Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression

Yipeng Gao[1,2], Lei Li[3], Christopher I. Amos[2], Wei Li[3*]

[1]Graduate Program in Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX USA

[2]Department of Medicine, Baylor College of Medicine, Houston, TX USA

[3]Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA USA

* Correspondence: wei.li@uci.edu

Abstract

Alternative polyadenylation (APA) is a major mechanism of post-transcriptional regulation in various cellular processes including cell proliferation and differentiation, but the APA heterogeneity among single cells remains largely unknown. Single-cell RNA sequencing (scRNA-seq) has been extensively used to define cell subpopulations at the transcription level. Yet, most scRNA-seq data have not been analyzed in an "APA-aware" manner. Here, we introduce scDaPars (**D**ynamic Analysis of **A**lternative **P**oly**A**denylation from **S**ingle-**c**ell **R**NA-**s**eq), a bioinformatics algorithm to accurately quantify APA events at both single-cell and single-gene resolution using either 3' end (10x Chromium) or full-length (Smart-seq2) scRNA-seq data. Validations in both real and simulated data indicate that scDaPars can robustly recover missing APA events caused by the low amounts of mRNA sequenced in single cells. When applied to cancer and human endoderm differentiation data, scDaPars not only revealed cell-type-specific APA regulation but also identified cell subpopulations that are otherwise invisible to conventional gene expression analysis. Thus, scDaPars will enable us to understand cellular heterogeneity at the post-transcriptional APA level.

Keywords

Alternative Polyadenylation, Single-cell RNA-sequencing, Single-cell Genomics, Imputation

Introduction

Alternative polyadenylation (APA) is a major mechanism of post-transcriptional regulation under diverse physiological and pathological conditions (Elkon et al. 2013; Tian and Manley 2017). The process of polyadenylation involves endonucleolytic cleavage of the nascent RNA followed by synthesis of a poly(A) tail on the 3' terminus (Tian and Manley 2017). By using different polyadenylation sites (poly(A) sites), which are defined by flanking RNA sequence motifs, APA can generate mRNA isoforms with various 3'-untranslated regions (3' UTRs) in the majority of human genes (Derti et al. 2012; Tian and Manley 2017). While APA in most cases does not alter the protein-coding regions in those mRNA isoforms, it disrupts important *cis*-regulatory elements located in the 3' UTRs, including adenylate-uridylate-rich elements (ARE) and binding sites of miRNAs and RNA-binding proteins, resulting in altered mRNA stability, localization and translation efficiency (Garneau et al. 2007; An et al. 2008; Hoffman et al. 2016).

High-throughput sequencing technologies have revolutionized our understanding of APA over the last decade, illustrating both the pervasiveness of dynamic APA events and complexity of the APA regulatory processes. Recently, multiple studies have shed light on the global regulation of APA in response to changes in cell proliferation and cell differentiation in human diseases including cancer (Tian and Manley 2017; Gruber and Zavolan 2019). Both proliferating cells and transformed cells often express a multitude of alternative mRNA isoforms with shortened 3' UTRs through APA (Sandberg et al. 2008), leading to the activation of several proto-oncogenes such as *CCND1*, by escaping miRNA-mediated repression (Mayr and Bartel 2009). On the other hand, 3' UTR lengthening is more prevalent in cell differentiation (Ji et al. 2009; Ji and Tian 2009). For example, progressive 3' UTR lengthening is observed during mouse embryonic development (Ji et al. 2009), and the generation of induced pluripotent stem cells (iPSCs) (dedifferentiation) is accompanied by global 3' UTR shortening (Ji and Tian 2009). Besides regulating cognate transcripts in *cis*, APA-induced 3' UTR changes can also disrupt competing endogenous RNA (ceRNA) regulation in *trans*, thus repressing several crucial tumor suppressors such as *PTEN* in breast cancer (Park et al. 2018). Although these observations imply a possible cell-state- or cell-type-dependent manner of APA regulation, the variability of APA among individual cells and the utility of APA in revealing novel cell subpopulations remain largely unknown.

Single-cell RNA sequencing (scRNA-seq) has become one of the most widely used technologies in biomedical research by providing an unprecedented opportunity to quantify the abundance of diverse transcript isoforms among individual cells (Shapiro et al. 2013; Saliba et al. 2014). However, methods to quantify relative APA usage across single cells remain underdeveloped. Recently, Velten et al. (Velten et al. 2015) developed an experimental protocol BATseq to quantify various 3'-UTR isoforms at the single-cell resolution. By integrating the standard scRNA-seq protocol and the 3' enriched bulk RNA-seq protocol, Velten et al. found that cell types can be well separated based exclusively on their 3'-UTR isoform usage, indicating that APA is a molecular feature intrinsic to cell states (Velten et al. 2015). While a compelling method, BATseq is hampered by its low sensitivity (~5%) and high procedural complexity (Chen et al. 2017), thereby not being widely adopted in practice. In contrast, standard scRNA-seq data is widely available, yet most of the scRNA-seq data has not been analyzed in an "APA-aware" manner. Since scRNA-seq only captures a small fraction (typically 5%-15%) of the total mRNAs in each cell (Stegle et al. 2015), it can falsely quantify genes, especially lowly expressed ones, as unexpressed; this phenomenon is termed as "dropout". Existing bulk RNA-seq based APA methods such as DaPars (Xia et al. 2014) cannot overcome this vexing challenge when applied directly to scRNA-seq data, as they would lead to a high degree of sparsity in the resulting APA profiles. To address this sparsity, recently published computational approaches such as scDAPA (Ye et al. 2020) and scAPA (Shulman and Elkon 2019) extract and combine reads from cells

aggregated based on pre-defined cell types. Alternatively, another study (Kim et al. 2019) aggregates individual genes into "meta-genes" with reference to common functionality. While these strategies cope with the problem of sparsity to some extent, they fail to retain the single-cell or single-gene resolution (Supplemental Table S1).

To fill this knowledge gap, we developed scDaPars (**D**ynamic analysis of **A**lternative **P**oly**A**denylation from **scR**NA-**S**eq), a bioinformatics algorithm for quantifying and recovering APA usage at the single-cell and single-gene resolution using standard scRNA-seq data. Since APA is reported to be regulated in a cell-state- or cell-type-specific manner, scDaPars employs a regression model that enables sharing of APA information across related cells to tackle the sparsity, achieving considerable robustness when applied to noisy scRNA-seq data. In addition, unlike scDAPA and scAPA which are only applicable to 3' end scRNA-seq datasets, scDaPars can be applied to both 3' end and full-length scRNA-seq data. To the best of our knowledge, scDaPars is the first single-cell- and single-gene- level APA quantification method for analyzing standard scRNA-seq data.

Results

Overview of the scDaPars algorithm

Figure 1 presents a schematic illustration of the scDaPars algorithm (see "Methods" for detailed definition and computational procedures). Given a scRNA-seq dataset, scDaPars first calculates raw relative APA usage, measured by the percentage of distal poly(A) site usage index (PDUI), based on the two-Poly(A)-site model introduced in DaPars (Xia et al. 2014). scDaPars takes scRNA-seq genome coverage data as input and forms a linear regression model to jointly infer the exact location of proximal poly(A) sites by minimizing the deviation between the observed read density and the expected read density in all single cells. The relative APA usage is then quantified as the proportion of the estimated abundances of transcripts with distal poly(A) sites (longer 3' UTRs) out of all transcripts (longer and shorter 3' UTRs), and therefore, genes favoring distal poly(A) site usage (long 3′ UTRs) will have PDUI values near 1, whereas genes favoring proximal poly(A) site usage (short 3′ UTRs) will have PDUI values near 0. This step (step (I)) will generate a PDUI matrix with rows representing genes and columns representing single cells. Of note, the raw PDUI values can only be estimated for genes with sufficient read coverages (default coverage of 5 reads per base), which automatically separates genes into robust genes (genes unaffected by dropout events) and dropout genes for further analysis. Due to the intrinsically low coverage of scRNA-seq data (Brennecke et al. 2013), the resulting PDUI matrix from step (I) is overly sparse with widespread missing data. To further recover the complete PDUI matrix independent of gene expression, we develop a new imputation method by sharing APA information across different cells. For a given cell, scDaPars begins by constructing a

nearest neighbor graph based on the sparse PDUI matrix generated in step (I) (Fig.1) to identify a pool of candidate neighboring cells that have similar APA profiles (step (II)). Finally, scDaPars uses a non-negative least square (NNLS) regression model to refine neighboring cells based on robust genes and then borrow APA information in these neighboring cells to impute PDUIs of dropout genes in each cell (step (III)).

Evaluation of the Accuracy and Robustness of scDaPars

To quantitatively evaluate the accuracy of imputed APA usage by scDaPars, we used 384 scRNA-seq libraries of individual human peripheral blood cells (PBMCs) sequenced by Smart-seq2 (Picelli et al. 2013) protocol and a matched bulk RNA-seq library from a benchmark study by Ding et al. (Ding et al. 2020). Since we can estimate poly(A) sites and quantify differential poly(A) sites usage with high sensitivity and specificity in bulk RNA-seq datasets (Xia et al. 2014), we treated the results from the matched bulk sample as pseudo-gold standard for the following evaluation.

