Identifying genes and cell types with dynamic alternative polyadenylation in multi-cluster single-cell transcriptomics data

YuLong Bai[1], Yidi Qin[1], Soyeon Kim[2,3], Zhenjiang Fan[4], KyongNyon Nam[5], Radosveta Koldamova[5], Quasar Saleem Padiath[1,6], Hyun Jung Park[1†]

[1] Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States

[2]Department of Pediatrics, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, USA

[3]Division of Pulmonary Medicine, Children’s Hospital of Pittsburgh of UPMC, Pittsburgh, Pennsylvania, USA

[4]Department of Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

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

[6]Department of Neurobiology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

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

ABSTRACT

Alternative polyadenylation (APA) causes shortening or lengthening of the 3′-untranslated region (3′-UTR), widespread in complex tissues. Bioinformatic tools have been developed to identify dynamic APA in single cell RNA-Seq (scRNA-Seq) data, but with relatively low power and the lack of interpretability for multiple cell types. To address these limitations, we developed a model-based method, scMAPA. scMAPA increases power by building a regression model based on a sensitive quantification of 3′-UTR short and long isoforms. By developing a *de novo* simulation platform, we demonstrated that scMAPA shows a markedly better sensitivity (9.5% on average) than the previous method with a negligible loss in specificity (0.3% on average). scMAPA improves interpretability by modeling the direction and the degree of dynamic APA for each cell type for each gene, while allowing flexibility to control potential confounders. scMAPA enables to bring a systematic understanding of APA dynamics in complex tissues in human Peripheral Blood Monocellular Cells and mouse brain data.

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 shortened or lengthened 3′-UTRs. These APA events are widespread in diverse physiological and pathological processes such as cancer, viral infection, cardiac hypertrophy, heart failure, muscular dystrophy, and sclerosis [3]. For example, 3′-UTR lengthening was reported to regulate cell senescence [4] that is linked with important biological functions such as cell cycle inhibitors and DNA damage markers [5–7]. 3′-UTR shortening was reported widespread in cancers [8]. We recently identified a trans effect of 3′-UTR shortening in TCGA breast cancer data[9–11], many of which were further validated in wet-bench experiments. However, understanding of their functions is generally very limited in most of the processes. To understand the functions of 3′-UTR shortening and lengthening in a finer scale, it is critical to consider tissue-specificity of APA events [1,12] suggesting its role in tissue-specific gene regulation [13]. To consider the tissue-specificity, it is reasonable to identify APA events in single-cell RNA sequencing (scRNA-Seq) data, since scRNA-Seq data allow to investigate transcriptomic dynamics in the single-cell resolution. In contrast to scRNA-Seq data, the original RNA-Seq data from a mixture of cells is referred to as bulk RNA-Seq data.

Several bioinformatic tools have been developed to identify APA events based on RNA-Seq data. Before scRNA-Seq data became widely available, one of the early developments for bulk RNA-Seq data includes Dynamic analyses of Alternative PolyAdenylation from RNA-Seq (DaPars) [8]. With its case-control design, it has been used to make significant discoveries in disease vs. healthy samples[8,14]. Those methods with the case-control setting, such as DaPars, can be extended for scRNA-Seq data consisting of two cell types or clusters. However, those methods are not suitable for scRNA-Seq data typically of multiple cell clusters. Recently, several tools have been developed for scRNA-Seq data, namely scDAPA[15] and scAPA[16]. However, they also come with several limitations to handle the multi-cluster setting of scRNA-Seq data. First, although scDAPA takes scRNA-Seq data as input, it identifies dynamic APA in comparison of two cell clusters in the data, thus not directly applicable for more than two clusters. On the other hand, while scAPA can identify genes with dynamic APA (gene-level identification) in more than two clusters using a statistical test (Pearson's χ^2), it raises two challenges for further analyses. First, after it identifies each gene with dynamic APA, it does not estimate in which clusters the gene undergoes dynamic APA in which direction (3′-UTR shortening or lengthening) and how much degree. Second, the test is based on a contingency table that explicitly divides the input samples. Thus, scAPA cannot directly control confounding factors when the samples differ in multiple aspects. This becomes problematic in studying complex tissues. For example, brain transcriptomic dynamics is known to be specific to regions (e.g. cortex and dorsal midbrain) and cell types (e.g. neuron and astrocyte) [17–19]. Thus, depending on how brain cells are clustered and the research question, it may be necessary to identify dynamic APA while controlling either brain region or cell type information. To address these challenges and bring a more systematic understanding of dynamic APA in scRNA-Seq data, we developed scMAPA that explicitly models 3′-UTR long and short isoforms (those using distal and proximal pA sites, respectively). This modeling of scMAPA improves interpretability by introducing a novel layer of analysis identifying cell clusters with different patterns of dynamic APA. Combined with the flexibility to control confounding factors, scMAPA brings a more comprehensive understanding of dynamic APA not only in terms of APA genes, but also clusters with different dynamics of APA for the genes.

With several bioinformatics tools developed to identify APA events in scRNA-Seq data, it is necessary to evaluate their performances in simulations where true APA and true non-APA genes are tested. However, it has been challenging due to the nature of the tasks. APA detection tools based on RNA-Seq alignment density shape perform two steps: 1) quantifying 3′-UTR long and short isoforms from the density shape and 2) estimating statistical significance of the bias in the long/short isoform proportion. It is difficult to model and simulate the density shape for true APA genes, since RNA alignment density shape is most dynamic on the 3′-UTRs. We believe that this is partly why current tools use heuristics for the first step and validate the findings mainly through the annotated polyadenylated sites[8,15,16]. On the other hand, it is relatively achievable to simulate and test the second step of estimating the statistical significance once the long and short isoforms are identified. Also, it is worth evaluating the APA detection tools at this step, since the second step largely determines the statistical power of the tools. Motivated by this rationale, we develop a simulation platform where the APA detection tools are compared in terms of their statistical power. In simulation data of diverse simulation scenarios and biological data of different contexts, scMAPA outperforms a previous method both in statistical power and interpretability.

MATERIALS AND METHODS

Data sets

PBMC data. Aligned BAM file and K-means clustering results were downloaded from the 10X genomics repository. Only 5 clusters with most cell numbers were used in the analysis.

Mouse brain data. Aligned BAM file and clustering result of cortex and midbrain dorsal from two donors were downloaded from [20]. We included only neurons, immune cells, astrocytes, oligos, and vascular cells in the analysis. On the data, scAPA previously identified widespread dynamic APA events (2,506 (14.8%) with a significant change in pA usage out of 16,942 3′-UTR peaks). We ran scAPA with the parameters specified in paper [16], and obtained 1,084 transcripts with significant (B-H P-val < 0.05) dynamic APA. We believe the number difference should be due to the difference in the unit of experiments (either change in pA usage or PDUI as estimated by DaPars) and of APA events (either 3′-UTR peak or transcript), the version of genomes, preprocessing steps, or other parameters that were not described.

scMAPA algorithm

Estimation of abundance of long/short isoforms.

For this step, we employed a module of DaPars, a widely used method estimating significance of dynamic APA events in the bulk-RNA Seq data between two conditions, such as case and control [8]. Before estimating significance of dynamic APA between two conditions, it first identifies the 3′-UTR long/short isoforms for each gene in each condition. We extended this module to estimate abundance of the isoforms for each cell cluster of scRNA-Seq data.

Model fitting and APA detection.

To quantifying 3′-UTR long and short isoforms from the density shape for our computational framework, we successfully redesigned an APA detection tool originally developed for bulk RNA-Seq, DaPars to solve the following optimization problem:

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

where w_{kL} and w_{kS} are the transcript abundances of long and short 3′-UTR isoforms for cell cluster k, respectively. $R_{ki} = \left[R_{ki1}, \ldots, R_{kij}, \ldots, R_{kiL}\right]^T$ 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, and I_{kP} has P_k times of 1 and $L - P_k$ times of 0. Using the model of DaPars, we take the optimization of this linear regression model using quadratic programming [21]. In order to model the relationship between long/short isoform and cell type, 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 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 possible confounding variables and their coefficients, respectively. m is the number of confounding factors.

