Published online 14 March 2022

Nucleic Acids Research, 2022, Vol. 50, No. 11 ***e66***
https://doi.org/10.1093/nar/gkac167

SCAPE: a mixture model revealing single-cell polyadenylation diversity and cellular dynamics during cell differentiation and reprogramming

Ran Zhou[1], Xia Xiao[1], Ping He[1], Yuancun Zhao[1], Mengying Xu[1], Xiuran Zheng[1], Ruirui Yang[1], Shasha Chen[1], Lifang Zhou[1], Dan Zhang[1], Qingxin Yang[1], Junwei Song[1], Chao Tang[1], Yiming Zhang[1], Jing-wen Lin[1,*], Lu Cheng[2,3,*] and Lu Chen[1,*]

[1]Key Laboratory of Birth Defects and Related Diseases of Women and Children of MOE, Department of Laboratory Medicine, State Key Laboratory of Biotherapy, West China Second University Hospital, Sichuan University, Chengdu, Sichuan 610041, China, [2]Microbiomes, Microbes and Informatics Group, Organisms and Environment Division, School of Biosciences, Cardiff University, Cardiff CF10 3AX, UK and [3]Department of Computer Science, School of Science, Aalto University, FI-00076, Aalto, Finland

Received November 05, 2021; Revised January 26, 2022; Editorial Decision February 23, 2022; Accepted March 09, 2022

ABSTRACT

Alternative polyadenylation increases transcript diversities at the 3' end, regulating biological processes including cell differentiation, embryonic development and cancer progression. Here, we present a Bayesian method SCAPE, which enables *de novo* identification and quantification of polyadenylation (pA) sites at single-cell level by utilizing insert size information. We demonstrated its accuracy and robustness and identified 31 558 sites from 36 mouse organs, 43.8% (13 807) of which were novel. We illustrated that APA isoforms were associated with miRNAs binding and regulated in tissue-, cell type- and tumor-specific manners where no difference was found at gene expression level, providing an extra layer of information for cell clustering. Furthermore, we found genome-wide dynamic changes of APA usage during erythropoiesis and induced pluripotent stem cell (iPSC) differentiation, suggesting APA contributes to the functional flexibility and diversity of single cells. We expect SCAPE to aid the analyses of cellular dynamics and diversities in health and disease.

INTRODUCTION

When appending poly(A) tails to the 3' end of an mRNA, the untranslated region (UTR) of the corresponding gene may be cleaved at alternative polyadenylation (APA) sites to generate different isoforms. APA occurring in the exonic or intronic region may disrupt the functional or structural domains (1). The 3' UTRs often contain binding sites of micro RNAs (miRNA) and RNA binding proteins, which are involved in the control of mRNA translation (2), stability (3) and localization (4,5). APA is observed in >50% of human genes (6) and plays important role in cellular reprogramming and cell fate. For instance, 3' UTRs are globally lengthened during differentiation (7) but shortened during de-differentiation (8) and tumorigenesis (9).

To better understand how APA is involved in these biological processes, the technologies of mapping transcriptome-wide APA utilized 3' end sequencing protocols were developed at the bulk level. This 3' UTR enriched sequencing is similar to single-cell RNA-sequencing (scRNA-seq) methods based on oligo-dT enrichment, including CEL-seq (10), 10× (11) and Microwell-seq (12). Several methods such as scAPA (13), scAPAtrap (14), Sierra (15), scDaPars (16), SCAPTURE (17) and MAAPER (18) have been proposed to detect APA events by finding peaks from the read coverage profile (Supplementary Table S1). However, since the majority of scRNA-seq reads did not cover the pA sites (Supplementary Figure S1F), these methods had limitations in (i) inferring the locations of pA sites accurately, (ii) separating overlapping peaks and (iii) discriminating weak signals from technical noise.

Here, we developed a Bayesian method SCAPE (Single Cell Alternative Polyadenylation using Expectation-maximization) that aims to solve the aforementioned challenges by utilizing the insert size information introduced during the preparation of sequencing libraries. We demonstrate that SCAPE exhibited superior performance compared to the current methods in various aspects. We further

*To whom correspondence should be addressed. Tel: +86 028 8546 8389; Email: lin.jingwen@scu.edu.cn
Correspondence may also be addressed to Lu Cheng. Email: lu.cheng.ac@gmail.com
Correspondence may also be addressed to Lu Chen. Email: luchen@scu.edu.cn

showcased its usage on multiple datasets and technical platforms, illustrating the global APA landscape in mouse cell atlas and human glioblastoma. We also examined the dynamics of the varied pA length during erythropoiesis and iPSC in the mouse model, in which genes with multiple pA sites (multi-pA gene) exhibited general increasing or decreasing trends in their APA usage during cell differentiation and reprogramming. Furthermore, we identified differentially expressed APA isoforms and found they were associated with expressions of miRNA binding in their 3' UTR. Taken together, our results established APA as a significant contributor to the cell identity at the single-cell level, consistent with an important role of transcript diversification through APA as a means to increase functional diversity.

MATERIALS AND METHODS

SECTION 1: SCAPE

SCAPE bioinformatics pipeline

The input data of SCAPE is a genome-aligned bam file. For 10× or bam-only scRNA-seq datasets, the genome-aligned bam file could be directly used by SCAPE.

For non-10x scRNA-seq datasets like Microwell-seq and Drop-seq, the procedure consists of three steps.

a) The raw fastq files are converted into unmapped bam files by Drop-seq tools (19). Cell barcode, unique molecular identifiers (UMI) and the length of polyT on R1 are recorded into unmapped bam files as tag information at the same time.
b) Then unmapped bam are converted to fastq, which are then aligned it to the reference genome with STAR (20). The unmapped and mapped bam files are merged by MergeBamAlignment function in Picard (2.9.3).
c) The gene expression matrix is retrieved by DigitalExpression in Dropseq tools, which is used to filter low-quality cells. The bam file generated in step (b) is fed into SCAPE to infer the pA sites and other parameters.

Estimation of insert size distribution and poly(A) length distribution

In pair-end scRNA-seq data, a small proportion of reads contains the cleavage site (junction between 3' UTR and poly(A) part), which we call cleavage reads (Supplementary Figure S1F). Both read 1 (R1) and read 2 (R2) of a cleavage read could be uniquely mapped to the genome. We select genes with only one exon and pick cleavage reads mapped to these one-exon genes, which excludes the biases brought by splicing in insert size calculation. Distances between R1 and R2 of these reads, plus the length of poly(A) part in R1, provide the insert size distribution. In practice, we follow several forward computational procedures of Drop-seq (19) to clean up the reads, including trimming and gathering barcode, UMI and poly(T) information from R1. The cleaned pair reads are then aligned to the genome with STAR (20).