First, we showed that scDaPars reliably identified the location of proximal poly(A) sites in single cells. We found that ~84% of poly(A) sites predicted from scRNA-seq data are within 100bp of those predicted in bulk, whereas only ~44% of randomly selected sites from 3' UTR regions are within 100bp of bulk predictions (Fig.2A). We found that ~66.2% of poly(A) sites predicted from scRNA-seq data also overlapped with annotated poly(A) sites complied from RefSeq, Ensembl, UCSC gene models and poly(A)_DB (Wang et al. 2017) within 100bp, and this overlap showed an approximately fivefold enrichment compared with random sites (Fig.2B). In addition, canonical poly(A) signal (PAS) AATAAA was successfully identified by *de novo* motif analysis (Bailey 2011) within the upstream (-100bp) sequence of single-cell predicted poly(A) sites with a p-value ($P = 1.2\times10^{-44}$) similar to that of bulk samples ($P = 5.4\times10^{-48}$) (Fig.2C, Supplemental Fig.S1), supporting the validity of scDaPars's prediction of poly(A) sites.

Next, we showed that scDaPars was able to recover APA usage for genes affected by dropouts in scRNA-seq data. APA is found to be uniquely regulated in distinct immune cell types in PBMCs (Kim et al. 2019). Yet the median Pearson's correlation between APA (PDUI values) of single-cell pairs in the same B cell cluster is only 0.46 when PDUI values were calculated by DaPars (our previous method for bulk RNA-seq) due to dropout effects (Fig. 2D). In contrast, scDaPars successfully recovered PDUI values for most of the affected dropout genes (Supplemental Fig.S2) and increased the median cell-cell correlation by a large margin (0.79) ($P < 2.2\times10^{-16}$) (Fig.2D). We further compared the average APA usage of all single cells with the bulk results. The Pearson's correlation between the average PDUI values of single cells and those of the bulk increased from 0.74 to 0.82 after scDaPars imputation (Fig.2E). Notably, even though the correlation increase was not large, the regression slope increased significantly from 0.59 (DaPars) to 0.8 (scDaPars) ($P = 4.89\times10^{-26}$), indicating

APA usage quantified by scDaPars better represents the linear relationship between the average of single cells and the corresponding bulk.

Finally, we used a simulation study to illustrate scDaPars's ability to identify dynamic APA events (see "Methods") between two cell types. We created a synthetic PDUI matrix of naive and activated CD4 T cells based on bulk RNA-seq data from the DICE project (Schmiedel et al. 2018) (see "Methods"). The naive and activated CD4 T cells are clearly distinguishable using the reference APA profiles estimated from bulk samples (Fig.3A). Additionally, the reference data showed a strong inclination of 3' UTR shortening in activated CD4 T cells ($P = 3.8\times10^{-4}$) (Fig.3D), in line with previous reports that 3' UTR shortening is widely observed upon activation of T cells (Sandberg et al. 2008). However, manually introduced dropout events obscured this differential 3' UTR pattern, in which only ~38% of differential APA genes remained, and the two cell types became less separated by their APA profiles (Fig.3B, E). After we applied the imputation steps of scDaPars, ~79% of differential APA genes are recovered and the clear separation of these two cell types was restored (Fig.3C, F). We further examined the robustness of scDaPars against varying dropout rates. Even though the accuracy of dynamic APA events identified by scDaPars decreased as the dropout rate increased, scDaPars could still achieve > 0.75 area under the receiver operating characteristics (ROC) curve when the proportion of dropout events was as high as 70% (Supplemental Fig.S3).

scDaPars outperforms existing methods by providing single-cell-resolution APA quantification applicable to both 3' end and full-length scRNA-seq data

Several bioinformatics tools have been developed to analyze APA usage using scRNA-seq data (i.e., scDAPA (Ye et al. 2020) and scAPA (Shulman and Elkon 2019)), yet, unlike scDaPars, they were not designed to quantify APA usage at the single-cell resolution. During the preparation of this manuscript, we noticed another method Sierra (Patrick et al. 2020), which detects differential transcript usage in scRNA-seq data, may also be used for quantifying dynamic APA events. To illustrate the superiority of scDaPars over these existing methods, we applied scDaPars, scAPA and Sierra to a benchmark 10x Chromium dataset containing 902 single cells from three lung adenocarcinoma cell lines (Tian et al. 2019) (see "Methods"). scDAPA was excluded from this study since it identifies APA events by pair-wise comparison without quantifying APA usage. scDaPars outperformed both scAPA and Sierra by generating clear and compact cell clusters according to annotated cell lines (UMAP (McInnes et al. 2018) visualization in Supplemental Fig.S4A, B and C). We used silhouette analysis to quantitatively assess the resulting clusters. Compared with scAPA and Sierra, scDaPars showed higher silhouette coefficients which indicated the clustering results from scDaPars are more congruent with the true cell-line labels (Supplemental Fig.S4D, E and F). To further benchmark scDaPars in more complex biological systems, we

applied scDaPars, scAPA and Sierra to an immune dataset containing 3362 PBMCs (Ding et al. 2020) (see “Methods”). Again, the APA usage quantified by scDaPars generated compact and accurate immune cell clusters (Fig.4A, D). In contrast, although Sierra outperformed scAPA and was able to separate B cell and CD14+ monocytes (Fig.4B, C), both Sierra and scAPA failed to accurately distinguish the five immune cell types (Fig.4E, F). Besides generating accurate cell clusters, scDaPars also identified 169 dynamic APA genes (genes with differential poly(A) site usage) among the five immune cell types, most of which (96%) were unseen by existing methods. For example, scDaPars identified EIF1 as a dynamic APA gene between B cells and CD14+ monocytes. Both cluster- and single-cell level coverage plots corroborated that EIF1 exhibits 3’ UTR lengthening in B cells compared to CD14+ monocytes (Supplemental Fig.S5). Yet, EIF1 was not captured by previous methods (i.e., scAPA), indicating the advantage of scDaPars. More importantly, scDAPA, scAPA and Sierra rely on peak calling using 3’ end enriched reads in 10x Chromium to quantify APA usage and thus are not applicable to data generated by full-length sequencing protocols like Smart-seq2 which do not contain enriched peaks in the 3’ UTR regions (Picelli et al. 2013).

scDaPars revealed intrinsic tumor APA variations and immune cell subpopulations in primary breast cancer

Global-scale coordinated APA events are commonly observed in cancers (Xia et al. 2014), and APA induced 3’ UTR shortening was shown to be associated with tumor aggressiveness and poor survival of cancer patients (Lembo et al. 2012; Xia et al. 2014). However, knowledge of APA regulations in cancer has been largely derived from bulk RNA-seq studies. Therefore, while global APA variations between tumor and normal cells have been well characterized, little is known about the intertumoral APA heterogeneity at the single-cell resolution. To illustrate scDaPars’ capacity of characterizing single-cell APA variations in cancers, we applied scDaPars to a Smart-seq2 (Picelli et al. 2013) scRNA-seq dataset containing 563 single cells from 11 breast cancer patients (Chung et al. 2017). In consistent with bulk results, 3’ UTRs were shortened in tumor cells compared to normal cells ($P < 2.2\times10^{-16}$) (Fig.5A). Even PDUI values before scDaPars imputation could separate tumor cells from non-tumor cells with effectiveness comparable to that of gene expression values (Supplemental Fig.S6A), suggesting an important role of dynamic APA events in breast cancer progression. As expected, scDaPars imputed APA profiles showed a better separation between tumor and non-tumor groups (Fig.5B, Supplemental Fig.S7).

To further elucidate APA variations among cell subgroups, we analyzed APA profiles of tumor and non-tumor cells separately. On the one hand, contrary to a previous single-cell APA analysis performed on aggregated “meta-genes” in the same breast cancer dataset (Chung et al. 2017), which showed that no differences in APA were associated with cancer subtypes or patients (Kim et al. 2019), we found that tumor

cells were not only separated into patient-specific clusters based on scDaPars-imputed APA profiles (Fig.5C), but also further classified into different molecular subtypes (Supplemental Fig.S8), showing evidence of both intertumoral and cancer-subtype-specific APA heterogeneity as well as scDaPars's advantage over existing method. On the other hand, non-tumor cells, which were derived from the same group of patients as tumor cells, were clustered mainly according to their cell types (B cells, Myeloid cells and T cells) instead of patients (Fig.5D, Supplemental Fig.S6B). This result not only reaffirmed that dynamic APA events are cell-type specific characteristics of immune cells, but also indicated that the patient-specific APA profiles observed in tumor cells were unlikely due to batch effects in patient samples but rather reflected true intertumoral variations in APA.

In addition, in consistent with prior knowledge of two B cell subclasses (proliferating and naive/memory B cells) in this dataset, we observed that B cells were classified into two cell subgroups based on scDaPars-imputed APA profiles (Fig.5E) with group 2 B cells showed global 3' UTR shortening compared with group 1 B cells ($P = 2\times10^{-3}$) (Fig.5F). We found that most B cell proliferation signature genes from the literature (Chung et al. 2017) were upregulated in group 2 B cells compared to group 1 B cells (Supplemental Fig.S9, Supplemental Table S2), suggesting that group 2 B cells may represent proliferating B cells. Indeed, the proliferating and naive/memory B cells determined by the expression of B cell proliferating marker genes are highly congruent with scDaPars derived cell subgroups (Supplemental Fig.S10A, B). These results are also in line with previous reports that proliferating cells (i.e., group 2 cells) express more isoforms with shortened 3' UTRs through APA (Sandberg et al. 2008). However, expression analysis of all genes failed to identify these B cell subgroups (Supplemental Fig.S10C), revealing the potential benefits of APA analysis in delineating cell subpopulations. In summary, scDaPars improves the characterization of APA variations and cell subpopulations in single cells.

scDaPars enables identification of novel cell subpopulations invisible to conventional gene expression analysis in endoderm differentiation

As APA patterns appear to be globally regulated in cell differentiation (Ji et al. 2009; Tian and Manley 2017) (i.e., decreased proximal poly(A) site usage in more differentiated states of embryonic development), we hypothesized that they could provide a new aspect to identify cell subpopulations during differentiation. To test this hypothesis, we applied scDaPars to a time-course Smart-seq2 (Picelli et al. 2013) scRNA-seq dataset containing 758 cells sequenced at 0, 12, 24, 36, 72 and 96 h of differentiation during human definitive endoderm (DE) emergence (Chu et al. 2016). scDaPars revealed clear and accurate cell clusters from each time point along the differentiation process (Fig.6A). Dimension 2 of the UMAP projection of raw PDUI

values reconstructed single-cell orders matching the true differentiation time points, reflecting the global APA dynamics during cell differentiation (Supplemental Fig.S11).