When there is no confounding variable, 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%.

Model fitting and APA detection of scMAPA is compatible to >2 peak detection result. When only 2 peaks are detected for a gene, a binary logistic regression model would be fitted. 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 are 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.

$$3'UTR\ lengthening \begin{cases} FDR \leq 0.05 \\ coefficient \geq \ln(2) \end{cases}$$

$$3'UTR\ shortening \begin{cases} FDR \leq 0.05 \\ coefficient \leq -\ln(2) \end{cases}$$

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 0.19 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 (model-based scMAPA), Fisher's exact test (test-based 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.

Application on PBMC and mouse brain data

Model-based scMAPA. In the application on PBMC and mouse brain data with only cell type as independent variable, only genes with sum CPM of all cell types greater than 10 and expression detected in at least three clusters were kept. In the application on mouse brain data with cell type and tissue type as independent variables, we kept genes detected in at least three cell types where in each cell type, it must be detected in both tissue types.

Test-based scMAPA. Instead of using logistic regression to build a model. We employed Fisher’s exact test on each pairwise comparisons (e.g. 10 comparisons for 5 clusters). P-values from all comparisons were adjusted by Benjamini–Hochberg procedure. Significant APA genes were defined by having at least one adjusted P-value less than 0.05 among all pairwise comparisons.

scAPA. scAPA was ran with default parameters and intronic regions omitted. The genes with CPM less than 10 were filtered out.

RESULTS

Alternative Polyadenylation identification in multi-cluster setting of single-cell RNA-Seq data (scMAPA)

To increase power and enhance interpretability in detecting dynamic APA in scRNA-Seq data, we developed scMAPA. To increase power in the two steps explained above, scMAPA employs a sensitive quantifier of 3′-UTR isoforms and build a regression model respectively. In step 1, given the input scRNA-Seq data and the cell cluster definition, scMAPA first divides the aligned read data by the cell cluster, which we refer to as cluster-bulk data. Then, in each cluster-bulk data, scMAPA estimates the abundance of 3′-UTR long and short isoform of genes using linear regression and quadratic programming[21] implemented in DaPars[8]. DaPars demonstrated its sensitivity in quantifying 3′-UTR long and short isoform in multiple analyses[8,9,22]. In step 2, scMAPA enhances its power by explicitly modeling the relationship among the ratio of the long/short isoforms, the cell cluster identity, and other possible confounding factors in logistic regression (see Methods). In comparison to previous methods only identifying genes with dynamic APA (gene-level identification), scMAPA introduces a novel layer of analysis, identifying clusters where the APA event occurs in different direction (3′-UTR shortening or lengthening) and in different degree (see Methods, **Fig. 1A, B**). Since this step will identify the clusters with different APA patterns for the gene, it is called gene-cluster-level identification.

To assess the impact of the modeling, we developed another approach based on the same module of DaPars, termed test-based scMAPA. Based on the estimates of the long/short isoforms in each cluster-bulk data, test-based scMAPA tests independence of the isoforms in all pairwise comparisons of the clusters. Finally, it defines significant APA genes if they pass significance test in any pair of the clusters. To distinguish from test-based scMAPA, we will refer to scMAPA with regression model as model-based scMAPA. All the tests used FDR (B-H) < 0.05 in the subsequent analyses.

scMAPA identify true APA events with an enhanced statistical power

To compare statistical power of scMAPA (**S. Fig. 1**), we first considered two previous methods for scRNA-Seq data, scAPA and scDAPA. However, scDAPA assumes case-control setting, so cannot be directly tested in the multi-cluster setting. Thus, we will compare two versions of scMAPA (test- and model-based) only to scAPA in the simulation data generated as follows. We first generated the gene expression matrix of 5,000 genes over 3,000 cells in 5 clusters using Splatter [23] by estimating the parameters from mouse brain data consisting of 5 main cell types collected from brain cortex and dorsal midbrain [20]. Then, we determined the standard deviation of the 3′-UTR long/short isoform proportion across the clusters separately for genes with dynamic APA (APA genes) and non-APA genes. Based on the long/short isoform proportion and the gene expression values, we generated abundance of long and short isoforms (see Methods) for each gene in each cluster under different simulation scenarios. The three simulation scenarios vary three factors: 1) standard deviation (SD) of the isoform proportion values across clusters ($SD_{isoprop}$), 2) number of APA genes, and 3) distribution of cell cluster size.

In the first scenario, we varied only the first factor, SD of the isoform proportion values across clusters ($SD_{isoprop}$), for APA genes while fixing i) $SD_{isoprop}$ for non-APA genes to be the same across the clusters, ii) the number of APA genes to be 500 (10 % of the total genes) and iii) the uniform distribution of cluster size (20% of the cells in each cell group). This scenario is motivated by our analysis as follows. In the mouse brain data, we selected the APA genes that were identified by both scAPA and model-based scMAPA. Then, we estimated their $SD_{isoprop}$ values. We did the same for non-APA genes. We found that $SD_{isoprop}$ is significantly (p value < 2.2×10^{-16}) higher in APA genes than in non-APA genes (**S. Fig. 2A**), indicating that the 3′-UTR long and short isoform proportion values spread more drastically across the clusters in APA genes than in non-APA genes. On the simulated APA genes and non-APA genes with $SD_{isoprop}$ values separately estimated from mouse brain data (see Methods), we ran the statistical tests employed by scMAPA (a logistic regression model + Likelihood ratio test for model-based or Fisher's exact test for pairwise) with scAPA (Pearson's χ^2) (**Fig. 2A**). The result shows that the statistical components of scMAPA identify more true APA genes than that of scAPA in all simulated $SD_{isoprop}$ values. Especially, the sensitivity of scMAPA is around 20% higher than that of scAPA except in very high $SD_{isoprop}$ values (e.g. 0.18), showing that statistical components of scMAPA are more sensitive at identifying subtle dynamic APA. Since all three methods perform equally good at identifying true non-APA genes in simulated $SD_{isoprop}$ values (scAPA 0.26% higher than model-based scMAPA and 0.33% higher than test-based scMAPA on average, **Fig. 2B**), scMAPA, either model-based or test-based, outperforms scAPA overall. In the second scenario, we varied the number of true APA genes and the distribution of cell cluster size simultaneously while fixing $SD_{isoprop}$ values for APA and non-APA genes (see Methods). With 500 (10% of the total genes) true APA genes, the statistical components of scMAPA consistently outperform that of scAPA in all three cluster size distributions (a: 20% of the cells for all clusters, b: 50% for one and 12.5% for all the others, and c: 30% for one and 17.5% for all the others) (**Fig. 2C**) with a slight loss of specificity (**Fig. 2D**). This trend holds true with 250 and 1,000 true APA genes simulated (**S. Fig. 2B, C, D, E**). Together, the statistical components of scMAPA identify true APA events with an enhanced statistical power compared to that of scAPA.

The enhanced statistical power of scMAPA facilitates more detailed understanding of APA dynamics

To compare the performance as a whole, we ran all three methods on the public data generated from 10x Chromium scRNA-Seq experiment on human Peripheral Blood Monocellular Cells (PBMC) (https://www.10xgenomics.com/, 10k PBMCs from a Healthy Donor in v3 chemistry). PBMCs are parts of the immune system critical to cell-mediated and humoral immunity, including T-Cells, B-Cells, monocytes and NK-Cells. Together with the definition of five cell clusters available in the 10x database as input, model-based scMAPA identifies dynamic APA genes distinctive to those of scAPA (31.3-fold less in common than unique findings combined, **Fig. 3A**). To test whether this result is due to scMAPA's high statistical power or high false positive, we inspected the APA genes and non-APA genes identified by scMAPA and scAPA. With the proportion of the long/short isoforms across clusters ($SD_{isoprop}$) as an indicator of APA heterogeneity across cluster, APA genes by scMAPA show significantly higher $SD_{isoprop}$ than those by scAPA (p value < 2.2×10^{-16}, **Fig. 3B**). PBMCs are known for extreme heterogeneity partly due to its nature being a mixture of different cell types [24,25]. Since this heterogeneity likely also affects the APA level, we believe that there should be APA genes with high $SD_{isoprop}$ values. Further, our simulation showed that scAPA is not as sensitive as scMAPA to identify true APA genes in a wide range of $SD_{isoprop}$ values (0.06 to 0.18, **Fig. 2A**). Since this range coincides with the range of $SD_{isoprop}$ values where scMAPA found much more true APA genes, we believe that more findings of scMAPA in the PBMC data are attributable to its greater sensitivity, not high false positive.