Poly(A) part of cleavage reads across all genes could be used to estimate the distribution of poly(A) length. We generated a histogram for the lengths of the poly(A) parts of cleavage reads, which could be used as the empirical distribution of poly(A) lengths. If single-end scRNA-seq data is used, for which there is no information regarding poly(A) parts, we used a uniform distribution from 20 to 150bp.

Statistical model of SCAPE

SCAPE is a probabilistic mixture model for inferring pA sites. Approximate expectation maximization (EM) is used for parameter inference. This mixture model contains K components for reads generated from APA isoforms and one noise component that accounts for random reads.

Figure 1A illustrates the model for an APA component. The pair-end reads (R2 and R1) are mapped to the 3' UTR and poly(A) parts, whose lengths are l_n and r_n, respectively. x_n is the start position of R2. θ_{nk} is kth pA site on the nth DNA fragment, i.e. 3' UTR and poly(A) are connected at this site. α_k and $\beta_k$are hyperparameters that control the distribution of the kth pA site. The poly(A) part length of the nth DNA fragment is denoted by s_n.

There are three sources of uncertainty in the modelling: (i) the cDNA fragment size, (ii) fluctuation of pA sites and (iii) poly(A) part length of the DNA fragment. Based on empirical data, cDNA fragments are around 300 bp with 50 bp standard variation which was estimated by scRNA-seq dataset (Supplementary Figure S1A). According to observations from the limited junction reads (reads cover a pA site), we find the pA sites exhibit a certain degree of fluctuation, as shown in Supplementary Figure S1F. The mRNA poly(A) tails are captured by 30 bp poly(T) tails on the beads, which may bind to any part of the poly(A) and thus introduce uncertainty.

As shown in Figure 1A, we introduce the hidden variables θ_{nk} and s_n, as well as z_{nk} that indicates the component membership. With the help of these hidden variables, we are able to explicitly write out the likelihood. The hidden variables are marginalized out to account for the aforementioned uncertainties. The likelihood is given by

$$p\left(Z, \boldsymbol{x}, \boldsymbol{l}, \boldsymbol{r} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}\right) = \prod_{n=1}^{N} \prod_{k=0}^{K} \{\pi_k p(x_n, l_n, r_n | \alpha_k, \beta_k, M_{I(k)})\}^{z_{nk}}$$

where π_k is the weight for kth component; $I\left(k\right) = 1$ if $k \geq 1$ and $I\left(k\right) = 0$ if $k = 0$; M_1 stands for APA isoform model and M_0 stands for the noise model. The APA isoform model could be further written as

$$\begin{aligned} p(x_n, l_n, r_n | \alpha_k, \beta_k, M_1) &= \sum_{\theta_{nk}} p(x_n, l_n, r_n, \theta_{nk} | \alpha_k, \beta_k) \\ &= \sum_{\theta_{nk}} p(x_n, l_n, r_n | \theta_{nk}) p(\theta_{nk} | \alpha_k, \beta_k) \\ &= \sum_{\theta_{nk}} p(\theta_{nk} | \alpha_k, \beta_k) \sum_{s_n} p(l_n | x_n, \theta_{nk}) p(x_n | s_n, \theta_{nk}) p(r_n | s_n) p(s_n) \end{aligned}$$

where θ_{nk} and s_n are hidden variables and will be marginalized out. The explicit forms of the terms are given by

$$p(\theta_{nk} | \alpha_k, \beta_k) = N(\theta_{nk} | \alpha_k, \beta_k^2)$$

where a Gaussian prior is assumed for the kth pA site.

$$p(l_n | x_n, \theta_{nk}) = \frac{1}{\theta_{nk} - x_n + 1}$$

where we have assumed a uniform distribution for l_n.

$$p\left(x_n | s_n, \theta_{nk}\right) = N\left(x_n | \theta_{nk} + s_n + 1 - \mu_f, \sigma_f^2\right)$$

