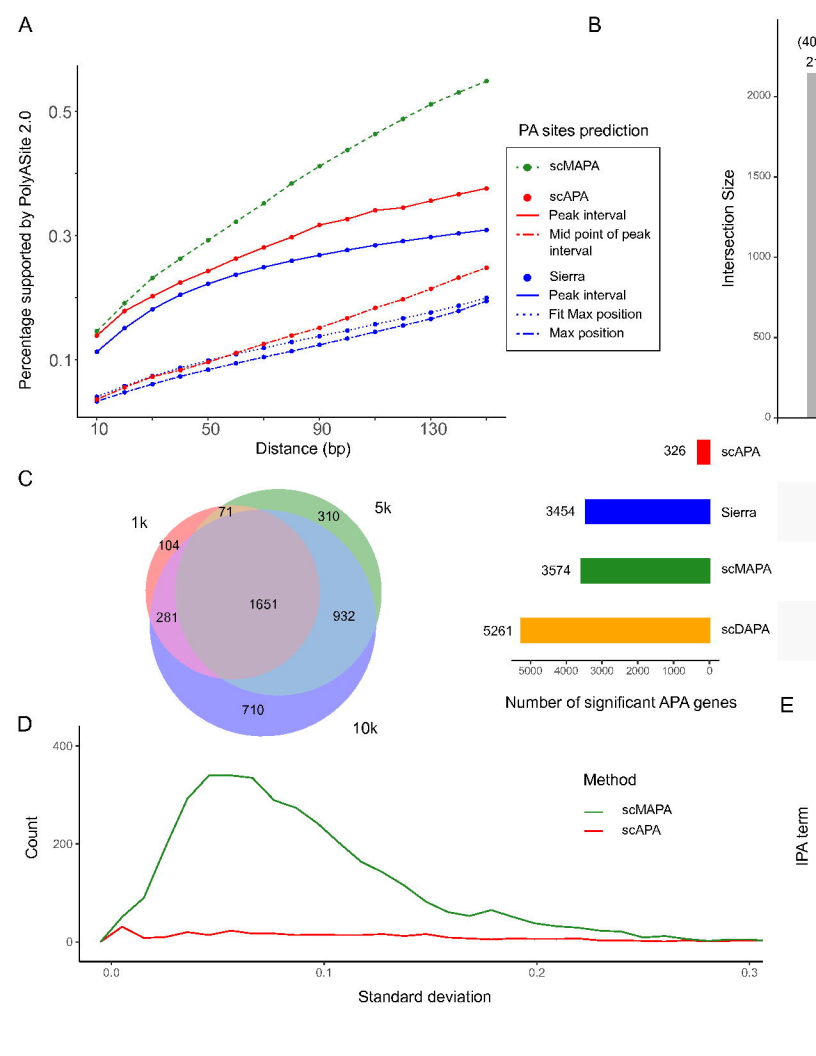
S. Table 5. Result of IPA comparison analysis on the "Disease & Function" terms enriched for APA genes identified uniquely in astrocyte, immune, oligos, vascular, and neuron cells.