To demonstrate biological implications brought by the high sensitivity of scMAPA, we ran Ingenuity Pathway Analysis (IPA) on 5,192 and 162 APA genes identified by scMAPA and scAPA respectively. While 500 IPA "Disease & Function" terms are significantly enriched (B-H P-Val < 0.05) for scMAPA APA genes (**S. Table 1**), 82 terms were enriched (B-H P-Val < 0.05) for scAPA APA genes (**S. Table 2**). Although this difference of the number of enriched terms is expected due to the difference of the number of input genes, it is interesting to note that only scMAPA APA genes include 14 terms with keyword "hemato" (**Fig. 3C**). Those terms are with additional keywords representing important biological function: 5 with "hematopoiesis", "development" or "differentiation", 5 with "cell death" or "apoptosis", 4 with "cancer" or "neoplasm". This result shows that scMAPA will help elucidate dynamic APA contributing to those important functions in hematology.

Further analyses attribute some of the scMAPA sensitivity to the underlying model. Since both of the scMAPA versions (model- or test-based) use DaPars modules to quantify 3′-UTR isoforms, they identify a high overlap of significant APA genes (6.5-fold more in common than unique findings combined, **Fig. 3D**). However, model-based scMAPA uniquely identifies 13.5-fold more significant APA genes than test-based scMAPA's unique identification. For example, model-based scMAPA uniquely identifies Myocyte Enhancer Factor 2D (MEF2D) with significant dynamic APA (B-H P-val < 0.05), as the RNA read density on the leftmost (the last 3′-UTR region) part of the backward stranded gene is 1.73-fold higher in cluster 4 vs. cluster 1 (**Fig. 3E**). The MEF2D transcription factor has essential roles in diverse biological conditions including blood and immune cell development[26] with its alteration implied for blood disorders (e.g. [27–29]). Since different proportions of the long and short 3′-UTR isoforms characterize certain pathological conditions, e.g. glioblastoma multiforme vs. normal brain tissues [30], we believe that the different proportion of the isoforms in cluster 4 vs. cluster 1 is associated with the different functions of the clusters. For 672 other APA genes uniquely identified by model-based scMAPA (e.g. PECAM1 in **S. Fig. 3**), our manual inspection suggests that the corresponding significance estimation module (regression + LRT) detects such difference in RNA alignment density in the 3′-UTR with more sensitivity than test-based scMAPA for real biological data.

scMAPA estimates multiple clusters for significant APA events

Model-based scMAPA has a novel layer of dynamic APA analysis not available in scAPA and scDAPA, identifying cell clusters with different patterns of dynamic APA (**Fig. 1B**). To do this for the genes identified with dynamic APA, we first identified such genes in mouse brain data (cortex and dorsal midbrain)[20] using scMAPA. Among 6 cell types defined in the data, we selected five main cell types with large sample size: neurons, astrocytes, immune cells, oligodendrocytes and vascular (see Methods). Across the five cell types, scMAPA identified 2,682 (35.5%) significant dynamic APA out of 7,560 transcripts expressed in > 3 cells and of which the sum of CPM > 10 across cells (**S. Table 3**). Our result is consistent with the report of scAPA on the same data (2,506 transcripts reported with significant dynamic APA). In the gene-level identification, model-based scMAPA identified significant APA genes in a high overlap with test-based scMAPA (6.5-fold more in common than unique findings combined, **Fig. 4A**), distinct to scAPA (8.2-fold less in common than unique findings combined, **Fig. 4B**), consistent with the PBMC data analysis.

On the significant APA genes returned from the gene-level identification, we ran the gene-cluster-level identification module of scMAPA that estimates the coefficients representing the degree and the direction of APA events in each cluster. Running hierarchical clustering on the resulting coefficients by cell type, we found that immune cells and neuron cells are most distinguished from the other cell types (**Fig. 4C**). While this finding supports the previous finding of scAPA that neuronal cells and brain immune cells are most different in the APA pattern [31], scMAPA uniquely identified a distinct tendency of either 3′-UTR shortening or lengthening in each cell type (**Fig. 4D**). Neuron cells are characterized with 3′-UTR lengthening, while immune and vascular cells are characterized with 3′-UTR shortening. This tendency not only reiterates the reported dominance of 3′-UTR lengthening in neuron cells [32–35], but also shows how the cell type specificity of the APA landscape[1,12] actually appears in mouse brain. To understand the functional implication of the cell type specificity, we selected APA genes (3′-UTR shortening or lengthening separately) uniquely identified in each cell type. By running g:Profiler [36] on them, 3′-UTR shortening or lengthening together, we identified significantly enriched terms (B-H p-val. < 0.05, **S. Table 4**) for each cell type. In either GO biological process (BP) or cellular component (CC), several enriched terms suggest that the APA genes are involved in the biological functions unique to each cell type. For example, APA genes unique in neuron cells are enriched for 3 GO cellular component (CC) terms with keyword “endoplasmic reticulum” and 1 with “myelin sheath”. Neuronal cytoplasm embeds various types of organelles, a major component of which is endoplasmic reticulum. Also, the myelin sheath is a lipid that wraps around nerve fibers and serves to increase the speed of neuronal electrical communication. So, APA genes' enrichment for "endoplasmic reticulum" and "myelin sheath" cellular components suggests that dynamic APAs add another dimension of regulation to maintain complex endoplasmic reticulum and myelin sheath biology in neurons. For example, CDC42 is known to affect myelin stabilization [37]. Its significant 3′-UTR lengthening found only in neurons suggests that CDC42 may play its roles partly in association with the APA signal. Quite strikingly, the GO terms with keyword "endoplasmic reticulum" did not come up in other cell types except oligodendrocytes. Since oligodendrocytes must synthesize an enormous amount of myelin membrane proteins, cholesterol, and membrane lipids through the secretory pathway, particularly the homeostasis of the endoplasmic reticulum [38], we believe that dynamic APA helps facilitate the specific biological functions of myelin formation involving the endoplasmic reticulum. Since this APA signal is independent of the expression information of the genes (**S. Fig. 4A, B, C, D, E, F**), the results suggest that the dynamic APA plays biological roles to differentiate the cell types in an independent manner on expression signals.

scMAPA controls confounding factors

Model-based scMAPA enables to control confounders. Confounding arises when cells are affected by the factors that are not a part of the research hypothesis under investigation. In that sense, it is often important to consider confounders in the scRNA-Seq data analysis where multiple factors affect the molecular dynamics of individual cells differently. To demonstrate the effect of such a consideration, we first split 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 these two factors are non-ordinal categorical variables that are addressed usually by splitting data. In the resulting cluster-bulk RNA-Seq data, scMAPA identified 1,018 transcripts with significant dynamic APA (**S. Table 5**) with all the other parameters same as previous. Note that the previous run of scMAPA, based on the cluster-bulk data split only by cell type, identified 2,682 APA transcripts **(S. Table 3**). Since the runs differ only in how the data are divided, it is reasonable to see a high overlap between the runs (the left Venn diagram in **Fig. 5A**). Using the run with the further split as reference, we ran another scMAPA that considers a confounder variable for brain region, identifying 881 transcripts with significant dynamic APA (the right Venn diagram in **Fig. 5A, S. Table 6**). Comparing to the dynamic APA transcripts in the reference, 163 transcripts are left representing the association caused by the confounding effect of brain region. To check if they are indeed associated with brain region, we ran Ingenuity Pathway Analysis (IPA) upstream regulator analysis on the 108 and 682 genes corresponding to the 163 and 881 transcripts, respectively (region-associated and type-associated APA genes respectively, **S. Table 7**). Since the enrichment significance would represent how statistically confident particular IPA upstream regulators regulate the input genes, we hypothesized that the region-associated APA genes are more enriched for the IPA upstream regulators that play region-specific functions. We found that the 5 IPA upstream regulators more enriched for the region-based APA genes support our hypothesis (**Fig. 5B**). For example, NUAK1 is heavily involved in the development of a specific mouse brain region, the cerebral cortex[39] likely by targeting downstream target genes. Also, TAF1 also regulates downstream genes in regulating the morphology and function of mouse brain regions, the cerebellum and the cerebral cortex[40]. Our result shows that the target genes undergo dynamic APA in a region specific manner. Since this data were collected from mouse brain regions including the cortex, these results collectively show the contribution of dynamic APA specific to the cortex region.