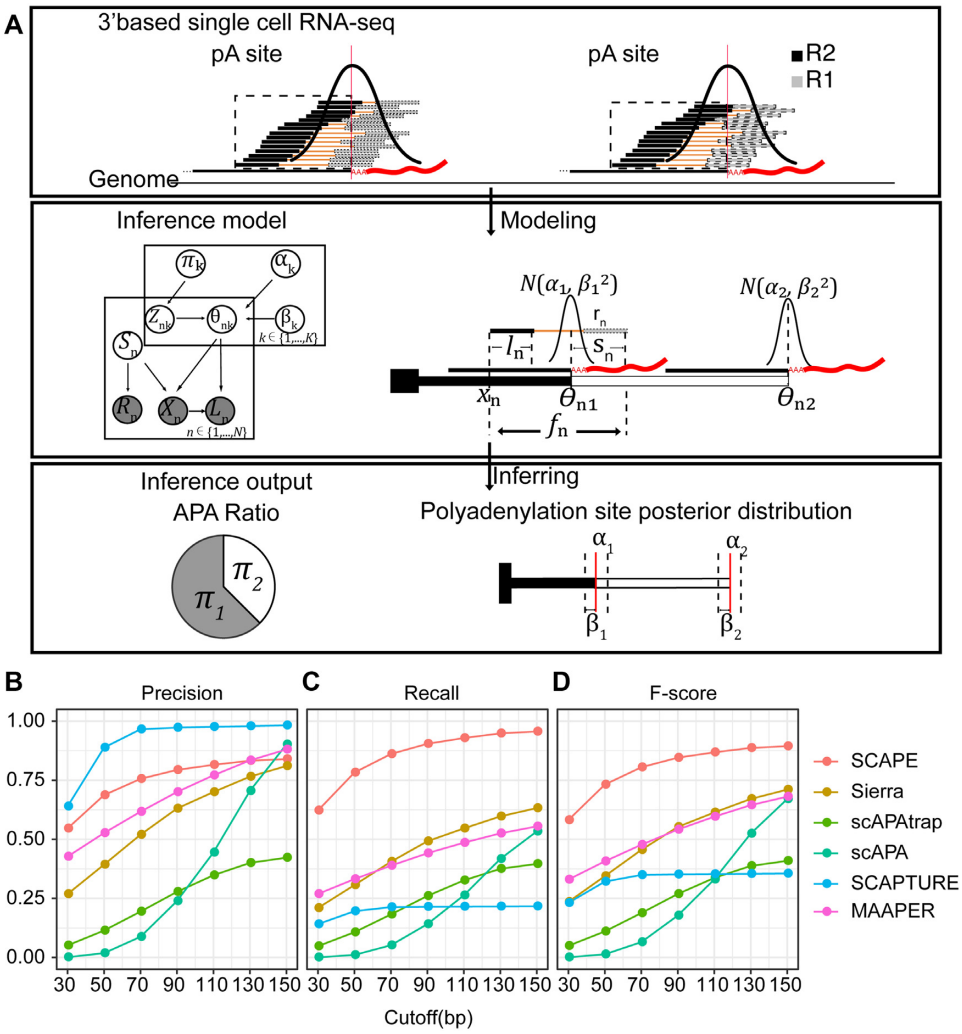

Figure 1. Overview of SCAPE model and method comparison. (**A**) Statistical model of SCAPE. The pair-end reads provide the start positions (x_n) of the DNA fragments on 3' UTR (top). Uncertainties of insert size (f_n) and poly(A) length (s_n) are considered by the statistical model, where l_n and r_n are the length of pair end reads R2 and R1. z_{nk} indicates the membership of nth read to kth isoform (middle). The outputs include the location of each pA site (α_k) and its confidence interval (β_k) and weights (π_k) of isoforms (bottom). The kth potential pA site for nth fragment is denoted by θ_{nk} ($k = 1, 2$), whose Gaussian fluctuation is defined by $\alpha_k$andβ_k . (**B**–**D**) Method comparison for detecting pA sites with SCAPE, Sierra, scAPAtrap, scAPA, SCAPTURE and MAAPER. Precision (B), recall (C) and *F*-score (D) of different methods on a simulation with 12 255 pA sites from 3533 genes with varying cutoff (x-axis) for matched pA sites. Matched pA sites are predicted pA sites that are within x bp (cutoff) of the ground truth. Precision is the proportion of matched pA sites out of all predicted pA sites. Recall is the proportion of matched sites out of all pA sites in the ground truth. F-score is a summarized measure of accuracy calculated from precision and recall with higher value indicating better performance.

where μ_f and σ_f are the mean and standard deviation of fragment length distribution (Gaussian). Note that the fragment length is the sum of the 3' UTR part length $\theta_{nk} - x_n + 1$ and the poly(A) length s_n.

$$p(r_n|s_n) = \frac{1}{s_n}$$

$$p(s_n) = Unif(20, 150)$$

The noise model is given by

$$p(x_n, l_n, r_n, \alpha_0, \beta_0|M_0) = p(x_n, l_n, r_n|M_0)$$
$$= p(x_n|M_0)\, p(l_n|s_n, M_0)\, p(r_n|s_n, M_0) = \frac{1}{L}\frac{1}{L}\frac{1}{LA}$$

where α_0 and β_0 serve as place holders for notational simplicity; L and LA refer to the length of 3' UTR part and maximum poly(A) part of the given gene, respectively.

We use an EM algorithm to maximize the marginalized log-likelihood and Bayesian Information Criterion (BIC) to select the number of components. Note there are no analytic solutions for α_k and β_k in the M-step, we use numeric computation for the maximization instead.

Detailed description of the mixture model is provided in Supplementary Note 1.

Differential APA analysis

Differential APA analysis is essentially the same as differential gene expression analysis except that we replace gene

expression values with pA counts. Here, we present three differential expression approaches.

The first approach is suitable for scRNA-seq dataset where the cells are grouped into multiple clusters. SCAPE is used to assign the reads to different APA isoforms. As a result, we get the number of reads for each pA site, which we term as pA counts. We then treat the pA counts as gene counts and utilize the 'FindMarkers' function in Seurat (v3.1.0) (21) to perform the differential analysis.

The second approach is to use DEXseq (22) to perform differential expression analysis between two groups, which is suitable for scRNA-seq data with high dropout rate. Cells from each group were randomly shuffled and divided into six equal size subgroups, whose expressions are summed to serve as six pseudo-replicates. The expression matrix of the pseudo-replicates and metadata of samples are then fed into DEXseq to perform the differential 3' UTR expression analysis. This approach is implemented as the 'FindDE' in the SCAPE package.

The third approach is suitable for differential analysis between two groups with low dropout rates in the data. We use SCAPE to quantify the weights of pA sites of a gene in each cell. We then test if there exist differences between two groups. For a given pA site, we assume a Gaussian distribution for its weights in different cells in each group. Then we test if the means of the two Gaussian distributions are the same, for which we could calculate the Bayes factor. The function 'ttestBF' from R package 'BayesFactor' (23) is used for this purpose. This approach is implemented as the 'FindBF' in the SCAPE package.

Category of pA usage

To better understand the heterogeneity of APA patterns among cell populations, we first calculated the usage of pA sites for each gene at single cell level, and used previous studies to classify alternative splicing methods to classify the pA usage (24,25), mainly divided into the following five categories: (1) L shape (average of pA usage less than 0.2), (2) J shape (average of pA usage greater than 0.8). When $0.2 \leq$ average pA usage ≤ 0.8, we first calculate the expected variance of each pA using the binomial distribution, and then we compare the observed variance of each pA site with the expected variance of all pA sites, when the observed variance is greater than third quartile of the expected variance is defined as (3) overdispersed, and the first quartile less than the expected variance is defined as (4) underdispersed, and the others are (5) Multimodal.

Expected pA length

Given a gene with $K(K > 1)$pA sites, let us denote the K pA sites by $\theta_1, \theta_2, \cdots, \theta_k$and their corresponding weights by $\pi_1, \pi_2, \ldots, \pi_k$. Note that $\theta_1 > \theta_2 > \cdots > \theta_k$, $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$. The expected pA length is given by

$$\bar{\theta} = \sum_{k=1}^{K} \pi_k \frac{\theta_k - \theta_1}{\theta_K - \theta_1}$$

which normalizes the pA length to the range of $[0, 1]$. If proximal pA sites are used more frequently, the expected pA length will be close to 0; if distal pA sites are preferred, its value will be close to 1. Thus, the expected pA length serves as an indicator of pA site usage. If weights of all pA sites are 0, then the expected pA length is not defined, which should be represented using a missing value such 'NA'.