Next, we investigated whether APA could help delineate novel cell subpopulations invisible to gene expression analysis alone. Imputation based on observed gene expression has been shown to enhance the identification of cell subpopulations (Li and Li 2018). Therefore, to ensure APA is providing additional information beyond expression, we first recovered plausible single-cell gene expression data using scImpute (Li and Li 2018), a state-of-the-art gene expression imputation method. Notably, although the imputed gene expression profile outputs more compact clusters than the raw expression, single cells collected from 72 and 96 h of differentiation were still largely overlapped (Supplemental Fig.S12). To characterize additional cellular heterogeneity, we integrated APA information with imputed gene expression using similarity network fusion (SNF) (Wang et al. 2014). By creating and converging separate similarity networks for APA and gene expression, SNF reduced noisy inter-cluster similarities among cells in 12 and 24 h of differentiation and enhanced intra-cluster similarities observed in one or both similarity networks (Fig.6B). We then quantitatively compared the clustering results by using spectral clustering algorithm (Ng et al. 2002) on different similarity networks with the number of clusters $k = 6$. The clustering results are evaluated by normalized mutual information (NMI) (Witten et al. 2016) where $NMI = 1$ indicates a perfect match between the clustering results and the known differentiation time points. While gene expression imputation increased NMI from 0.76 to 0.85, integration of APA usages with imputed gene expression further increased NMI from 0.85 to 0.89, suggesting the benefits of adding APA information.

Besides unifying the clustering results of APA and gene expression, the fused similarity network also revealed novel and potentially meaningful subpopulations. For example, cells at 96 h of differentiation were divided into two previously unidentified subpopulations (Fig.6B). Through analyzing APA and gene expression between the two subpopulations, we found that APA usage alone can accurately separate the two subpopulations (Fig.6C, Supplemental Fig.S13) and subpopulation 2, which was more distinct from cells in 72 h of differentiation than subpopulation 1, exhibited global 3' UTR lengthening compared to subpopulation 1 ($P = 3.64\times10^{-8}$) (Fig.6D); whereas the imputed gene expression profile alone failed to distinguish the two subpopulations (Fig.6C). The APA profile quantified by DaPars also failed to identify the 2 subgroups (Supplemental Fig.S14), indicating the superiority of scDaPars.

Since subpopulation 2 showed global 3' UTR lengthening, we hypothesized it may represent a more differentiated cell subgroup. To test our hypothesis, we performed differential gene expression analysis between subpopulation 1 and 2 using DESeq2 (Love et al. 2014). As a result, subpopulation 2 was characterized by higher expression of endoderm development marker genes including *GATA6*, *EOMES*, and

SOX17 (Chu et al. 2016) (Fig.6F, Supplemental Table S3). In addition, the transcriptional profile of subpopulation 2 also included significantly upregulated endoderm development related genes like *LHX1*, which is important for renal development (Reidy and Rosenblum 2009), and *HMGA2*, which is required for epithelium differentiation during embryonic lung development (Singh et al. 2014), suggesting subpopulation 2 has a more differentiated phenotype than subpopulation 1. To further elucidate the global biological differences between the two subpopulations, we performed gene oncology (GO) analysis (Luo et al. 2009). We found that several endoderm development related GO terms were highly enriched in the upregulated genes in subpopulation 2 (Fig.6E). Furthermore, using the expression of differential APA genes, we were able to separate the two subpopulations (Supplemental Fig.S15), indicating that some biologically meaningful subpopulations were masked by overall gene expression analysis. Finally, we conducted a trajectory analysis by STREAM (Chen et al. 2019) to independently show the validity of the identified subpopulations. Using cells at 0 h of differentiation as a natural starting point (root), we found that most cells are projected onto the inferred branches according to their corresponding differentiation time points (Supplemental Fig.S16A, B), and the derived pseudotime progression corroborated that cells in subpopulation 2 are more differentiated than those in subpopulation 1 (Fig.6G, Supplemental Fig.S16C). Considered collectively, scDaPars calculated APA usage offered an additional layer of information in characterizing cellular heterogeneity that was otherwise invisible in gene expression analysis.

Discussion

Here, we developed scDaPars, a novel bioinformatics algorithm to *de novo* identify and quantify single-cell dynamic APA events using standard scRNA-seq data. Many methods have been developed to measure the relative APA usages in RNA-seq data from bulk samples (Xia et al. 2014). However, the widespread dropout events in scRNA-seq data impede these bulk-sample based methods to quantify APA usage among single cells (Figs.2D and 2E). To address this technical challenge in scRNA-seq, scDaPars first quantifies raw APA usage based on the two-poly(A)-site model introduced in DaPars (Xia et al. 2014). Since APA exhibits a cell-type specific pattern (Velten et al. 2015; Kim et al. 2019), scDaPars then clusters cells into different cell neighbors based on their calculated raw APA profiles. Next, scDaPars imputes missing APA usage by borrowing APA information of the same gene from neighboring cells. Benchmarking on both real and simulated data show the accuracy of scDaPars in predicting poly(A) sites, the ability in recovering missing APA usages, and the robustness in identifying dynamic APA events across different cell types (Fig.2 and 3).

Previously, methods for analyzing APA usage using scRNA-seq data mostly address the high technical noise in scRNA-seq by creating pseudo-bulk RNA-seq data (i.e. pooled reads from cells that are assigned to the same cell cluster) (Shulman and Elkon 2019; Ye et al. 2020). Unlike scDaPars, even though these methods perform on scRNA-seq data, they do not quantify APA usage at the single-cell resolution but rather measure cell-cluster APA usage, which contradicts the purpose of single-cell sequencing (Supplemental Table S1). Additionally, previous methods are confined by cell cluster assignments determined by conventional gene expression analysis. In contrast, scDaPars quantifies single-cell APA usage independent of gene expression, which provides an additional layer of APA information that helps identify hidden cell states. (Fig.6C).

Finally, unlike existing methods, we expect scDaPars to be widely applicable to any scRNA-seq datasets. While the main analysis presented in this paper builds on scRNA-seq data generated by low-throughput Smart-seq2 (Picelli et al. 2013) protocol and the accuracy of scDaPars decreases as the dropout rate increases (Supplemental Fig.S3), scDaPars can also be applied to datasets generated by high-throughput high-dropout-rate droplet-based methods, e.g. 10x Chromium (Zheng et al. 2017). For example, scDaPars successfully revealed cell-type specific APA patterns in 3362 PBMCs sequenced by 10x Chromium (Ding et al. 2020) (Fig.4A). Together, scDaPars provides an additional layer of APA information that helps identify cell subpopulations invisible to conventional gene expression analysis.

Methods

De novo quantification of dynamic APA events

scDaPars first performs *de novo* identification and quantification of dynamic APA events based on the two-poly(A)-site model introduced in DaPars. The bedGraph files for each single cell were used as input and jointly analyzed to calculate the APA usage measured as the Percentage of Distal poly(A) site Usage Index (PDUI). For each gene, the distal poly(A) site was identified as the end point of the longest 3' UTR among all scRNA-seq samples, and the proximal poly(A) site was inferred by optimizing the following linear regression model:

$$\left(\overline{W_L^{1,2,3,\ldots,m}}, \overline{W_S^{1,2,3,\ldots,m}}, \bar{P}\right) = \underset{W_L^{1,2,3,\ldots,m}, W_S^{1,2,3,\ldots,m} \geq 0,\ 1<P<L}{argmin} \sum_{i=1}^{m} \left\| C_i - \left(W_L^i I_L + W_S^i I_P \right) \right\|_2^2 \quad (1)$$

where W_L^i and W_S^i are the abundances of transcripts with distal and proximal poly(A) sites for cell i, C_i is the read coverage of cell i normalized by total sequencing depth, L is the length of the longest 3' UTR, P is the length of the alternative proximal 3' UTR to be inferred, I_L and I_P are two indicator functions for long and short 3' UTRs such that $I_L = \underset{L}{[1, \cdots, 1]}$ and $I_P = \underset{P,\quad L-P}{[1, \cdots, 1, 0, \cdots, 0]}$. The optimal proximal poly(A) site is selected by minimizing the deviation between the observed read density C_i and the expected read density $W_L^i I_L + W_S^i I_P$ in all single cells. The APA usage is then quantified as PDUI for each gene in each single cell, with PDUI defined as:

$$PDUI_i = \frac{W_L^{i*}}{} \quad (2)$$

where W_L^{i*} and W_S^{i*} are the optimal expression levels of transcripts with distal and proximal poly(A) site for cell i. The smaller the PDUI is, the less distal poly(A) site is used, the shorter the 3' UTRs. The final output is a PDUI matrix in which rows represent genes and columns represent cells. Additionally, PDUIs can only be calculated in this step for genes with sufficient read coverage (default coverage of 5 reads per base), which automatically separate genes into robust genes and dropout genes for future analysis. On average, 50% of the genes in a cell are robust genes after quality control and if the dropout rate in the dataset is higher (e.g., in 10x Chromium datasets), the average number of robust genes in the data will decrease. There are overlaps between robust genes of different cells: in the benchmark dataset in Figure 2, the overlap of robust genes between any two cells is ~40%.