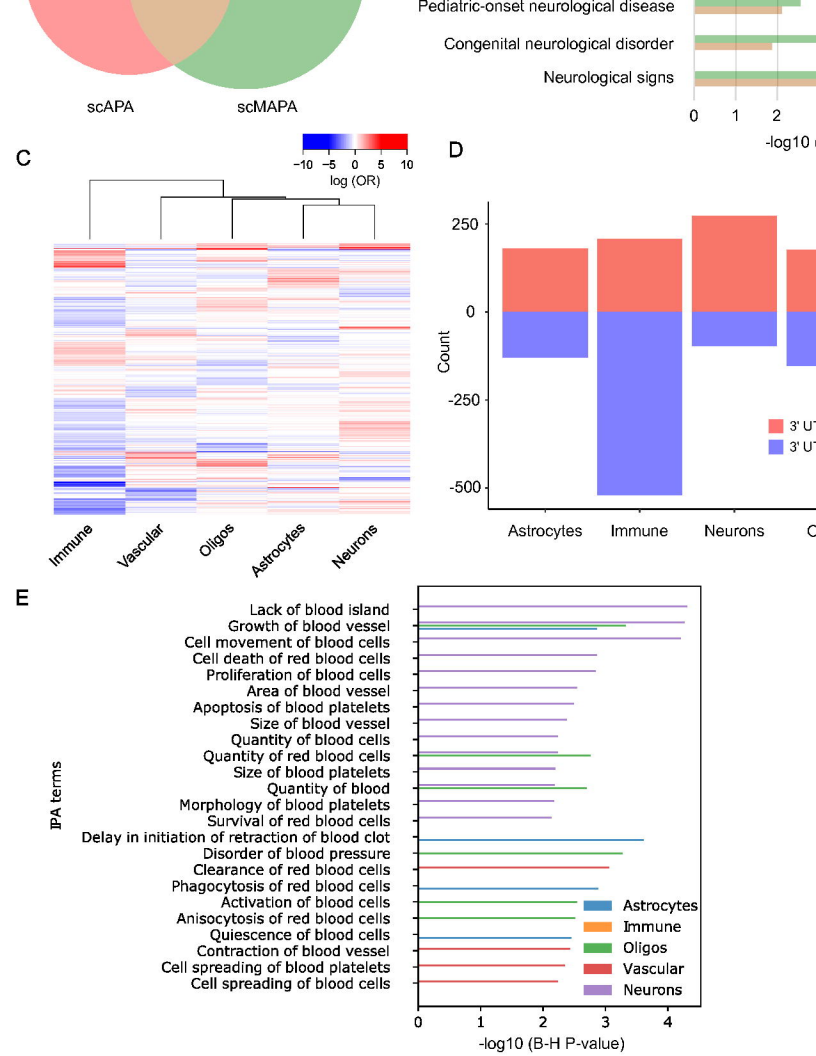
S. Table 6. scMAPA estimates on the input data that are split by cell type and brain region either with brain region as a confounder or not.

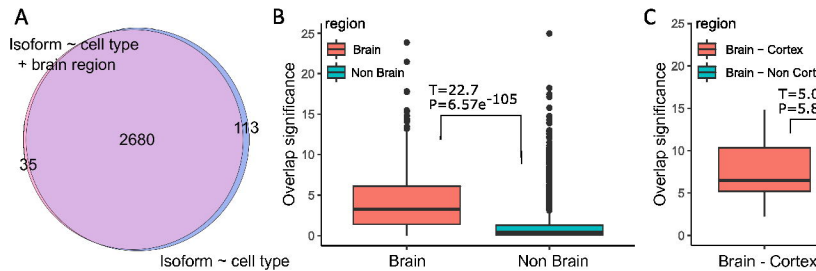
S. Table 7. IPA upstream regulator analysis result (enrichment p-value) on 113 and 2,715 APA genes that are supposed to be brain-region-specific and non-specific, respectively.

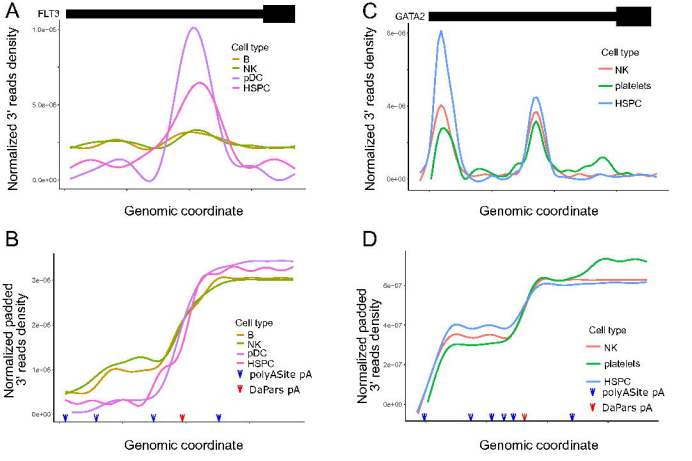

Step 0. split reads by cell clusters (e.g. cell type)
cluster 1
cluster i
cluster n
RNA-Seq density
Genomic coordinate
Step 1. pad reads along the 3'UTR
padded RNA-Seq density
Step 2. quantify 3'UTR isoforms
Step 3. estimate APA significance across cell clusters
***
Relative abundance
3'UTR isoforms
Step 4. define and visualize cluster-specific APA
*
**