SIMULATED DATASETS

Theoretical validation on simulated datasets

To assess the performance of SCAPE, we simulate data for a single gene that mimics real scRNA-seq as follows:

1) the number of APA isoforms or pA sites $K \in \{1, 2, 3\}$ is uniformly sampled.
2) the 3' UTR length is set to $L = 2000$ and the poly(A) length is uniformly sampled from $[50, 200]$.
3) K pA sites are randomly selected on the 3' UTR such that adjacent sites are at least 500 bp away.
4) standard deviation for each pA site is randomly picked from {5, 10, 15, 20, 25, 30}.
5) weights for K isoforms and the noise component are generated using a Dirichlet distribution such that isoform weights are 0.2 to 1 and noise weight is around 0.1.
6) fragment length mean and standard deviation are uniformly sampled from $[250, 350]$ and $[20, 40]$, respectively.
7) poly(A) length is sampled from an empirical distribution estimated from a mouse dataset, which falls in the range of $[10, 130]$ with its mode located at 50.
8) mean and std of pair-end reads are set to 120 and 10, respectively.
9) 10% of R1 reads (with poly(A)) are kept to represent reads that can be mapped to the genome, while other unmapped reads are represented using a stretch of 'N'.
10) the number of reads is uniformly sampled from 0 to 5000.
11) 2000 datasets are generated by repeating the previous steps. For each dataset, we performed SCAPE analysis by setting the maximum number of isoforms to 3, the mean and std of fragment length to 300 and 30, respectively. We filter out low weight APA components by varying the cutoff from 0 to 0.2 with a step size of 0.01.

In this simulation, the noise weight is 0.1 and APA components weights are larger than 0.2 in general. Supplementary Figure S1 B-E shows the parameter comparisons between SCAPE results and the ground truth. The x-axis represents the component weight cutoff.

Panel (B) shows the precision and recall of the number of matched pA sites (within 50 bp of a pA site in the ground truth). The overall precision and recall are higher than 0.75, which supports the correctness of the prediction. The best performance is achieved at a cutoff between 0.05 and 0.1, where both precision and recall are around 0.9, which shows a high consistency with the ground truth.

Panel (C) shows the root-mean-square error (RMSE) between inferred pA site location and the ground truth. If no

pA site (ground truth) is found within 200bp of an inferred pA site, the difference is set to 200bp. It can be seen that the average difference drops as the cutoff increases. When the cutoff is 0.1, the average difference between inferred pA sites and the ground truth is around 30bp.

Panel (D) shows the Spearman Rank correlation between inferred pA site std and the ground truth. Due to the high noise-to-signal ratio, the estimated std of pA site shows a low consistency with the ground truth. Thus, we check if the relative order of std of different components is kept instead. Given a component weight cutoff, we first identify matched pA sites (within 50bp of a pA site in the ground truth); then we perform Spearman rank correlation for genes with more than 2 matched pA sites; after that we discard genes that return an invalid correlation (NA); finally, we derive the mean and median of the correlations over all genes. At cutoff 0.1, the mean is 0.5 and the median is 0.9, which shows a relative strong agreement with the ground truth.

Panel (E) shows the RMSE between inferred APA component weights and the ground truth. If no pA site (ground truth) is found within 200 bp of an inferred pA site, the difference is set to the weight of the inferred component. In general, the difference decreases as the cutoff increases. The difference ranges from 0.0215 to 0.0245, which is very low.

In summary, we demonstrate that SCAPE is able to accurately identify and quantify APA isoforms from the theoretical perspective.

Method comparison on simulated datasets

We compare SCAPE with several state-of-the-art methods including Sierra (15), scAPA (13), scAPAtrap (14), SCAPTURE (17) and MAAPER (18) on a simulated dataset, which is generated as follows:

1) 3' UTR annotations of 3533 genes are taken from QAPA (26).
2) for each gene, the number of pA sites $K \in \{1, 2, 3\}$ is uniformly sampled.
3) K pA sites are randomly selected on the UTR such that adjacent sites are 250–350 bp away.
4) standard deviation for each pA site is randomly picked from {5, 10, 15, 20, 25, 30}.
5) poly(A) length is sampled from an empirical distribution estimated from a mouse dataset, which falls in the range of [10, 130] with its mode located at 50.
6) weights for K isoforms and the noise component are uniformly sampled such that summed isoform weights is 0.9 and noise weight is 0.1.
7) mean and standard deviation (std) of fragment length (insert size) are uniformly sampled from [250, 350] and [20, 40], respectively.
8) mean and std of pair-end reads are set to 120 and 10, respectively.
9) 10% of R1 reads (with poly(A)) are kept to represent reads that can be mapped to the genome, while other unmapped reads are represented using a stretch of 'N'.
10) with the above information, we could infer the exact locations of pair-end reads on the 3' UTR and use the corresponding DNA fragments to generate the reads.
11) the number of pair-end reads is uniformly sampled from 0 to 5000.

We analyzed the simulated scRNA-seq dataset using the chosen tools with the following specifications. For SCAPE, mean and std of the fragment length are set to 300 and 50; maximum number of pA sites is set to 5. Default parameters are used in the analysis of Sierra, SCAPTURE, MAAPER, scAPA and scAPAtrap. Since scAPAtrap does not filter out low weight components, we set the component weight cutoff to 0 for SCAPE. Sierra, scAPAtrap and scAPA provide intervals of the peaks of corresponding pA sites, but do not provide the exact locations of pA sites. We use the end point (near poly(A)) of the interval as the location of the predicted pA site for these methods.

Generation of the mouse bone marrow dataset

This section describes sample preparation and data generation of the mouse bone marrow (MBM) dataset.

Mice. C57BL/6J mice were purchased from Beijing HFK Bioscience Co. Ltd (Beijing, China), housed and bred under SPF condition (Specific Pathogen Free) at Laboratory Animal Center of West China Second University Hospital and were allowed access to diet and water ad libitum. All animal experiments were carried out following the protocols approved by the Institutional Animal Care and Use Committee of West China Second University Hospital [(2018) Animal Ethics Approval No.004].