Detection of potential neighboring cells and outliers

Since APA exhibits alterations in different cell types and cell states in a global scale, scDaPars recovers missing single-cell level APA usage by borrowing APA information of

the same gene from neighboring cells. A critical step here is to determine which cells are from the same cell subpopulation and therefore are neighboring cells. Instead of using observed gene expression, scDaPars uses raw APA usage for this task because (1) APA is a feature intrinsic to cell types or cell states; (2) scDaPars quantifies APA usage independent of gene expression. We first performed a quantitative comparison of clustering using raw APA usage and observed gene expression from the hESC dataset in Figure 6 (Supplemental Fig.S17). We found that clustering of raw APA usage outperformed that of observed gene expression (Supplemental Fig.S17C, D) partly because differentiation is one of the biological processes with the most dramatic APA changes. To further illustrate the benefits of quantifying APA independent of gene expression, we modified our original scDaPars algorithm so that the initial clustering is performed using observed gene expression instead of raw APA usage and re-quantified the APA usage of cells from the hESC dataset in Figure 6. We found that the two subpopulations identified by original scDaPars were obscured by the modified version (Supplemental Fig.S18), indicating the advantage of quantifying APA independent of gene expression.

Due to the technical limitation of scRNA-seq data, it is unlikely to completely cluster cells into true subpopulations based on the sparse PDUI matrix generated in last step. Instead, the goal of this step is to determine a set of potential neighboring cells which scDaPars will fine-tune in the following imputation step.

To increase the robustness and reliability of the clustering results and to find more plausible neighboring cells, scDaPars applies principal component analysis (PCA) to the raw PDUI matrix. While the PDUI matrix is sparse, the modularity of dynamic APA provides redundancy in gene dimensions, which can be exploited. Therefore, scDaPars selects principal components (PCs) that can together explain at least 40% of the variance in the data. Note that the neighboring cells are identified in these PCA dimensions while the imputation is performed on the full PDUI matrix.

$$PDUI_{pca} = pca(PDUI, 0.4) \qquad (3)$$

Next, scDaPars identifies and removes outlier cells from the analysis. The outlier cells may be the result of technical errors or may represent true rare biological variations, in either case, scDaPars will not use these outlier cells to impute missing APA usage in other cells. We calculate the distance matrix $D_{N\times N}$ between cells based on the PCA transformed data $PDUI_{pca}$. For each cell m, we define the Euclidean distance of cell m to its nearest neighbor as d_m, resulting a set $\boldsymbol{d} = \{d_1, \cdots, d_N\}$. We denote the first quantile of $\boldsymbol{d}$ as Q_1 and its third quantile as Q_3 and the distance between Q_1 and Q_3 as interquartile range IQR. The outlier cells are defined as cells which are separated by more than 1.5 IQR to the third quantile Q_3.

$$Outlier = \{m : d_m > Q_3 + 1.5 \times IQR\} \qquad (4)$$

The remaining non-outlier cells $\{1, \cdots, N\} \backslash Outlier$ are then clustered into subpopulations using graph-based community detection algorithm. The single cells are the vertices in the graph, and community detection in graphs will identify groups of vertices with high probability of being connected to each other than to members of other groups. We use R package *RANN* with default parameters to first identify the approximate nearest neighbors and convert neighbor relation matrix into an adjacency matrix. We then use *igraph (Csardi and Nepusz 2006)* to represent the resulting adjacency matrix as a graph and apply *walkstrap (Pons and Latapy 2005) algorithm* to identify communities of vertices (cells) that are densely connected. Suppose scDaPars divides cells into K subpopulations in this step, for each cell m, its potential neighboring cells N_m are the other cells in the same cell subpopulation k.

$$N_m = \{i \in k, i \neq m\} \qquad (5)$$

Imputation of missing APA usage

After potential neighboring cells N_m for each cell are determined, we impute APA usage cell by cell. Recall that PDUIs can only be estimated for genes with sufficient read coverage, scDaPars thereby automatically separates genes into robust genes and dropout genes when calculating the PDUI matrix. Here, we denote the set of robust genes for cell m as R_m and the set of dropout genes that will be imputed in this step as D_m. scDaPars then learns the cells' similarities through the robust gene set $G_{Robust,m}$ and impute the APA usage of D_m by borrowing information from the same gene's APA usage in other neighboring cells learned from R_m . To fine-tune the grouping of neighboring cells from N_m, we use non-negative least squares (NNLS) regression:

$$\overline{\theta_m} = argmin_{\theta_m} \left\| PDUI_{R_m,m} - PDUI_{R_m,N_m}\theta_m \right\|_2^2, \ \theta_m > 0 \qquad (6)$$

where N_m represents the indices of cells that are potential neighboring cells of cell m, $PDUI_{Gene_{robust},m}$ is a vector of response variables representing R_m rows in the m-th column (cell m) of the original PDUI matrix, $PDUI_{R_m,N_m}$ is a sub-matrix of the original PDUI matrix with dimensions $|R_m| \times |N_m|$. The goal is to find the optimal coefficients $\overline{\theta_m}$ of length $|N_m|$ that can minimize the deviation between APA usage of R_m in cell m and those in potential neighboring cells. The advantage of using NNLS is that it has the property of leading to a sparse estimate of θ_m, whose components may have exact zeros, so that true neighboring cells of cell m are conveniently selected from N_m. Once $\overline{\theta_m}$ is computed, we have a vector of weighted neighbors associated with each cell in

our data. scDaPars use this coefficient $\overline{\theta_m}$ estimated from the set R_m to impute the APA usage of genes in the set D_m in cell m. All of the above analyses are conducted in *R* (R Core Team 2020).

$$\overline{PDUI_{g,m}} = \begin{cases} PDUI_{g,m}, & \text{if } g \in R_m \\ PDUI_{g,N_m} \cdot \overline{\theta_m}, & \text{if } g \in D_m \end{cases} \quad (7)$$

Differential percentage of distal APA usage index (PDUI) (Dynamic APA events)
We used the following two criteria to define the significant dynamic APA events: first, given the PDUI values for cells in two cell types, the Benjamini-Hochberg corrected Mann-Whitney *U* p-value between two cell types (FDR) is less than 0.05; second, the absolute difference of mean PDUIs in cell type 1 and cell type 2 is greater than 0.2.

$$\begin{cases} FDR \leq 0.05 \\ \left| PDUI_{cell\ type\ 1} - PDUI_{cell\ type\ 2} \right| \geq 0.2 \end{cases} \quad (8)$$

Preprocessing of scRNA-seq data
The scRNA-seq datasets used in this manuscript are all publicly available and are summarized in Supplemental Table S4. The 2 single-cell PBMC data are available at the Gene Expression Omnibus (GEO) under accession code GSE132044. The breast cancer data are available at GEO under accession code GSE75688. The time-course definitive endoderm data are available at GEO under accession code GSE75748. The lung adenocarcinoma cell line data are available at GEO under accession code GSE118767. The DICE immune data used to generate synthetic dataset were obtained from dbGaP under study accession code phs001703.v1.p1. For low-throughput datasets generated by Smart-seq2 (Picelli et al. 2013) protocol, we downloaded the publicly available FASTQ files from GEO database and aligned the reads using STAR 2.5.2 (Dobin et al. 2013) with default parameters, generating one BAM file for each single cell. For high-throughput datasets generated by 10x Chromium (Zheng et al. 2017), we downloaded the FASTQ files and aligned the reads using Cell Ranger 3.0.2. We then selected reads with correct unique molecular identifier (UMI) using Drop-seq tools *FilterBAM (Macosko et al. 2015)* and remove reads with duplicated UMIs using UMI-tools *dedup* (Smith et al. 2017). We next merged reads originated from same cells together and generated one BAM file for each single cell. The BAM files are used as inputs for subsequent scDaPars analysis. The average dropout rate (Percentage of missing data) for Smart-seq2 datasets is ~50% in our study. The 10x Chromium dataset in our study has a dropout rate of ~65%.

Generation of synthetic dataset

The synthetic dataset was created based on bulk RNA-seq data generated from 13 immune cell types (Schmiedel et al. 2018). The different immune cell types are isolated so that each sample only contains cells from one cell type. We used DaPars to estimate the APA usage in these bulk samples and generated an APA matrix, in which rows represent genes and columns represent samples. Since widespread dynamic APA events were reported in Naïve and activated CD4 T cells, we selected only samples that belong to these two cell types for the following simulation.

We down-sampled the resulting bulk APA matrix to emulate the APA profiles generated from single-cell data. We first calculated the dropout rate for each gene in the benchmark immune dataset (Ding et al. 2020). Next, for each gene in the bulk APA matrix, the dropout rate is randomly selected from the set of real dropout rates with replacement. Finally, we used Bernoulli distribution with p equals to the selected dropout rate and n equals to the number of samples to introduce dropouts into the synthetic dataset. The final dropout introduced data has a ~50% dropout rate which is similar to the dropout rate of real datasets. Notice that the generation of the synthetic dataset is independent from the models of scDaPars, so that it can be used to evaluate scDaPars in a fair way.