Since these results are unique findings of scMAPA due to its model-based property, they demonstrate the importance of the confounder consideration and, thus, highlight a unique contribution of scMAPA.

DISCUSSION

APA is a type of post-transcriptional regulation emerging as an important layer for transcriptomic diversity in physiological and pathological conditions. With the cell type specificity [1,41], there have been a couple of computational methods identifying APA events in scRNA-Seq data. However, we realized the need for improvement for better statistical power and cell cluster-wise interpretation. In this work, we bring the improvement in three major ways. First, we developed scMAPA that identifies genes with significant APA events (gene-level identification) with a better sensitivity than the previous method. Second, we devised a *de novo* simulation framework where an essential part of the APA detection methods, the statistical component, can be objectively compared. Although scMAPA consistently outperforms the previous method, it is important to note that this result does not necessarily imply that scMAPA would outperform scAPA as a whole, since this simulation does not cover the first step of the methods of quantifying 3′-UTR long and short isoforms. Third, we enable a new type of APA analysis in multi-cluster setting, identifying multiple clusters in which each gene shortens or lengthens the 3′-UTRs with significance (gene-cluster-level identification). To demonstrate these improvements, we used simulation data of various simulation scenarios and biological data of different biological contexts (human PBMC and mouse brain data).

Another contribution of this work is the decomposition of the APA detection algorithms into two steps: quantifying 3′-UTR long/short isoforms and estimating statistical significance of the isoform proportion. This decomposition brings out two advantages for further development of scRNA-Seq APA detection tools. First, it enables to compare the statistical power of various APA detection methods step by step, which establishes our simulation framework. Second, although scMAPA currently utilizes the DaPars module for the isoform quantification, it can easily adapt other pairwise APA detection methods to carry out the step. This implies that scMAPA can improve simply by replacing the DaPars module with other tools with a better sensitivity whenever such tools emerge.

Currently, both the simulation platform and scMAPA have limitations. First, our simulation platform does not cover how APA detection tools quantify 3′-UTR long/short isoforms from RNA alignment density information. However, it is critical to evaluate efficiency of the step to evaluate the tools in totality. Secondly, we studied scRNA-Seq data of 5 cell clusters at most, partly because many scRNA-Seq data have around that number of main cell types investigated. It is important to study performance of both scMAPA and our simulation platform in scRNA-Seq data with more than 5 cell clusters. Thirdly, scMAPA, like other scRNA-Seq APA detection tool, bases its identification on the pooling of scRNA-Seq data with respect to cell clusters that we call bulk RNA-Seq data. To study the APA dynamics in the single cell resolution, much advance is needed both in the biochemistry and bioinformatics techniques. One of our main tasks for future research is to study the potential of the current chemistry of scRNA-Seq data (e.g. version 3 chemistry in 10x Chromium) in detecting dynamic APA in the single cell resolution.

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.

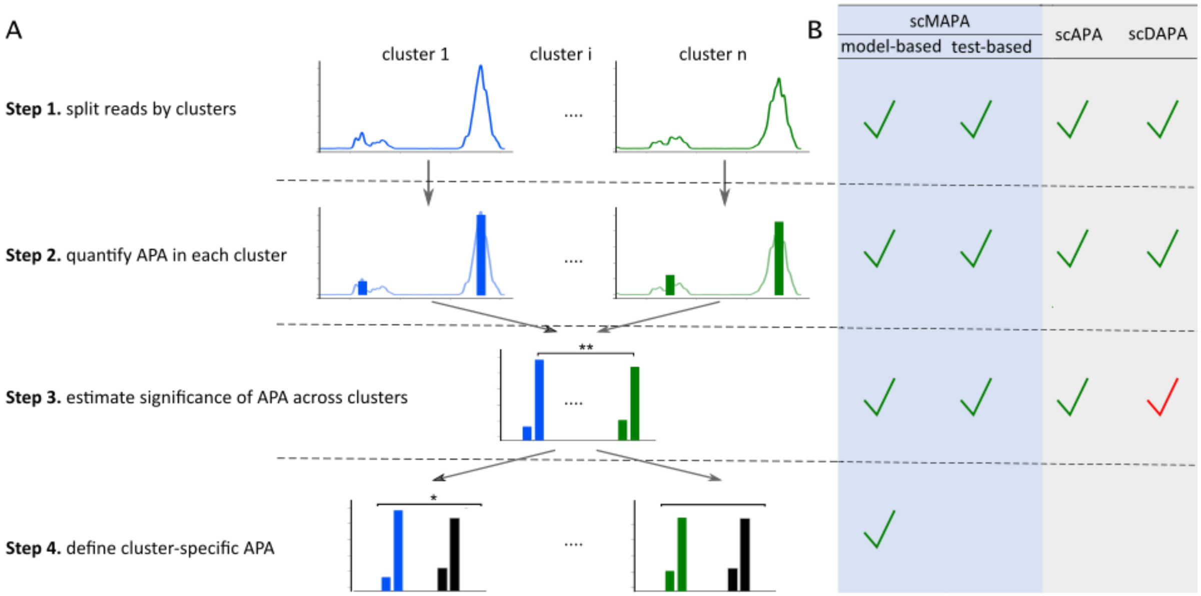
Acknowledgements We thank Daniel Weeks, Ph.D., Professor, Department of Human Genetics, University of Pittsburgh for valuable discussion on the novelty of scMAPA. This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided.

Author Contributions H.J.P and Y.B. conceived the project, designed the experiments and implemented the software. Y.B. and Y.Q. performed the analysis. S.K., K.N., 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. The Python script used to split bam file was shared by Dr. Ming Tang at http://doi.org/10.5281/zenodo.3946832.

Figures


A
cluster 1
cluster i
cluster n
Step 1. split reads by clusters
Step 2. quantify APA in each cluster
Step 3. estimate significance of APA across clusters
Step 4. define cluster-specific APA
....
**
*
B

	scMAPA		scAPA	scDAPA
	model-based	test-based		
Step 1	✓	✓	✓	✓
Step 2	✓	✓	✓	✓
Step 3	✓	✓	✓	✓
Step 4	✓			

Figure 1. Characteristics of APA detection tools for scRNA-Seq data. (A) 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. (B) table showing the capability of the scRNA-Seq APA detection algorithms for the corresponding step. The green checkmarks indicate the full capability and the red checkmark for scDAPA at Step 3 indicates a partial capability in that it is designed for two-cluster, not multi-cluster setting.