Tissue preparation. The bone marrow cells were obtained from the femur and tibia of C57BL/6 mice aged between 8 and 9 weeks and filtered via a 70-μm cell strainer (BD Biosciences). After centrifuging at 1,200 rpm for 3 min, the cells were re-suspended in PBS containing 2% FBS at a concentration of 1.2×10^7 cells/ml. Hematopoietic stem and progenitor cells (HSPCs) were negatively selected using the EasySep™ Mouse Hematopoietic Progenitor Cell Isolation Kit (StemCell, Cat No. 19856) with a lineage cocktail (biotinylated-CD11b, B220, Gr-1, TER-119 and CD3e), according to the manufacturer's instructions. Sorted cells were pelleted by centrifugation and lysed in TRIZOL reagent (Invitrogen).

Bulk RNA sequencing and analysis. Total RNA was purified using TRIzol reagent (Invitrogen). RNA purity was checked using the NanoPhotometer® spectrophotometer (IMPLEN, CA, USA) and the concentration was measured using Qubit® 2.0 Fluorometer (Life Technologies, CA, USA). The integrity of RNA was assessed using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). Sequencing libraries were generated using NEB Next® UltraTM RNA Library Prep Kit for Illumina® (NEB, USA) following the manufacturer's recommendations. The library fragments were purified with AMPure XP system (Beckman Counlter, Beverly, USA). The libraries were loaded on the Illumina HiSeq 2500 platform for 150 bp pair-end sequencing at the Annoroad Gene Technology Corporation (Beijing, China). Fastq files were initially subjected to a quality control step using FastQC (v0.10.1), and the reads were then trimmed using Trimmomatic (27). The filtered reads were mapped

to mouse (mm10) using STAR (v2.5.3) (20) with Ensembl GRCm38 (release-84) annotation.

Oxford Nanopore Technologies (ONT) library preparation and sequencing. Total RNA from bone marrow (cKit+) samples of 8-weeks-old C57BL/6J mice was extracted using TRIzol reagent (Invitrogen) following the manufacturer's protocol. Starting from 1500 ng total RNA, mRNA selection was performed using Dynabeads® Oligo (dT)25 (Invitrogen). The mRNA was converted to cDNA and 10 amplification cycles of PCR using PCR-cDNA Sequencing Kit (Oxford Nanopore Technology, SQK-PCS109). The cDNA PCR products were end-repaired and added dA-tailed using the NEBNext® Ultra II End Repair/dA-Tailing Module (NEB, E7546) and then ligated to Naïve barcode by using Blunt/TA Ligase Master (NEB, M0367) and tethered to the adapter by using Quick Ligation Module (NEB, E6056). The library was loaded into PromethION flowcells and sequenced over 72 h. Finally, used the 'high accuracy' basecalling mode of guppy_basecaller to convert the electrical signal of fast5 to the base sequence of fastq.

PacBio library preparation and sequencing. The full-length cDNA libraries were generated as described above, 5 μg of full-length cDNA were used for size selection using the BluePippin™ Size Selection System (Sage Science, Beverly, MA, USA). SMRTbell library was constructed using 1 μg size-selected (above 4 kb) cDNA with the Pacific Biosciences SMRTbell template prep kit. The binding of SMRTbell templates to polymerases was conducted using the Sequel II Binding Kit, and then primer annealing was performed. Sequencing was carried out on the Pacific Bioscience (PacBio) Sequel II platform. Sequencing library construction and sequencing were performed in Annoroad Gene Technology (Beijing, China).

Single-cell RNA Library preparation and sequencing. HSPCs were collected as described before. The collected cells were resuspended in PBS containing 1% BSA at the concentration at 1×10^6 cells/mL and prepared for single-cell library preparation. Single cells were prepared in the Chromium Single Cell Gene Expression Solution using the Chromium Single Cell 3′ Gel Bead, Chip and Library Kits v2 (10× Genomics) as per the manufacturer's protocol. The cells were then partitioned into Gel Beads in Emulsion in the Chromium instrument, where cell lysis and barcoded reverse transcription of RNA occurred, followed by amplification, shearing 5′ adaptor, and sample index attachment. Libraries were sequenced on the Illumina NovaSeq 6000 platform at Novogene, Beijing, China.

Analyses of full-length sequencing from PacBio and ONT. For PacBio dataset, the 142 686 raw reads were clustered and polished using the Iso-Seq pipeline (SMRTLink, v5.1.0.26412) (28) and 20,106 high-quality, full-length sequences were used for the follow-up analysis. The raw reads were aligned to Ensembl GRCm38 (release-84) using GMAP (version 2018-07-04) (29). For ONT dataset, raw reads were initially subjected to a quality control step using NanoComp (v1.33.1) (30). Then 32 117 065 high-quality reads of long-read sequencing were aligned to the same genome using Minimap2 (v2.17-r974-dirty) (31). We further used FLAIR (v1.5) (32) to construct full-length isoforms with default parameters for PacBio and ONT reads.

Validation of APA events on tissue sections

Fresh Frozen Sections. Brain tissues and femurs were isolated after perfusion using DEPC-PBST. Femurs were immersed in 4% paraformaldehyde (PFA) in 1 × PBS at 4°C for 24 h, transferred into 14% (w/v) EDTA (pH 8.0) at 4°C for 48 h, next into DEPC-treated water containing 30% (w/v) sucrose at 4°C for 24 h for dehydration and tissue clearing, followed by embedding and cryostat. Brains were directly embedded in O.C.T. and snapped frozen in the gas phase of liquid nitrogen following storage at -80 °C until use. All the fresh frozen sections were sectioned on a Leica CM1950 cryostat at a thickness of 10 μm. For brain sections, a fixation at room temperature for 45 min using 4% PFA in 1× PBS was performed following dehydration in a gradient series of ethanol (75%, 85% and 100% respectively, v/v). For femur sections, O.C.T. were washed away before dehydration, skipping the fixation step. All the sections were kept at -80 °C until use.

in situ RNA detection. RNA fluorescence *in situ* hybridization (RNA-FISH) employing padlock probes was performed according to the methods described previously (33–35) with some modifications. In brief, padlock probes were designed using in-house python scripts. The 5'- and 3'-arms were reverse complement to the mRNA, leaving a nick between 5'- and 3'-sites. All priming sites of the padlock probes were assessed by ultrafast alignment using Bowtie2, and the probes that bind specifically to the targets were shortlisted.