Benchmark comparison of scDaPars

To illustrate the advantage of scDaPars, we applied scDaPars, scAPA and Sierra to two benchmark 10x Chromium datasets. scAPA measures differential usage of poly(A) sites between different cell types by the proximal poly(A) site usage index (proximal PUI). Since we want to test scAPA's ability for quantifying single-cell-level APA usage, we input single-cell coverage into scAPA to generate a cell by transcript proximal PUIs matrix to perform the clustering analysis. The Sierra pipeline does not yield PDUI like measurements. Instead, it generates a peak count matrix in which peak coordinates are annotated according to the genomic features they fall on including UTRs, exons, or introns. In order to calculate APA usage from the peak count matrix, we first selected peaks falling on the 3' UTRs and only kept transcripts with more than one peak. We then transferred the peak count matrix into an APA matrix by calculating the relative usage of the most distal peak. The resulting APA matrix were used for the clustering analysis. Finally, we performed silhouette analysis by *silhouette ()* in R package *cluster* v2.1.0. to quantitatively evaluate the clustering accuracy of the three methods.

Software Availability

The source codes and the R package scDaPars are available as Supplemental Code. scDaPars is also freely available at GitHub (https://github.com/YiPeng-Gao/scDaPars).

Acknowledgements

We thank Yikai Luo, Dr. Joel Neilson at Baylor College of Medicine, members of the Li lab at University of California, Irvine, and Dr. Jingyi Jessica Li at University of California, Los Angeles for insightful discussions. We thank Dr. Chen Chao at Baylor College of Medicine for his suggestions on this manuscript. This work is supported by US National Institutes of Health (NIH) grants R01HG007538, R01CA193466 and R01CA228140 to W.L. and the Cancer Prevention Research Institute of Texas (CPRIT) grant RR170048 to C.A. C.A. is a CPRIT research scholar.

Author Contributions

W. L. conceived and supervised the project. Y.G. performed the data analysis. Y.G., L.L., W.L. interpreted the data. Y.G., L.L., W.L., C.A. wrote the manuscript.

Disclosure Declaration

The authors declare no competing financial interests.

Figure Legends

Figure 1. A schematic illustration of the scDaPars algorithm.
(I) scDaPars predicts both distal and proximal poly(A) sites by joint analysis of all single-cell samples and quantifies the raw relative APA usage by the proportion of estimated abundances of transcripts with distal poly(A) sites (long isoform). (II) scDaPars determines potential neighboring cells by applying community detection methods in APA profiles generated in step(I). (III) scDaPars uses NNLS regression model to refine neighboring cells and impute missing values by borrowing APA information from neighboring cells.

Figure 2. Evaluation of APA detection accuracy of scDaPars using human PBMCs datasets.
(A) Fraction of poly(A) sites predicted in matched bulk RNA-seq data recovered in single cells using scDaPars or random control. Poly(A) sites predicted in scRNA-seq are considered true if they are located within cutoff distance from the bulk results. The cutoffs range from 0 to 100bp with 10bp increment.

(B) Percentage of scDaPars predicted poly(A) sites or random control overlapped with annotated poly(A) sites from RefSeq, Ensembl, UCSC gene models and poly(A)_DB. The confidence interval was derived by taking random sites 10 times.
(C) The top-scoring signal identified by de novo motif analysis (DREME) from the upstream (-100bp) of scDaPars predicted poly(A) sites from single cells.
(D) Boxplot showing Pearson's correlations between PDUI values of B-cell pairs estimated by DaPars and scDaPars (Wilcoxon test $P < 2.2\times10^{-16}$).
(E) Scatter plots of PDUI values between average of all single cells and bulk results estimated by DaPars (left) and scDaPars (right). Red line represents the theoretical linear relationships between bulk and average of all single-cell PDUIs, and blue represents the actual linear relationships estimated from data.

Figure 3. Evaluation of scDaPars in identifying dynamic APA events between two cell types using naive and activated CD4 T cells.
(A) – (C) Scatterplots showing UMAP results of 54 naive CD4 T cells and 31 activated CD4 T cells based on (A) Reference APA profiles or (B) Dropout events introduced APA profiles or (C) scDaPars corrected APA profiles.
(D) – (F) Heatmaps showing APA profiles of 136 differential APA genes (FDR <= 0.05 and PDUI differences >= 0.2) in the (D) reference data (E) dropout events introduced data and (F) scDaPars corrected data. Rows represent differential APA genes and columns represent cells. 88 out of 136 differential APA genes have shorter 3' UTRs in activated CD4 T cells in the reference data.

Figure 4. scDaPars outperforms existing methods by quantifying APA usage in single-cell resolution.
(A) – (C) Scatterplots showing UMAP results of 3362 PBMCs based on (A) scDaPars quantified APA usage or (B) scAPA quantified APA usage or (C) Sierra quantified APA usage.
(D) – (F) Silhouette plots for clustering results from (D) scDaPars, (E) scAPA and (F) Sierra. The x-axis represents cells and y-axis is the corresponding silhouette coefficient Si for each cell. The silhouette coefficient measures how similar a cell is to its own cluster compared to other clusters, therefore a higher silhouette coefficient indicates a better clustering result and a negative coefficient may suggest the cell is assigned to the wrong cluster. The red dashed line is the average Si for all cells.

Figure 5. scDaPars reveals tumor-specific and immune-cell-type specific APA landscape in primary breast cancer.
(A) Scatter plot of PDUI values in Tumor and Normal cells. For each gene, the mean PDUI values in tumor cells (y-axis) are plotted against that in normal cells (x-axis). Genes with shortened or lengthened 3' UTR (FDR <= 0.05 and PDUI difference >= 0.2)

in tumor cells are shown in red and blue. Bar plot shows the number of shortening genes or lengthening genes in tumor cells and p-value is calculated using single-tailed binomial test.
(B) Scatter plot gives UMAP results calculated from scDaPars restored APA profiles. Each dot represents a cell, and cells are labeled based on cell index provided in the original publication.
(C) Scatter plot of UMAP results of tumor cells. Cells are labeled by patient information.
(D) Scatter plot of UMAP results of immune cells. Cells are labeled by cell type information.
(E) Scatter plot of UMAP results of B cells based on scDaPars results.
(F) Scatter plot of PDUI values in group 1 B cells and group 2 B cells. For each gene, the mean PDUI values in group 2 B cells (y-axis) are plotted against that in group 1 B cells (x-axis). Genes with shortened or lengthened 3' UTR (FDR <= 0.05 and PDUI difference >= 0.2) in group 2 B cells are shown in red and blue. Bar plot shows the number of shortening genes or lengthening genes in group 2 cells.

Figure 6. scDaPars helps identify novel cell subpopulations during human embryonic development.
(A) Scatter plot shows UMAP results of single cells based on scDaPars recovered APA profiles. Cells are labeled based on cell differentiation time points given in the original publication.
(B) Cell-by-cell similarities represented by similarity matrices generated by R package SNFtool.
(C) Scatter plots of UMAP results of cells in 96h of differentiation based on scDaPars results (left) and imputed gene expression (right). Cells are labeled by results from (B).
(D) Scatter plot shows mean PDUI values of genes in subpopulation 2 (x-axis) and sub-population 1 (y-axis). Genes with 3' UTR shortening and lengthening (FDR <= 0.05 and PDUI differences >= 0.2) in subpopulation 2 are labeled in blue and red respectively. Bar plot shows the number of genes with shortening or lengthening in subpopulation 2 and p-value is calculated using single-tailed binomial test.
(E) Selected GO terms enriched in the upregulated genes in subpopulation 2.
(F) Example gene expression levels in two subpopulations.
(G) Stream plot from STREAM which shows cell density along different trajectories at a given pseudotime.

References

An JJ, Gharami K, Liao GY, Woo NH, Lau AG, Vanevski F, Torre ER, Jones KR, Feng Y, Lu B et al. 2008. Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. *Cell* **134**: 175-187.

Bailey TL. 2011. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**: 1653-1659.

Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC. 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods* **10**: 1093.

Chen H, Albergante L, Hsu JY, Lareau CA, Lo Bosco G, Guan J, Zhou S, Gorban AN, Bauer DE, Aryee MJ et al. 2019. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat Commun* **10**: 1903.

Chen W, Jia Q, Song Y, Fu H, Wei G, Ni T. 2017. Alternative Polyadenylation: Methods, Findings, and Impacts. *Genomics Proteomics Bioinformatics* **15**: 287-300.

Chu L-F, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, Choi J, Kendziorski C, Stewart R, Thomson JA. 2016. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology* **17**: 173.

Chung W, Eum HH, Lee H-O, Lee K-M, Lee H-B, Kim K-T, Ryu HS, Kim S, Lee JE, Park YH. 2017. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications* **8**: 15081.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal, complex systems* **1695**: 1-9.

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

Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, Hughes TK, Wadsworth MH, Burks T, Nguyen LT. 2020. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature biotechnology* **38**: 737-746.

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

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

Garneau NL, Wilusz J, Wilusz CJ. 2007. The highways and byways of mRNA decay. *Nat Rev Mol Cell Biol* **8**: 113-126.

Gruber AJ, Zavolan M. 2019. Alternative cleavage and polyadenylation in health and disease. *Nat Rev Genet* **20**: 599-614.

Hoffman Y, Bublik DR, Ugalde AP, Elkon R, Biniashvili T, Agami R, Oren M, Pilpel Y. 2016. 3'UTR shortening potentiates microRNA-based repression of pro-differentiation genes in proliferating human cells. *PLoS genetics* **12**: e1005879.

Ji Z, Lee JY, Pan Z, Jiang B, Tian B. 2009. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci U S A* **106**: 7028-7033.

Ji Z, Tian B. 2009. Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS One* **4**: e8419.

Kim N, Chung W, Eum HH, Lee H-O, Park W-Y. 2019. Alternative polyadenylation of single cells delineates cell types and serves as a prognostic marker in early stage breast cancer. *PloS one* **14**: e0217196.

Lembo A, Di Cunto F, Provero P. 2012. Shortening of 3′ UTRs correlates with poor prognosis in breast and lung cancer. *PloS one* **7**: e31129.