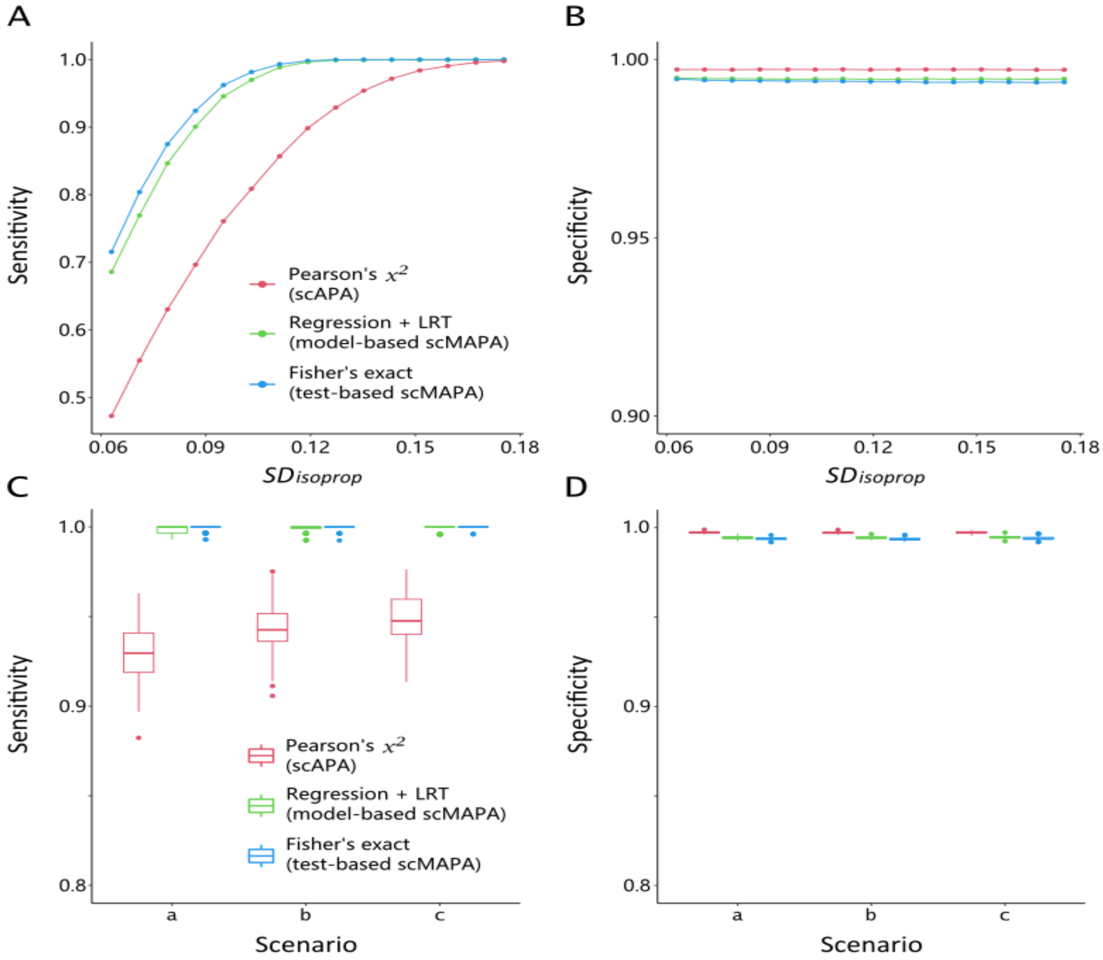


Figure 2. Performance assessment using simulated data. With fixed number of true APA events (500 out of 5000) and uniform distribution of cell cluster size (600 cells in each cell type), (A) sensitivity and (B) specificity were plotted against varying degree of standard deviation (SD) of PDUI values 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), (C) sensitivity and (D) 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.

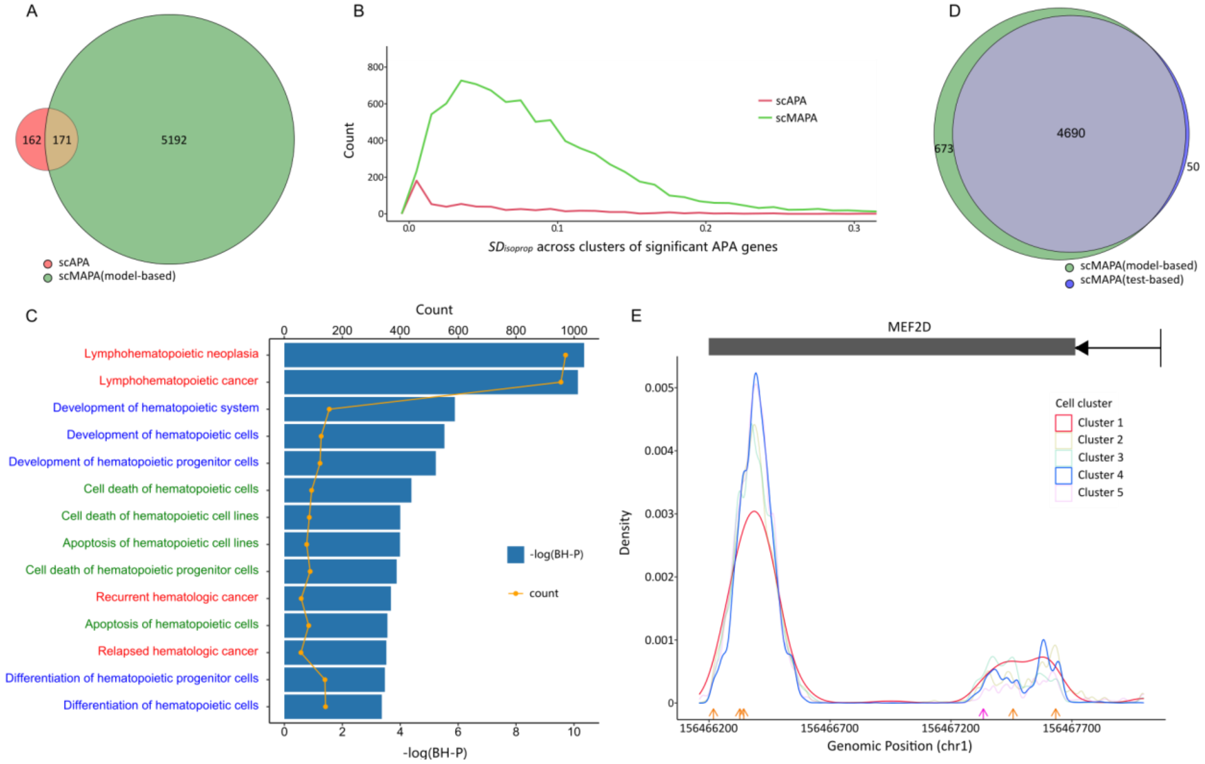


Figure 3. Performance assessment of scMAPA and scAPA using PBMC data. (A) Venn diagram of significant APA genes detected by scMAPA and scAPA. (B) Frequency polygon plot shows the distribution of standard deviations (SD) of PDUI values across clusters ($SD_{isoprop}$) for significant APA genes. (C) Bar plot shows that the significant APA genes identified by scMAPA is significantly enriched with “hemato”-related “Disease & Function” IPA terms. The yellow line represents the number of APA genes that fall into molecule list of each term. The blue bar represents the -log10(p values) from enrichment test. Terms are colored based on the additional associated keyword: red for neoplasm/cancer, blue for development/hematopoiesis, green for cell death/apoptosis. (D) Venn diagram of significant APA genes detected by model-based and test-based scMAPA. (E) Coverage plot of gene MEF2D illustrates the dynamic of APA among the cell clusters. Orange arrows on the bottom indicate the polyA site annotated in polyASite database. Bar on top marks the boundary of 3′-UTR region of MEF2D.