Sections were equilibrated to room temperature before permeabilization in 0.1 M HCl and 0.2 mg/mL pepsin at 37°C for 5 min. After washing twice with DEPC-PBST, a heat shock was performed in the TE buffer at 70°C for 10 min, according to the method described by Codeluppi *et al*. (36). Sections were then hybridized with a pool of padlock probes (100 nM) and RCA primers in 2× SSC and 20% formamide at 37°C for 4 h, briefly washed with DEPC-PBST, and incubated with SplintR Ligase mixture (250 nM SplintR Ligase with 1× BSA and 1 U/μl RiboLock RNase Inhibitor) at 37 °C for 1 h. The sections were then washed twice with DEPC-PBST, post-fixed with a 4% PFA in 1 × PBS at room temperature for 30 min. After washing twice with DEPC-PBST, the rolling-circle amplification (RCA) was performed using phi29 DNA polymerase mixture (1 U/μl phi29 DNA polymerase and 1 mM dNTP mixture in 1× phi29 buffer) at 30°C for 2 h. With a brief wash with DEPC-PBST, the fluorescence labeled detection probes were added at the condition of 20% formamide, 2× SSC at room temperature for 30 min. The sections were mounted in SlowFade with DAPI and covered with a cover slit. Images were acquired using a Leica DM6B widefield fluorescence microscope under 40× objective lens and DAPI, GFP, Cy3, TXR and Cy5 filter set. Area scan and images stitching were performed using LAS X suite. After exporting tiled images as TIFF format, bright compact FISH signals were detected using CellProfiler (3.1.8) (37) with a custom pipeline. In brief, signal dots were recognized using the 'IdentifyPrimaryObjects' module with a modified threshold, keeping the coordinates of each dot. Due to the FISH signals are very compact in a tiled large image, these

coordinates were virtually plotted under DAPI channel for better pattern observation.

Analysis of scRNA-seq data in mouse bone marrow (MBM) dataset

Alignment and quantification. The sequencing data were processed using Cell Ranger software (version 3.0.0) with default parameters, and mapped to the mouse (mm10) genome. We first removed outliers using isOuterlier from scater package (v1.12.2) (38), then removed low-quality cells (gene count < 500 or the mitochondrial gene ratio > 25%).

Clustering and annotation. We used Seurat (v3.1.0) (21) for downstream analyses including data normalization (NormalizeData, LogNormalize method, scaling factor 10,000), data feature scaling (ScaleData), variable gene detection (FindVariableGenes with vst method) and PCA of variable genes (RunPCA). Then the original Louvain algorithm (FindClusters) with clustering resolution 0.6 was performed to cluster the cells.

Weighted nearest neighbor (WNN) analysis. Gene and pA expression matrixes were integrated by weighted nearest-neighbor (WNN) model using the 'FindMultiModalNeighbors' function from R package Seurat (v. 4.0.5) (39).

Trajectory analysis. The STARsolo (40) was used to estimate proportions of spliced and unspliced reads in mouse bone marrow dataset. The spliced and unspliced count matrices were inputted to scVelo (41) to infer the pseudotime and driver genes in 'dynamic' mode with default parameters.

GO analysis. Based on a given classification of cells, differentially expressed pA sites were obtained using the FindAllMarkers function in the Seurat package with default parameters. These pA sites were further filtered using criteria P-value <0.01 and $\log_2$ fold change >0.25 to generate significant pA sites, the host genes of which were fed into clusterProfiler (42) to perform GO enrichment analysis.

Analysis of mouse cell atlas (MCA), mouse iPSC (MIC) and human Glioblastoma (HGB) datasets

We download the mouse cell atlas, mouse iPSC and human GBM (Glioblastoma) datasets from GSE108097 (12), GSE103221 (43) and PRJNA5795936 (44), respectively Following the same preprocessing steps described by the original publications, we retrieved high quality cells for downstream analysis in Seurat(v.3.1.0) (21), including variable gene detection (FindVariableGenes with vst method), PCA using variable genes (RunPCA) and UMAP using PCA matrix (RunUMAP).

Mouse cell atlas and iPSC datasets. Differentially expressed pA sites for each tissue were obtained using 'FindAllMarkers' in Seurat by inputting the pA count matrix. Top pA sites (P-value < 0.01 and $\log_2$ fold change > 0.25) were then selected for GO enrichment analysis using DAVID (45).

Human GBM datasets. The presence/absence of somatic copy-number aberrations (CNAs) was assessed with CONICSmat (46). Malignant and non-malignant cells were classified based on chromosome 1–22 with the default parameters as described in section 'CNA Analysis of Tumor versus Normal Cells' (44). Briefly, raw counts in cells of each patient were scaled to log(CPM/100 + 1) and centered by the average expression in each cell. Subsequently, we fit a two-component Gaussian mixture model on the average expression values across all cells for each chromosome, whereas only genes robustly (log(CPM/100 + 1)) > 1) expressed in more than 10 cells were considered. We then only focused on chromosomes with a significant deviation of the log-likelihood of the model compared to a one-component model (likelihood ratio test < 0.001) and a difference in Bayesian Inference Criterion (BIC) > 300. These chromosomes were considered to have somatic CNVs. Cell with CNV alterations with a cutoff on the posterior probability (pp > 0.8) were classified as tumor cells, whereas cells were classified as normal cells. Cells that could not be clearly assigned to a genotype (e.g. 0.2 < pp < 0.8) remained unclassified.

3'-Seq dataset analyses

The mouse 3'-seq dataset of *ex vivo* isolated mouse multipotent steady state hematopoietic stem cells (sHSC) and proliferating HSCs (16h pIC; pHSC) was downloaded from PRJEB29693 (47). Following the preprocessing steps described in the original publication, fastq files first went through a quality control step using FastQC (v0.10.1), then the reads were trimmed using Trimmomatic (27). The filtered reads were mapped to mouse genome (mm10) using bowtie2 (2.4.1) (48) with Ensembl GRCm38 (release-84) annotation.

Differential miRNA expression analysis

For mouse brain and bone marrow, the miRNA expression matrix was downloaded from miRBase (http://www.mirbase.org/). For human GBM and normal brain, miRNA expression matrices of all GBM patients were fetched by bioconductor package TCGAbiolinks (49) from TCGA with `GDCquery(project = 'TCGA-GBM',data.category = 'Transcriptome Profiling', data.type = 'miRNA Expression Quantification')`. Datasets of all normal human brain tissues were downloaded from R package microRNAome (50) with `data('microRNAome')`, samples annotated as astrocyte, cerebellum and brain were included for downstream analysis. For each dataset, we performed normalization and differential expression analysis using DESeq2 (51). Significance of differentially expressed miRNAs was set to adjusted P-value <0.05 and absolute $\log_2$(fold-change) > 1.