Li WV, Li JJ. 2018. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature communications* **9**: 997.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**: 550.

Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. 2009. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics* **10**: 161.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**: 1202-1214.

Mayr C, Bartel DP. 2009. Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**: 673-684.

McInnes L, Healy J, Melville J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426*.

Ng AY, Jordan MI, Weiss Y. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pp. 849-856.

Park HJ, Ji P, Kim S, Xia Z, Rodriguez B, Li L, Su J, Chen K, Masamha CP, Baillat D et al. 2018. 3' UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk. *Nat Genet* **50**: 783-789.

Patrick R, Humphreys DT, Janbandhu V, Oshlack A, Ho JW, Harvey RP, Lo KK. 2020. Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome biology* **21**: 1-27.

Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* **10**: 1096-1098.

Pons P, Latapy M. 2005. Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, pp. 284-293. Springer.

R Core Team. 2020. R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Reidy KJ, Rosenblum ND. 2009. Cell and molecular biology of kidney development. In *Seminars in nephrology*, Vol 29, pp. 321-337. Elsevier.

Saliba AE, Westermann AJ, Gorski SA, Vogel J. 2014. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* **42**: 8845-8860.

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3'untranslated regions and fewer microRNA target sites. *Science* **320**: 1643-1647.

Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, Ha B, Altay G, Greenbaum JA, McVicker G. 2018. Impact of genetic polymorphisms on human immune cell gene expression. *Cell* **175**: 1701-1715. e1716.

Shapiro E, Biezuner T, Linnarsson S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics* **14**: 618-630.

Shulman ED, Elkon R. 2019. Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic acids research* **47**: 10027-10039.

Singh I, Mehta A, Contreras A, Boettger T, Carraro G, Wheeler M, Cabrera-Fuentes HA, Bellusci S, Seeger W, Braun T. 2014. Hmga2 is required for canonical WNT signaling during lung development. *BMC biology* **12**: 21.

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

Stegle O, Teichmann SA, Marioni JC. 2015. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16**: 133-145.

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

Tian L, Dong X, Freytag S, Lê Cao K-A, Su S, JalalAbadi A, Amann-Zalcenstein D, Weber TS, Seidi A, Jabbari JS. 2019. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature methods* **16**: 479-487.

Velten L, Anders S, Pekowska A, Jarvelin AI, Huber W, Pelechano V, Steinmetz LM. 2015. Single-cell polyadenylation site mapping reveals 3' isoform choice variability. *Mol Syst Biol* **11**: 812.

Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. 2014. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods* **11**: 333.

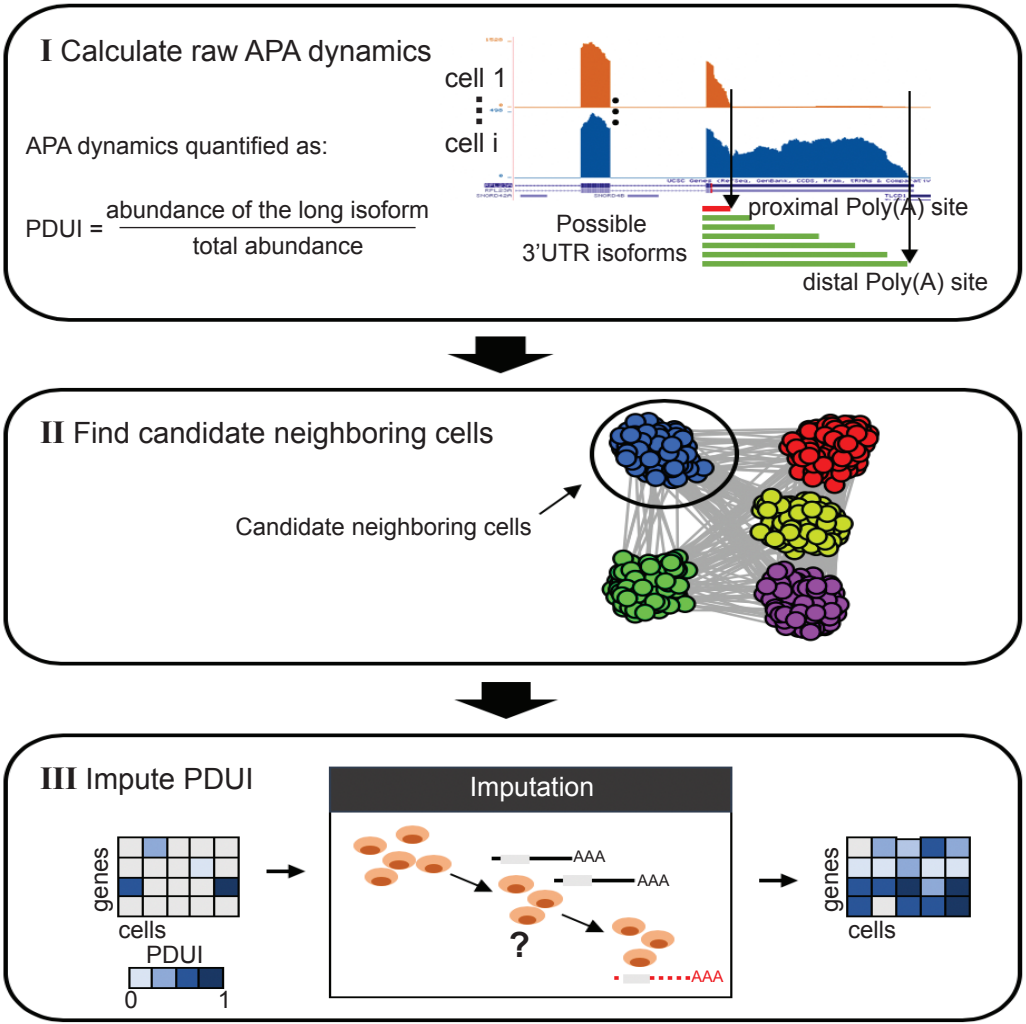
Wang R, Nambiar R, Zheng D, Tian B. 2017. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic acids research* **46**: D315-D319.

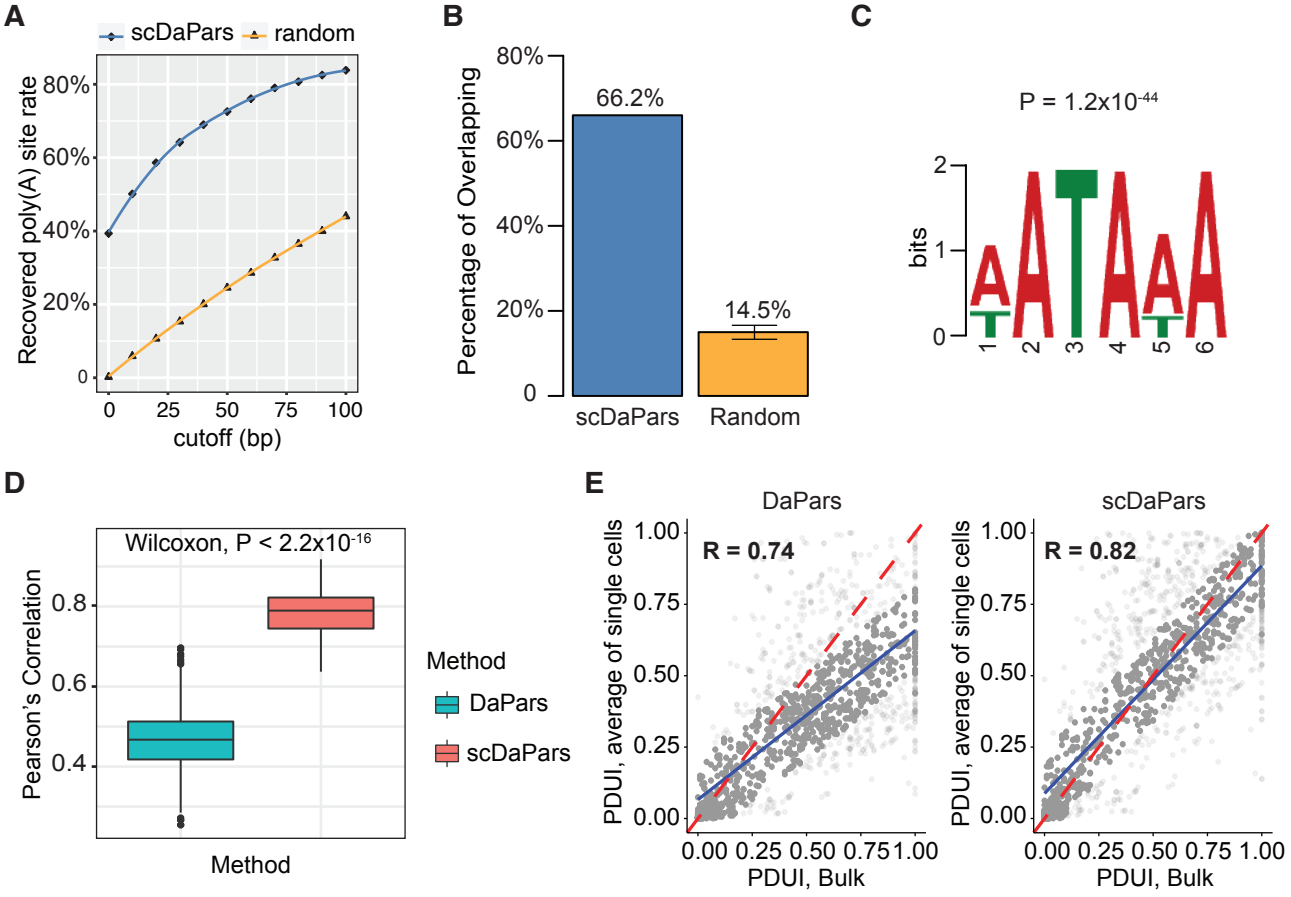
Witten IH, Frank E, Hall MA, Pal CJ. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