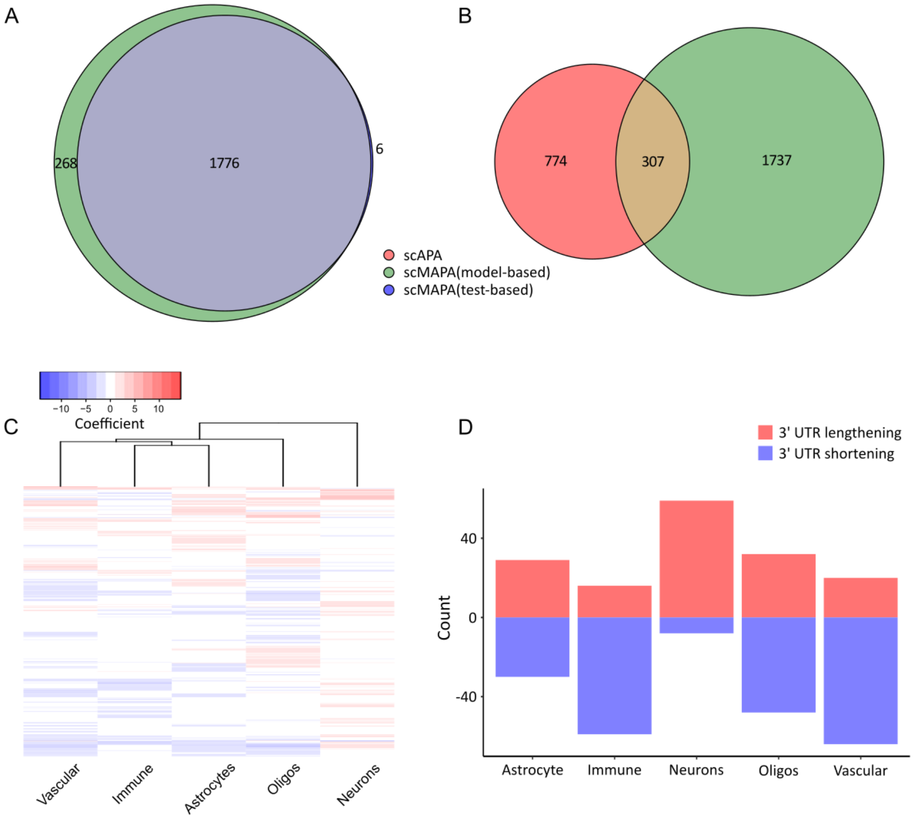


Figure 4. Gene-level and gene-cluster-level identification using mouse brain data. (A) Venn diagram of significant APA genes detected by scMAPA and scAPA. (B) Venn diagram of significant APA genes detected by model-based and test-based scMAPA. (C) Heatmap of coefficients of cell type-specific APA genes. Coefficients were estimated in logistic regression model. (D) Bar plot shows the number of 3’ UTR lengthening and shortening detected in each cell type.

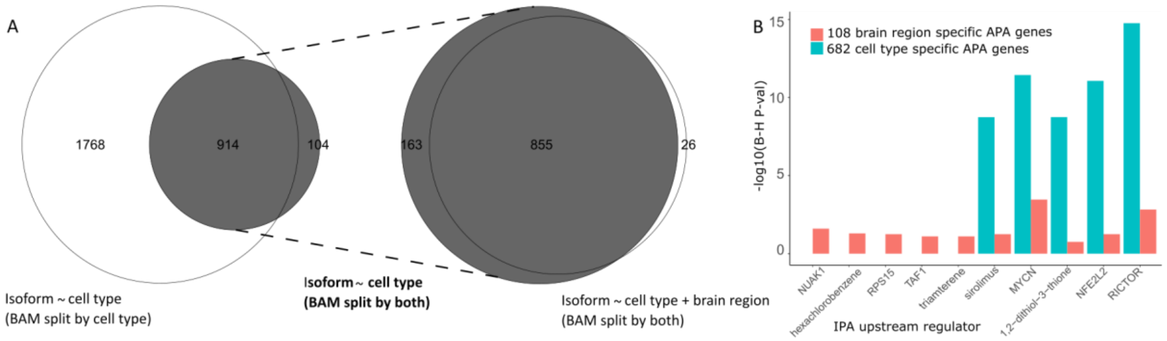
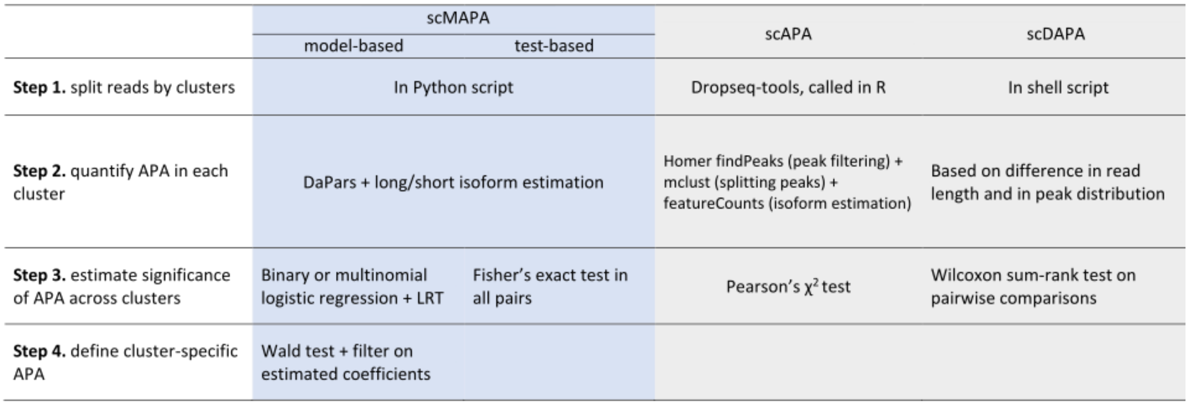


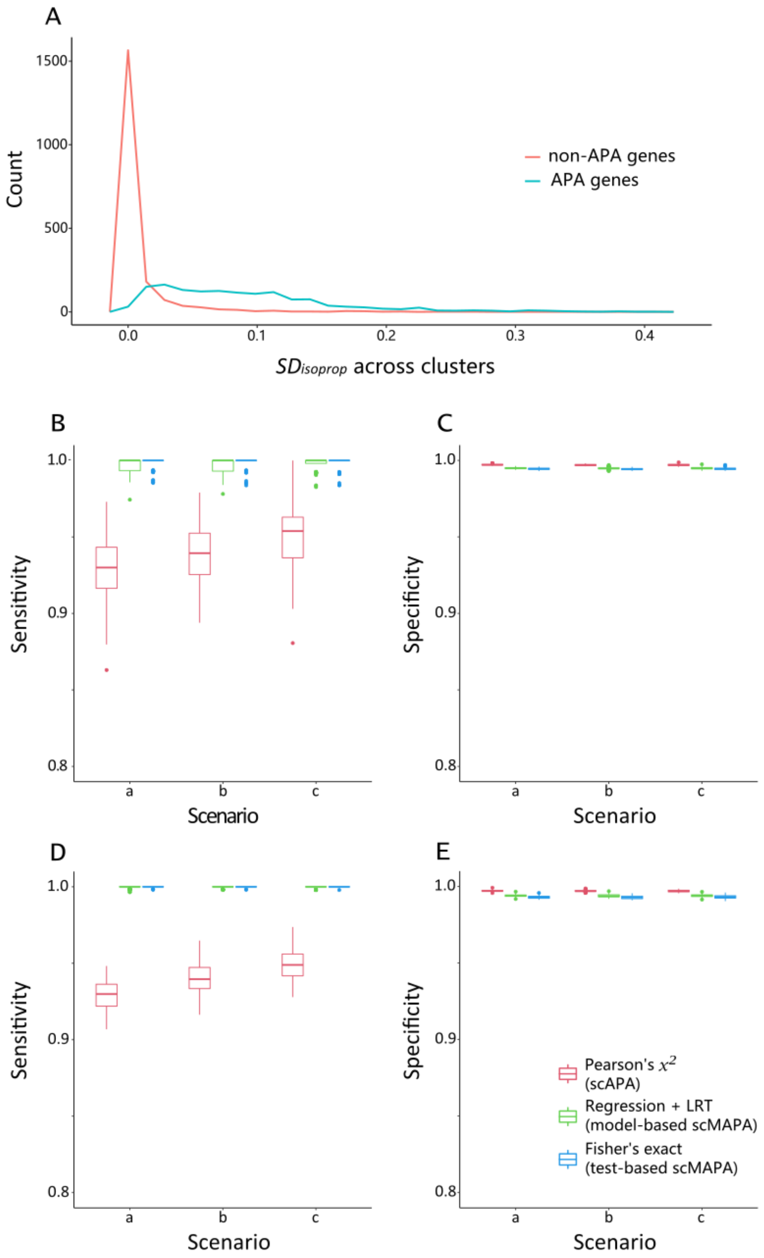
Figure 5. Venn diagrams show the significant APA identification by scMAPA before and after adjusting for the brain region. (A) In the left venn diagram, the left circle represents the identification with the input BAM file split by only cell type. The right circle represents the identification with the input BAM file split by both cell type and brain region. Only cell type was considered as the independent variable in both runs. With the right circle in the left venn diagram as the reference for further process in the right venn diagram (colored in gray), the right circle represents the identification with both cell type and brain region as independent variables. (B) Significance of IPA enrichment terms most distinguishing region- and type-associated APA genes ranked by the significance difference in region-based vs. type-based APA genes.