RESULTS

Overview of SCAPE

In scRNA-seq, each pair-read contains read 1 (R1) and read 2 (R2), the former consists of a cell barcode, a unique molecular identifier (UMI), and a poly(A) tail (Figure 1A), whereas the latter provides the sequence information on 3' UTR that is not directly adjacent to the pA site. Therefore, scRNA-seq data cannot directly pinpoint the APA location.

However, a size-selection step is used in the sequencing library preparation step to control the average length of inserted cDNA fragments, which provide key information to infer pA sites. By utilizing this insert size information, we developed a Bayesian model **SCAPE** that enables *de novo* identification of pA sites and quantification of APA events using scRNA-seq data (Figure 1A, Online Methods, Supplementary Note 1).

SCAPE models the scRNA-seq data as a mixture of K isoform components and one noise component, i.e. a read either arises from an isoform or noise. The ith ($i = 1 \ldots K$) isoform component (Figure 1A) has three parameters: (a) the mean position of the pA site α_i, (b) the standard deviation (std) of the pA site β_i, (c) the weight or proportion π_i. The idea is to infer the kth pA site (θ_{nk}) on nth fragment using the start position of R2 on 3' UTR (x_n), fragment length (f_n) and poly(A) length (s_n). As a fragment only consists of the 3' UTR and the poly(A), the length of 3' UTR is given by $f_n - s_{\mathrm{n}}$. Since the end position of the 3' UTR is the pA site, we could infer it by adding the length of 3' UTR ($f_n - s_{\mathrm{n}}$) to its start position (x_n), i.e. $\theta_{\mathrm{nk}} = x_n + (f_n - s_{\mathrm{n}}) - 1$. In practice, the start position (x_n) is known, while the fragment length (f_n) and poly(A) length (s_n) possess uncertainties that need to be modelled. In the pair-end sequencing, the insert size (f_n) distribution can be estimated from the read pairs by SCAPE that are mapped to large constitutive regions such as 3' UTR, which are typically intronless. Poly(A) length (s_n) distribution could be estimated from the reads that cover pA sites, or data from the literature (52,53) that directly measure the poly(A) length. With these specifications, SCAPE integrates out the uncertainties and infers the probabilities of a read arising from isoforms or noises. The number of components is automatically selected using Bayesian Information Criterion (BIC). We used the Expectation-Maximization (EM) algorithm and numerical optimization to infer the parameters (Online Methods and Supplementary Note 1).

SCAPE can be used to analyze single- or pair-end scRNA-seq data such as 10× genomics or Microwell. Users need to further specify the maximum number of isoforms in the data, the minimum weight of an isoform component, the mean and standard deviation (std) of fragment lengths. The outputs are K sets of parameters that specify the location (mean and std), weights of pA sites and the noise component, as well as the component membership of each input read (Figure 1A). With these parameters, SCAPE can calculate statistics of interest such as expected pA length and perform differential analysis (Online methods).

Robust identification of APA in simulated and real datasets

To validate the robustness of SCAPE, we simulated 2000 datasets according to the hypothesized statistical model. We demonstrated that the estimated parameters agreed well with the ground truth (Supplementary Figure S1B-E), in general estimated pA sites were within 30 bp of the ground truth and estimated weights varied less than 0.03), which validated the accuracy of our inference algorithm. Next, we compared SCAPE with several methods including Sierra (15), scAPA (13), scAPAtrap (14), SCAPTURE (17) and MAAPER (18) on a simulated scRNA-seq dataset that includes 3533 genes and 12 255 pA sites (Online Methods). SCAPE exhibited a higher precision, recall and F-score than other methods (Figure 1A-D) in terms of pA site identification. Moreover, the number of matched real pA sites of SCAPE ($N = 9592$) was much larger than MAAPER ($N = 4067$), Sierra ($N = 3759$), SCAPTURE ($N = 2404$), scAPAtrap ($N = 1313$) and scAPA ($N = 134$). In terms of isoform weight quantification, SCAPE showed the highest correlation with the ground truth ($R^2 = 0.97$), while the best performance of other methods was $R^2 = 0.53$ (Supplementary Figure S1G).

To further assess the performance of SCAPE in real datasets, we generated a comprehensive RNA-seq dataset from the same mouse bone marrow (MBM) sample to cross validate the predictions, including 10x scRNA-seq, bulk RNA-seq, ONT and PacBio data (Online Methods, Supplementary Table S2). Moreover, we used a mouse cell atlas (MCA) scRNA-seq data (Microwell-seq) (12) and bulk 3' tag sequencing data (3'-seq) from mouse bone marrow (47) as external validations. From 10x scRNA-seq data, SCAPE detected 15 576 pA sites, 78.1% of which have been annotated in GENCODE v13 (54) or polyA_DB (55) (Figure 2A, Supplementary Table S3). To further validate these predicted pA sites, we first searched for the poly(A) signals within 50 bp upstream region and compared the enrichment with the random 3'UTR sequences after 1000 permutations. We found that the majority of identified pA sites contained canonical poly(A) signals (49% and 12.2% for AATAAA and ATTAAA, respectively), or their one-base edit (20.6%) (Figure 2B), which are significantly higher than those of the random 3'UTR sequences (0.2% for AATAAA and 0.012% for ATTAAA, T test $P \leq 2.2e^{-16}$, Supplementary Figure S2A). These results showed that a large proportion of pA sites identified by SCAPE have annotations and poly(A) signals. The rest 3,411 (21.9%) predicted pA sites were not annotated. Notably, 93.9% of unannotated ones can be validated by ONT or PacBio (Supplementary Figure S2B), this validation rate is significantly higher than that for randomly-selected sites from 10 000 3' UTR regions (21.2%, Chi-square test, $P < 2.2e^{-16}$, Supplementary Figure S2C).For the rest of 6.1% unvalidated ones, we found that they were from genes expressed in fewer cells (Wilcox test $P = 0.017$, Supplementary Figure S2D), potentially explaining why they were not captured by the bulk long-read sequencing. Next, we used the high-confident pA sites identified by ONT sequencing as the 'ground truth' in mouse bone marrow, and compared the pA sites identified by SCAPE, scAPA, Sierra, scAPAtrap, SCAPTURE and MAAPER. 83.7% of pA sites identified by also were supported in one or more methods (Supplementary Figure S2E), suggesting a high consistency between SCAPE and other methods. Simultaneously, SCAPE exhibited higher precisions than other five methods, with comparable recalls to that of MAAPER and are higher than the other four methods. Overall, SCAPE has a consistently higher F-score than other methods even when using 30 bp as the cutoff (Figure 2C).