A
Cell clusters
Genes
Reference scRNA-Seq data
1. Identify APA and non-APA genes
2. Estimate PDUI (% of 3'UTR long isoform) of APA and non-APA genes
3. Estimate SD of the PDUI values across cell clusters (SDisoprop) for APA and non-APA genes
4. Simulate PDUI across cell clusters based on SDisoprop value
5. Simulate gene expression using Splatter
Simulated gene expr.
Cell clusters
6. Divide simulated gene expr. into 3'UTR isoform abundances
for APA genes
for non-APA genes
3'UTR long isoform
3'UTR short isoform
% of 3'UTR long isoform
3'UTR long/short isoform abundance
Cell clusters
B
Sensitivity
SDisoprop
scAPA (Pearson's χ²)
scMAPA (Regression + LRT)
C
Specificity
D
Sensitivity
Scenario
a
b
c
scAPA (Pearson's χ²)
scMAPA (Regression + LRT)
E
Specificity


A
Percentage supported by PolyASite 2.0
0.5
0.3
0.1
10
50
90
130
Distance (bp)
PA sites prediction
scMAPA
scAPA
Peak interval
Mid point of peak interval
Sierra
Peak interval
Fit Max position
Max position
B
Intersection Size
2000
1500
1000
500
0
326
scAPA
C
1k
5k
10k
71
310
104
1651
281
932
710
3454
Sierra
3574
scMAPA
5261
scDAPA
5000
4000
3000
2000
1000
0
Number of significant APA genes
D
Count
400
200
0
Method
scMAPA
scAPA
0.0
0.1
0.2
0.3
Standard deviation
E
IPA term

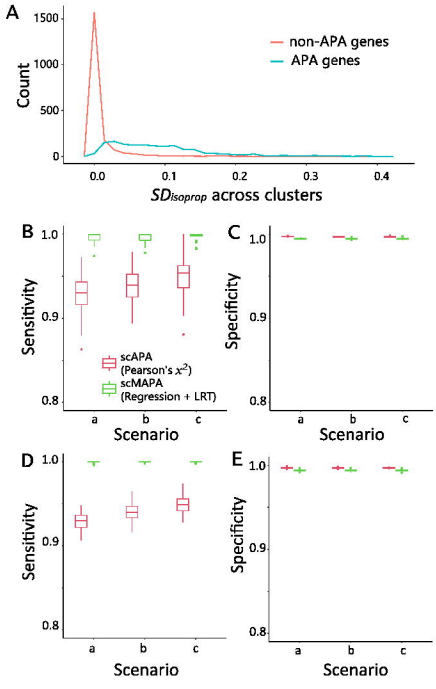

scAPA
scMAPA
Pediatric-onset neurological disease
Congenital neurological disorder
Neurological signs
0
1
2
-log10
C
−10 −5 0 5 10
log (OR)
Immune
Vascular
Oligos
Astrocytes
Neurons
D
250
0
-250
-500
Count
Astrocytes
Immune
Neurons
E
IPA terms
Lack of blood island
Growth of blood vessel
Cell movement of blood cells
Cell death of red blood cells
Proliferation of blood cells
Area of blood vessel
Apoptosis of blood platelets
Size of blood vessel
Quantity of blood cells
Quantity of red blood cells
Size of blood platelets
Quantity of blood
Morphology of blood platelets
Survival of red blood cells
Delay in initiation of retraction of blood clot
Disorder of blood pressure
Clearance of red blood cells
Phagocytosis of red blood cells
Activation of blood cells
Anisocytosis of red blood cells
Quiescence of blood cells
Contraction of blood vessel
Cell spreading of blood platelets
Cell spreading of blood cells
Astrocytes
Immune
Oligos
Vascular
Neurons
0
1
2
3
4
-log10 (B-H P-value)