Supplemental Figures

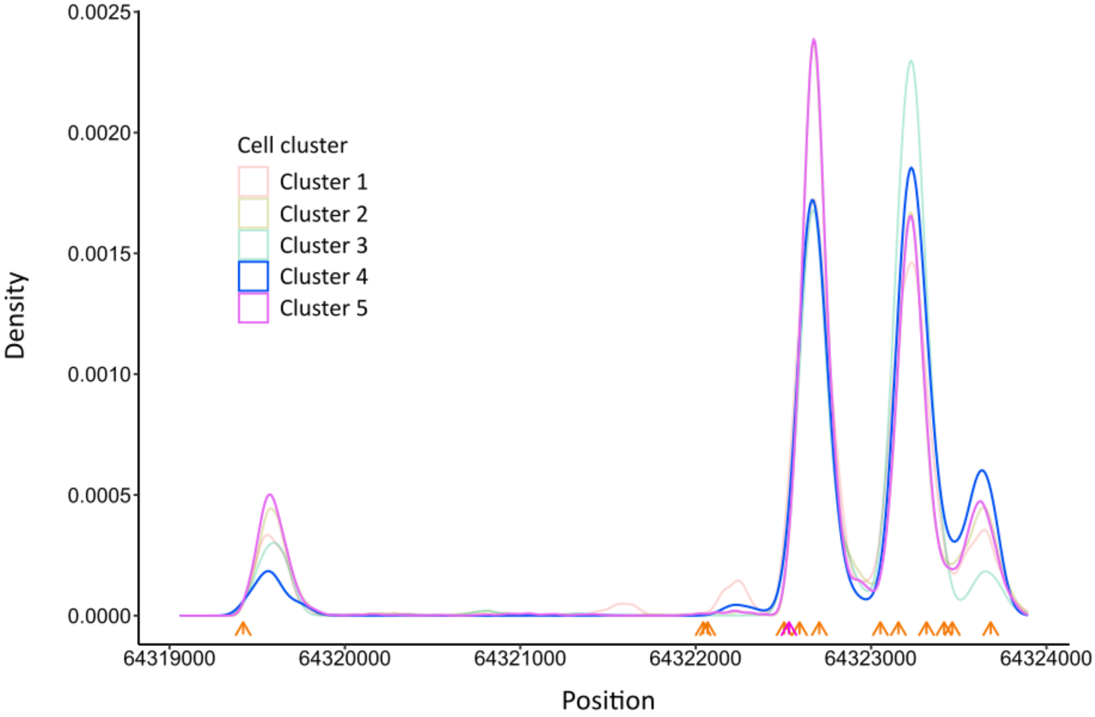


	scMAPA		scAPA	scDAPA
	model-based	test-based		
Step 1. split reads by clusters	In Python script		Dropseq-tools, called in R	In shell script
Step 2. quantify APA in each cluster	DaPars + long/short isoform estimation		Homer findPeaks (peak filtering) + mclust (splitting peaks) + featureCounts (isoform estimation)	Based on difference in read length and in peak distribution
Step 3. estimate significance of APA across clusters	Binary or multinomial logistic regression + LRT	Fisher's exact test in all pairs	Pearson's χ^2 test	Wilcoxon sum-rank test on pairwise comparisons
Step 4. define cluster-specific APA	Wald test + filter on estimated coefficients			

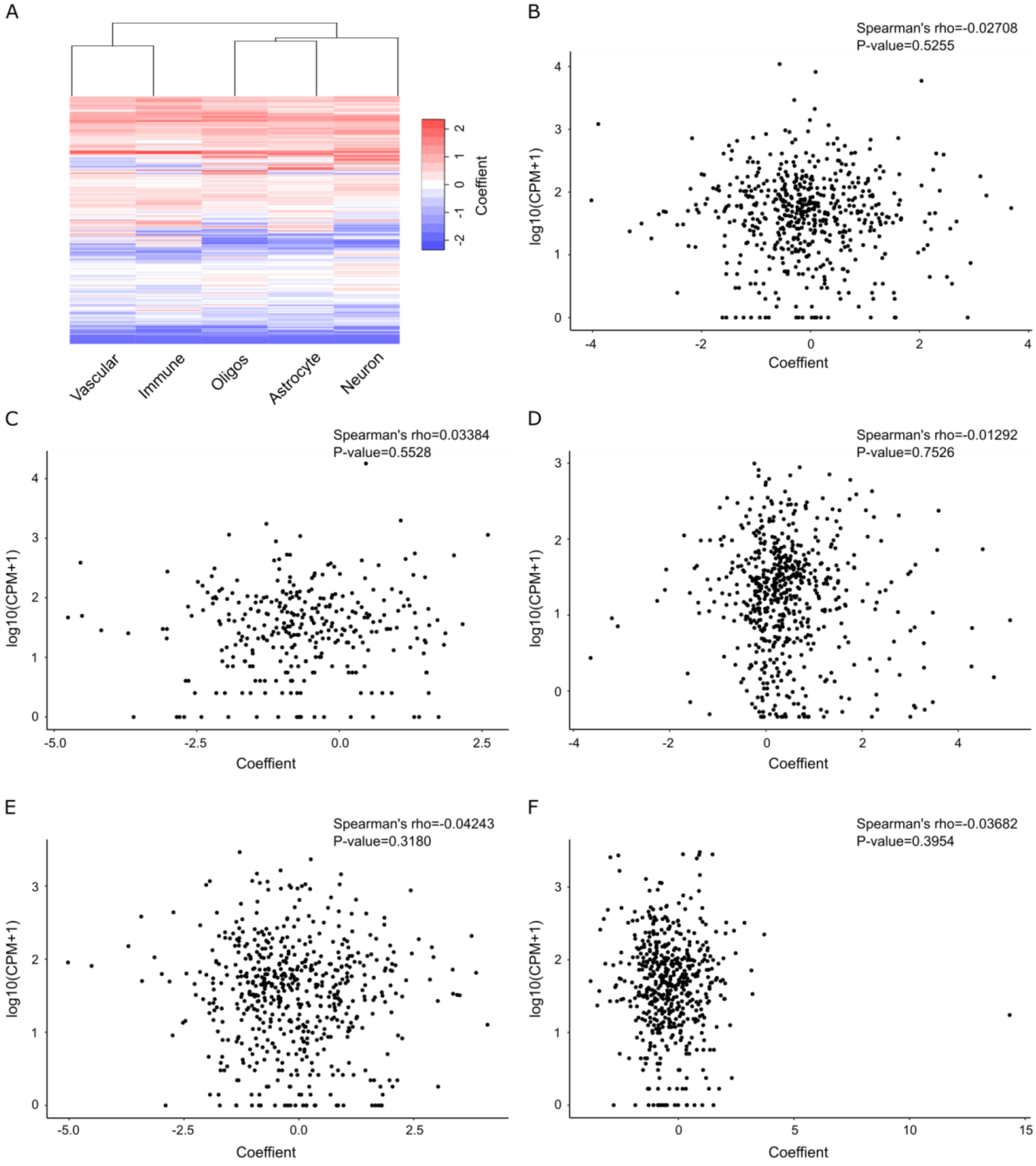
Supplemental Figure 1. Algorithm overview of bioinformatic tools and statistical methods to identify dynamic APAs in scRNA-Seq data



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 1000 and all other factors remain same.