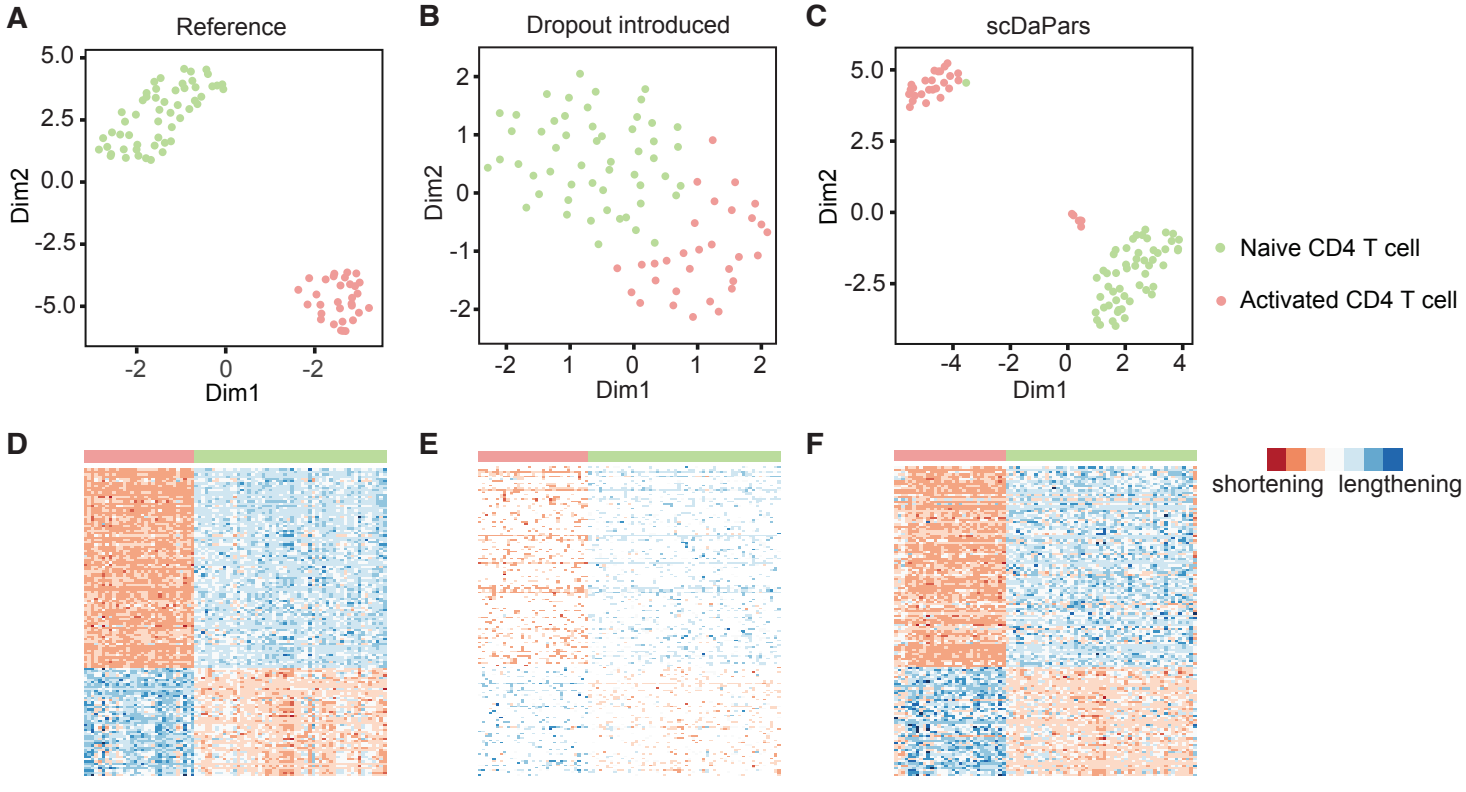
Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, Li W. 2014. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun* **5**: 5274.

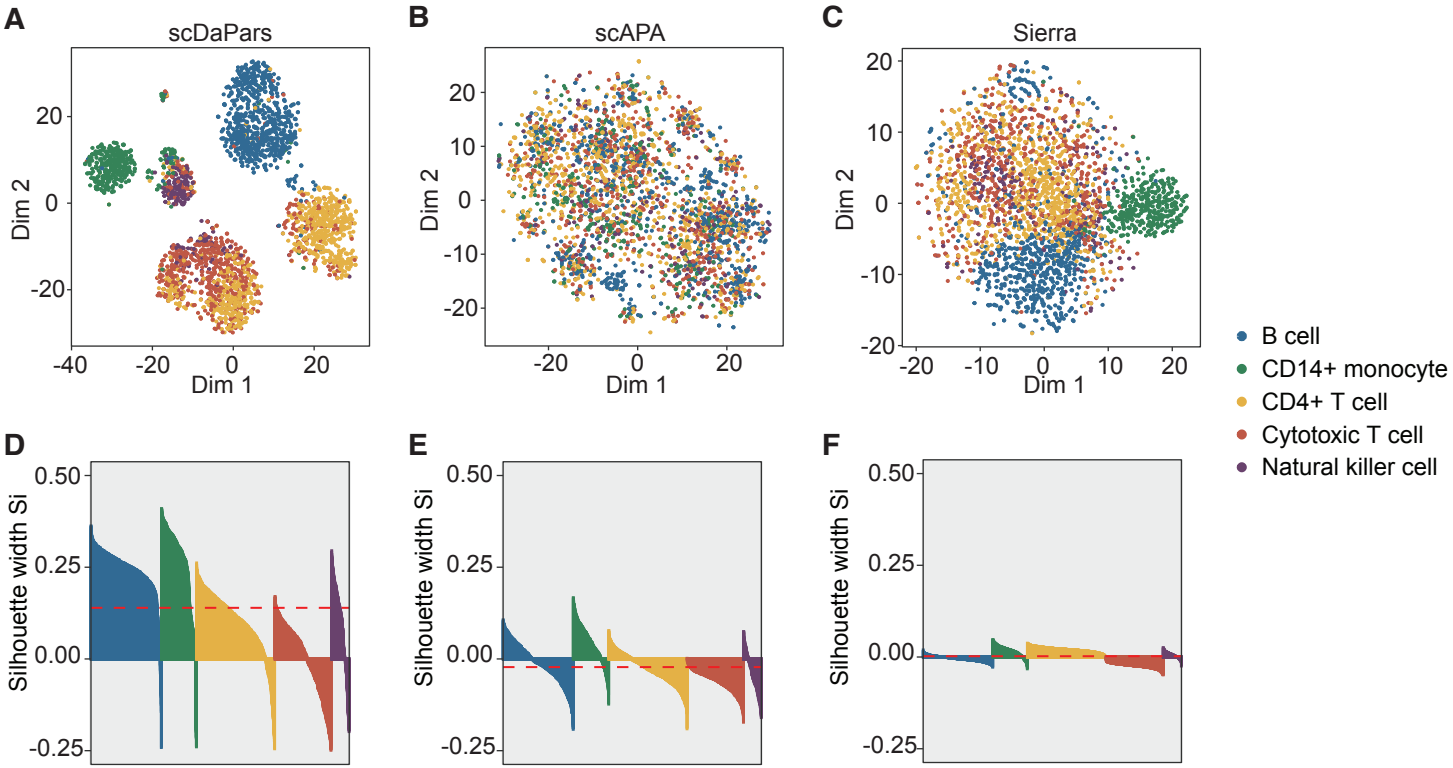
Ye C, Zhou Q, Wu X, Yu C, Ji G, Saban DR, Li QQ. 2020. scDAPA: detection and visualization of dynamic alternative polyadenylation from single cell RNA-seq data. *Bioinformatics* **36**: 1262-1264.