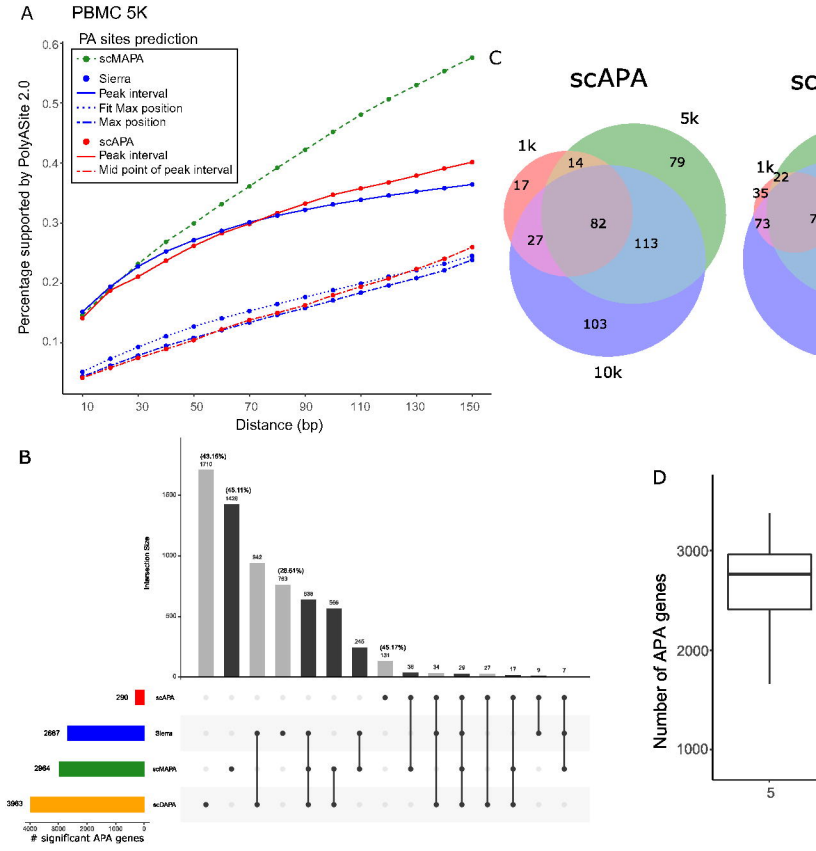

A
Isoform ~ cell type
+ brain region
2680
113
35
Isoform ~ cell type
B
region
Brain
Non Brain
T=22.7
P=6.57e-105
Overlap significance
Brain
Non Brain
C
region
Brain - Cortex
Brain - Non Cort
Overlap significance
Brain - Cortex