Supplemental Figure 3. Performance assessment of scMAPA and scAPA using PBMC data. (A) Coverage plot of gene PECAM1 illustrates the dynamic of APA among cell types. Orange arrows indicate the polyA site annotated in polyASite database. Purple arrow shows the proximal polyA site predicted by DaPars. Two numbers above the arrows mark the boundary of 3’UTR region.



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.

REFERENCES

1. Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22**, 1173–1183 (2012).

2. Masamha, C. P. *et al.* CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* **510**, 412–416 (2014).

3. Chen, W. *et al.* Alternative Polyadenylation: Methods, Findings, and Impacts. *Genomics. Proteomics Bioinformatics* **15**, 287–300 (2017).

4. Chen, M. *et al.* 3 ′ UTR lengthening as a novel mechanism in regulating cellular senescence. *Genome Res.* **28**, 285–294 (2018).

5. Dimri, G. P. *et al.* A biomarker that identifies senescent human cells in culture and in aging skin in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 9363–7 (1995).

6. Busuttil, R. A., Rubio, M., Dollé, M. E. T., Campisi, J. & Vijg, J. Oxygen accelerates the accumulation of mutations during the senescence and immortalization of murine cells in culture. *Aging Cell* **2**, 287–294 (2003).

7. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, (2013).

8. Xia, Z. *et al.* Dynamic Analyses of Alternative Polyadenylation from RNA- Seq Reveal Landscape of 3 ’ UTR Usage Across 7 Tumor Types. *Nat. Commun.* 1–38 (2014).

9. Park, H. J. *et al.* 3′ UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk. *Nat. Genet.* **50**, 783–789 (2018).

10. Kim, S., Bai, Y., Diergaarde, B., Tseng, G. C. & Park, H. J. Alternative Polyadenylation Modifies Target Sites of MicroRNAs with Clinical Potential for Breast Cancer Progression. *bioRxiv* 601518 (2019) doi:10.1101/601518.

11. Fan, Z. *et al.* 3′-UTR shortening contributes to subtype-specific cancer growth by breaking stable ceRNA crosstalk of housekeeping genes. *Front. Bioeng. Biotechnol.* **to appear**, 601526 (2020).

12. Zhang, H., Lee, J. Y. & Tian, B. Biased alternative polyadenylation in human tissues. *Genome Biol.* **6**, R100 (2005).

13. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* **27**, 2380–2396 (2013).

14. Xiang, Y. *et al.* Comprehensive Characterization of Alternative Polyadenylation in Human Cancer. *JNCI J. Natl. Cancer Inst.* **110**, 1–11 (2017).

15. Ye, C. *et al.* scDAPA: detection and visualization of dynamic alternative polyadenylation from single cell RNA-seq data. *Bioinformatics* **36**, 1262–1264 (2020).

16. Shulman, E. D. & Elkon, R. Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic Acids Res.* **47**, 10027–10039 (2019).

17. Nott, A. *et al.* Brain cell type–specific enhancer–promoter interactome maps and disease&lt;strong&gt;-&lt;/strong&gt;risk association. *Science (80-.).* **366**, 1134 LP – 1139 (2019).

18. Doorn, K. J. *et al.* Brain region-specific gene expression profiles in freshly isolated rat microglia. *Front. Cell. Neurosci.* **9**, 84 (2015).

19. McKenzie, A. T. *et al.* Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Sci. Rep.* **8**, 8868 (2018).

20. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014.e22 (2018).

21. Bohnert, R. & Rätsch, G. rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res.* **38**, W348-51 (2010).

22. Xiang, Y. *et al.* Comprehensive Characterization of Alternative Polyadenylation in Human Cancer. **110**, 1–11 (2018).

23. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).

24. Appleby, L. J. *et al.* Sources of heterogeneity in human monocyte subsets. *Immunol. Lett.* **152**, 32–41 (2013).

25. Schildberger, A., Rossmanith, E., Eichhorn, T., Strassl, K. & Weber, V. Monocytes, peripheral blood mononuclear cells, and THP-1 cells exhibit different cytokine expression patterns following stimulation with lipopolysaccharide. *Mediators Inflamm.* **2013**, 697972 (2013).

26. Pon, J. R. & Marra, M. A. MEF2 transcription factors: developmental regulators and emerging cancer genes. *Oncotarget* **7**, 2297–2312 (2016).

27. Gu, Z. *et al.* Genomic analyses identify recurrent MEF2D fusions in acute lymphoblastic leukaemia. *Nat. Commun.* **7**, 13331 (2016).

28. Herglotz, J. *et al.* Essential control of early B-cell development by Mef2 transcription factors. *Blood* **127**, 572–581 (2016).

29. Ohki, K. *et al.* Clinical and molecular characteristics of MEF2D fusion-positive B-cell precursor acute lymphoblastic leukemia in childhood, including a novel translocation resulting in MEF2D-HNRNPH1 gene fusion. *Haematologica* **104**, 128–137 (2019).

30. Shao, J. *et al.* Alternative polyadenylation in glioblastoma multiforme and changes in predicted RNA binding protein profiles. *OMICS* **17**, 136–149 (2013).

31. Hilgers, V., Lemke, S. B. & Levine, M. ELAV mediates 3’ UTR extension in the Drosophila nervous system. *Genes Dev.* **26**, 2259–2264 (2012).

32. Stark, A., Brennecke, J., Bushati, N., Russell, R. B. & Cohen, S. M. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3’UTR evolution. *Cell* **123**, 1133–1146 (2005).

33. Shepard, P. J. *et al.* Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**, 761–772 (2011).

34. Hilgers, V. *et al.* Neural-specific elongation of 3' UTRs during Drosophila development. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 15864–15869 (2011).

35. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).

36. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193-200 (2007).

37. Benninger, Y. *et al.* Essential and distinct roles for cdc42 and rac1 in the regulation of Schwann cell biology during peripheral nervous system development. *J. Cell Biol.* **177**, 1051–1061 (2007).

38. Lin, W. & Popko, B. Endoplasmic reticulum stress in disorders of myelinating cells. *Nat. Neurosci.* **12**, 379–385 (2009).

39. Courchet, V. *et al.* Haploinsufficiency of autism spectrum disorder candidate gene NUAK1 impairs cortical development and behavior in mice. *Nat. Commun.* **9**, 4289 (2018).

40. Janakiraman, U. *et al.* TAF1-gene editing alters the morphology and function of the cerebellum and cerebral cortex. *Neurobiol. Dis.* **132**, 104539 (2019).

41. Zhang, H., Lee, J. Y. & Tian, B. Biased alternative polyadenylation in human tissues. *Genome Biol.* **6**, R100 (2005).

S. Table 1. Detailed information of significantly enriched "Disease & Function" terms from Ingenuity Pathway Analysis (IPA) analysis on APA genes detected by scMAPA on the PBMC data.

S. Table 2. Detailed information of significantly enriched "Disease & Function" terms from Ingenuity Pathway Analysis (IPA) analysis on APA genes detected by scAPA on the PBMC data.

S. Table 3. scMAPA estimates on the 2,682 transcripts with significant dynamic APA in the gene level identification of the scMAPA run.

S. Table 4. Analysis results from g:Profiler on tissue-specific APA events identified by scMAPA on the mouse brain data.

S. Table 5. scMAPA estimates on the 1,018 transcripts with significant APA identified by scMAPA. Input BAM file is split by cell type and brain region, but the regression model included cell type as the only independent variable.

S. Table 6. scMAPA estimates on the 881 transcripts with significant APA identified by scMAPA. Input BAM file is split by cell type and brain region, and the regression model included cell type as the only independent variable and brain region as a confounding factor.

S. Table 7. IPA analysis result (enrichment p-value) on 108 and 682 genes corresponding to the 163 region-associated and 881 type-associated APA events.