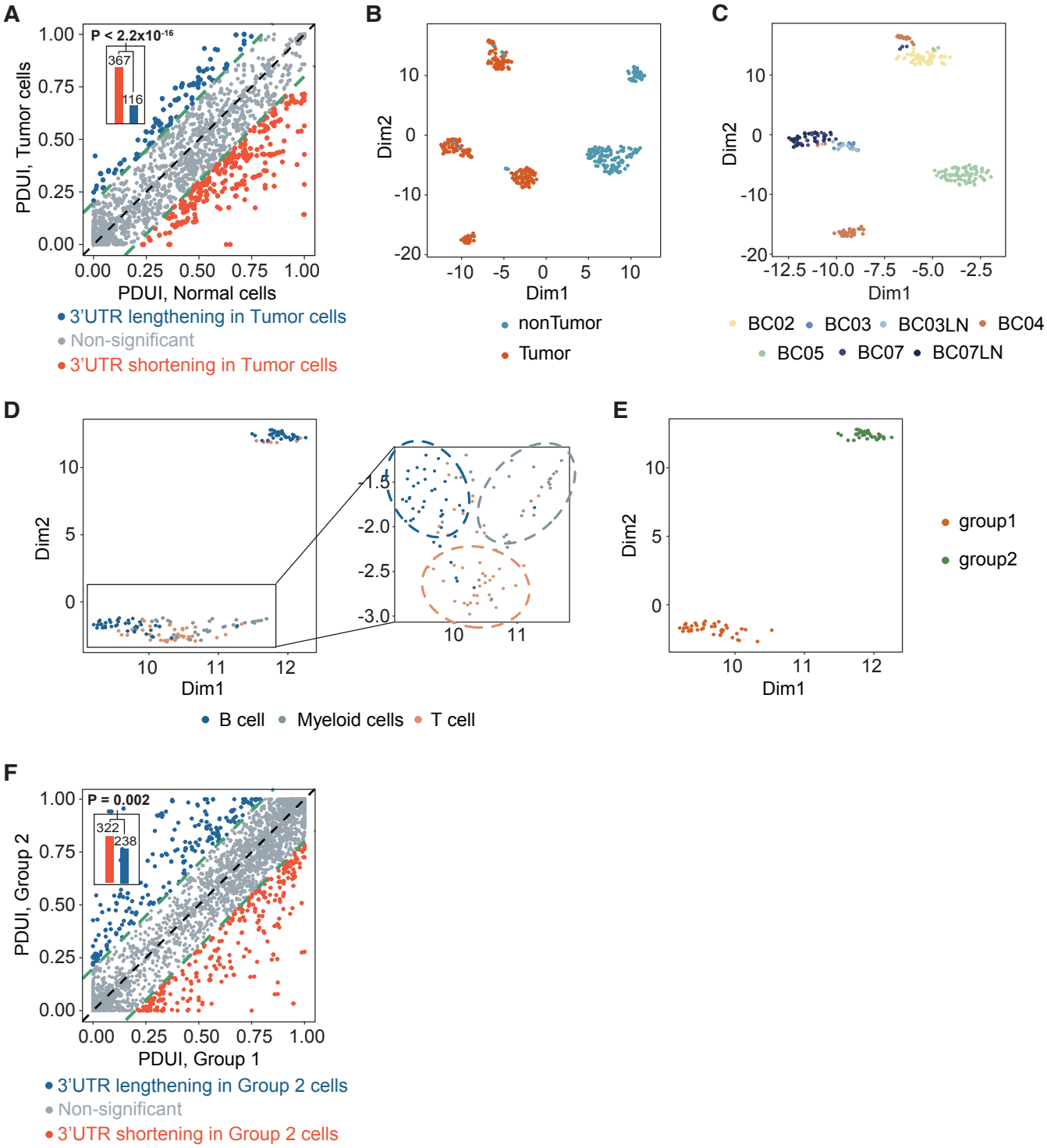
Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J. 2017. Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**: 14049.

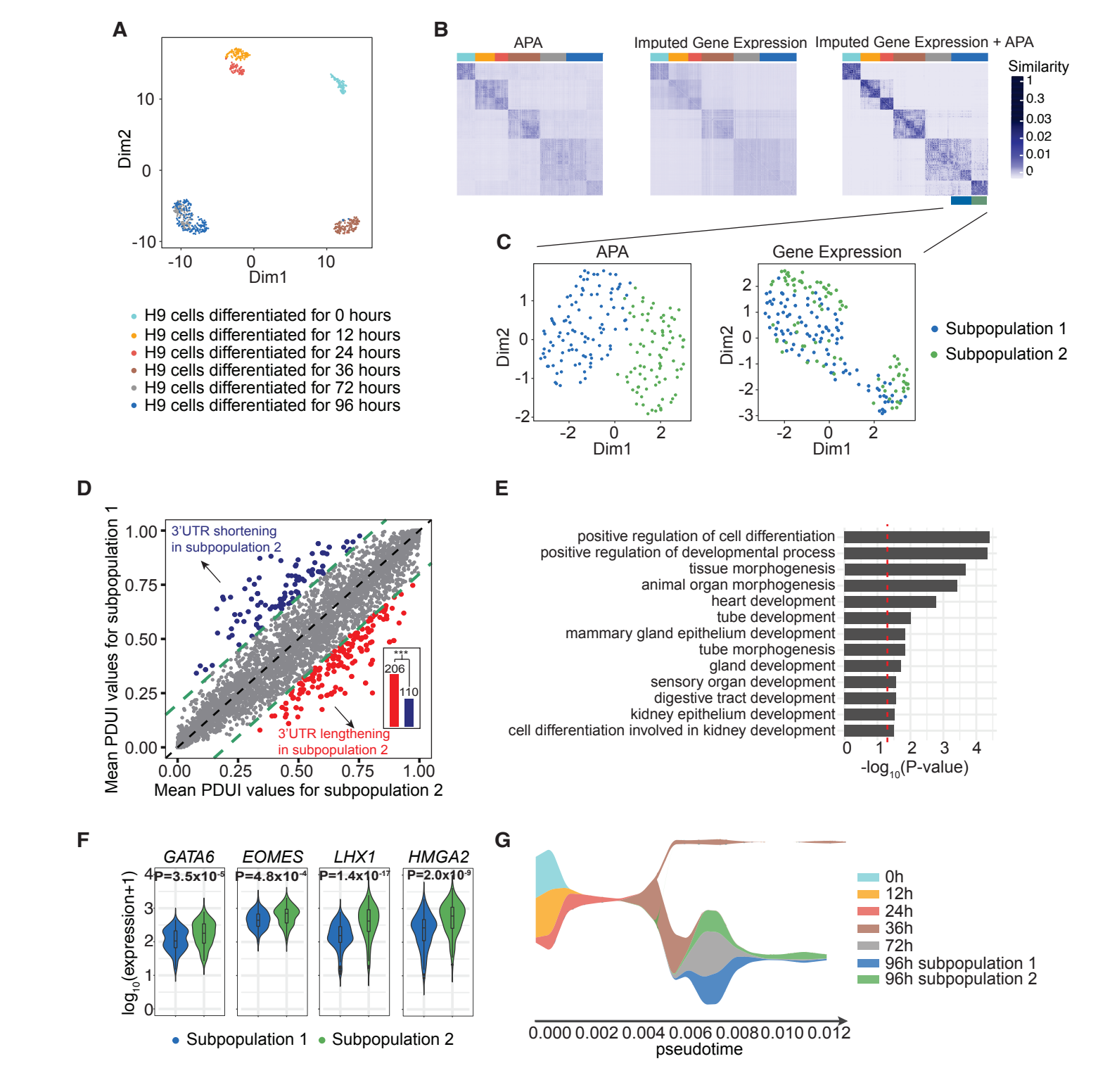

I Calculate raw APA dynamics
APA dynamics quantified as:
PDUI = abundance of the long isoform / total abundance
cell 1
cell i
Possible 3'UTR isoforms
proximal Poly(A) site
distal Poly(A) site
II Find candidate neighboring cells
Candidate neighboring cells
III Impute PDUI
genes
cells
PDUI
0
1
Imputation
AAA
AAA
?
AAA
genes
cells


A
scDaPars
random
Recovered poly(A) site rate
80%
60%
40%
20%
0
0
25
50
75
100
cutoff (bp)
B
Percentage of Overlapping
80%
60%
40%
20%
0
66.2%
14.5%
scDaPars
Random
C
P = 1.2x10-44
bits
2
1
0
1
2
3
4
5
6
D
Wilcoxon, P < 2.2x10-16
Pearson's Correlation
0.8
0.6
0.4
Method
Method
DaPars
scDaPars
E
DaPars
scDaPars
R = 0.74
R = 0.82
PDUI, average of single cells
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
PDUI, Bulk


A
Reference
Dim2
Dim1
B
Dropout introduced
Dim2
Dim1
C
scDaPars
Dim2
Dim1
Naive CD4 T cell
Activated CD4 T cell
D
E
F
shortening
lengthening


A
scDaPars
Dim 2
Dim 1
B
scAPA
Dim 2
Dim 1
C
Sierra
Dim 2
Dim 1
D
Silhouette width Si
E
Silhouette width Si
F
Silhouette width Si
B cell
CD14+ monocyte
CD4+ T cell
Cytotoxic T cell
Natural killer cell


A
P < 2.2x10-16
367
116
PDUI, Tumor cells
PDUI, Normal cells
3'UTR lengthening in Tumor cells
Non-significant
3'UTR shortening in Tumor cells
B
Dim2
Dim1
nonTumor
Tumor
C
Dim2
Dim1
BC02
BC03
BC03LN
BC04
BC05
BC07
BC07LN
D
Dim2
Dim1
B cell
Myeloid cells
T cell
E
Dim2
Dim1
group1
group2
F
P = 0.002
322
238
PDUI, Group 2
PDUI, Group 1
3'UTR lengthening in Group 2 cells
Non-significant
3'UTR shortening in Group 2 cells


A
Dim2
Dim1
H9 cells differentiated for 0 hours
H9 cells differentiated for 12 hours
H9 cells differentiated for 24 hours
H9 cells differentiated for 36 hours
H9 cells differentiated for 72 hours
H9 cells differentiated for 96 hours
B
APA
Imputed Gene Expression
Imputed Gene Expression + APA
Similarity
C
APA
Gene Expression
Subpopulation 1
Subpopulation 2
D
3'UTR shortening in subpopulation 2
3'UTR lengthening in subpopulation 2
Mean PDUI values for subpopulation 1
Mean PDUI values for subpopulation 2
E
positive regulation of cell differentiation
positive regulation of developmental process
tissue morphogenesis
animal organ morphogenesis
heart development
tube development
mammary gland epithelium development
tube morphogenesis
gland development
sensory organ development
digestive tract development
kidney epithelium development
cell differentiation involved in kidney development
-log10(P-value)
F
GATA6
EOMES
LHX1
HMGA2
log10(expression+1)
Subpopulation 1
Subpopulation 2
G
0h
12h
24h
36h
72h
96h subpopulation 1
96h subpopulation 2
pseudotime

Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression

Yipeng Gao, Lei Li, Christopher Ian Amos, et al.

Genome Res. published online May 25, 2021
Access the most recent version at doi:10.1101/gr.271346.120