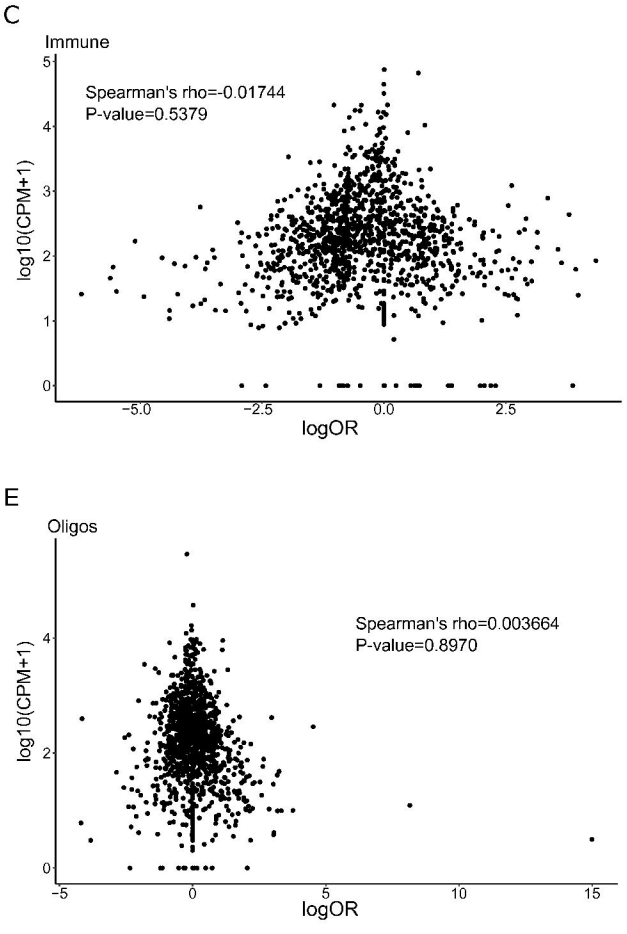


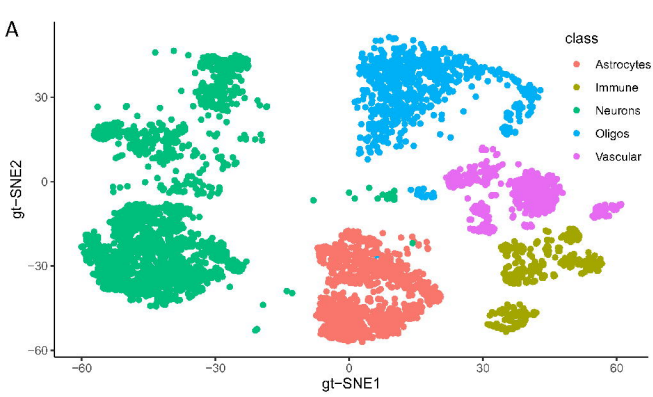
E

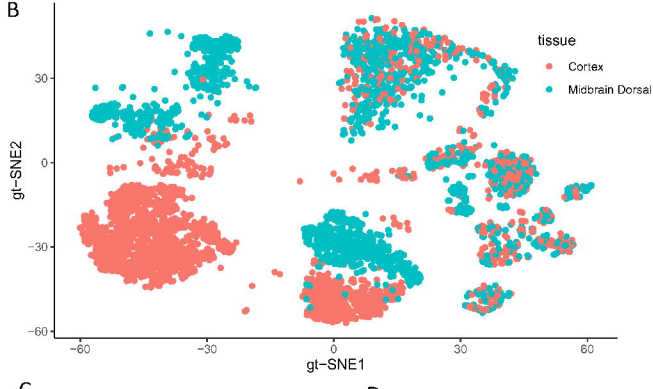
	scMAPA	scAPA	scDAPA	Sierra
Step 1. split reads by clusters	In Python script	Dropseq-tools, called in R	In shell script	User provided
Step 2. quantify APA in each cluster	DaPars + long/short isoform estimation	Homer findPeaks (peak filtering) + mclust (splitting peaks) + featureCounts (isoform estimation)	Based on difference in read length and in peak distribution	Junction awared peak calling Peaks merging for multiple datasets UMI counting
Step 3. estimate significance of APA across clusters	Binary or multinomial logistic regression + LRT	Pearson's χ^2 test	Wilcoxon sum-rank test on pairwise comparisons	DEXSeq pipline on pseudo bulk counts
Step 4. define cluster-specific APA	Wald test + filter on estimated coefficients			
Note	Identifying up to 2 peaks	Identifying > 2 peaks	Identifying up to 2 cell clusters	Identifying up to 2 cell clusters

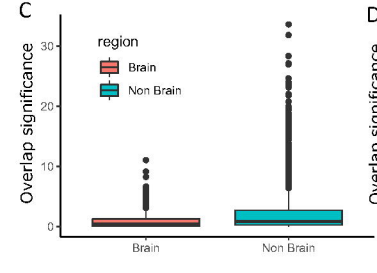

A
1500
1000
500
0
Count
non-APA genes
APA genes
0.0
0.1
0.2
0.3
0.4
SDisoprop across clusters
B
1.0
0.9
0.8
Sensitivity
scAPA
(Pearson's χ2)
scMAPA
(Regression + LRT)
a
b
c
Scenario
C
1.0
0.9
0.8
Specificity
a
b
c
Scenario
D
1.0
0.9
0.8
Sensitivity
a
b
c
Scenario
E
1.0
0.9
0.8
Specificity
a
b
c
Scenario

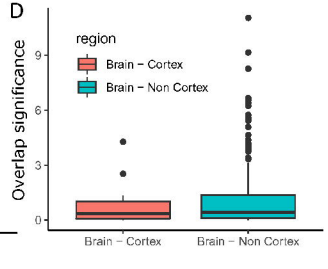

A
PBMC 5K
PA sites prediction
scMAPA
Sierra
Peak interval
Fit Max position
Max position
scAPA
Peak interval
Mid point of peak interval
Percentage supported by PolyASite 2.0
Distance (bp)
C
scAPA
5k
1k
10k
14
79
17
82
27
113
103
sc
1k
22
35
73
B
Intersection Size
(49.16%)
5710
(45.11%)
1438
842
(28.61%)
769
599
556
240
(45.17%)
131
36
34
29
27
17
9
7
scAPA
Sierra
scMAPA
scDAPA
290
2687
2964
3963
significant APA genes
D
Number of APA genes
3000
2000
1000
5


C
Immune
Spearman's rho=-0.01744
P-value=0.5379
log10(CPM+1)
logOR
D
Neurons
log10(CPM+1)
E
Oligos
Spearman's rho=0.003664
P-value=0.8970
log10(CPM+1)
logOR
F
Vascular
log10(CPM+1)


A
class
Astrocytes
Immune
Neurons
Oligos
Vascular
gt−SNE2
gt−SNE1
30
0
−30
−60
−60
−30
0
30
60


B
tissue
Cortex
Midbrain Dorsal
gt−SNE2
gt−SNE1
30
0
−30
−60
−60
−30
0
30
60


C
region
Brain
Non Brain
Overlap significance
30
20
10
0
Brain
Non Brain


D
region
Brain − Cortex
Brain − Non Cortex
Overlap significance
9
6
3
0
Brain − Cortex
Brain − Non Cortex