To demonstrate that the SCAPE can identify annotated, novel and multiple APA sites, we cross-validated the SCAPE-predicted pA sites using data generated by different sequencing platforms or methods. For example, two of

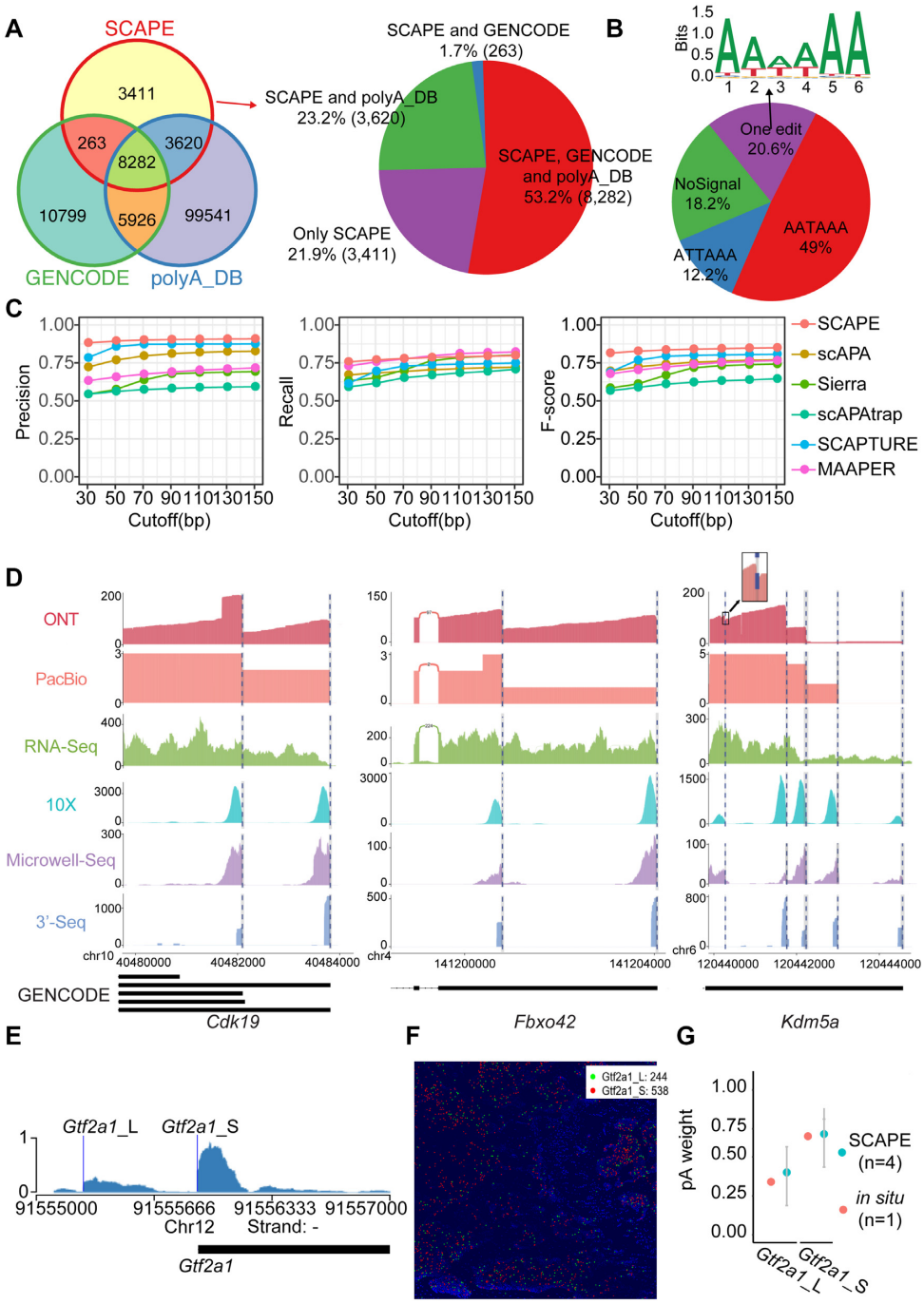

Figure 2. SCAPE discovers reproducible unannotated APA. (**A**) Comparison of SCAPE inferred pA sites from MBM 10x genomics dataset with pA annotations from GENCODE v13 and polyA_DB. Inferred pA sites within 50 bp of annotated pA sites are treated as the same pA site. (**B**) Proportion of poly(A) signals (AATAAA or ATTAAA) in the 50 bp upstream region of SCAPE inferred pA sites. The sequence logo shows the AT-rich 6-mers. (**C**) Method comparison for detecting pA sites with SCAPE, Sierra, scAPAtrap, scAPA, SCAPTURE and MAAPER. Precision (left), recall (middle) and *F*-score (right) of different methods on MBM datasets with varying cutoff (x-axis) for matched pA sites. Matched pA sites are predicted pA sites that are within *x* bp (cutoff) of the ground truth. (**D**) Sashimi plots of pA sites in *Cdk19*, *Fbxo42* and *Kdm5a* from mouse bone marrow inferred by SCAPE. PacBio full length sequencing, ONT, bulk RNA-seq, scRNA-seq (10x genomics) were generated from the mouse bone marrow sample, and Microwell and 3'-Seq are public datasets of mouse bone marrow. Dashed lines and the surrounding grey areas indicate the mean and standard deviation (std) of predicted pA sites. Transcript annotations from GENCODE v13 are given at the bottom. (**E**) Sashimi plots of SCAPE inferred pA sites of *Gtf2a1* in mouse bone marrow. Vertical blue lines indicate the estimated pA sites of SCAPE. Transcript annotation from GENCODE v13 is given at the bottom. (**F**) *in situ* hybridization of two APA isoforms of *Gtf2a1* in mouse bone marrow. The long (*Gtf2a1_L*) and short (*Gtf2a1_S*) isoforms are colored by green and red, and the signal counts are provided for each isoform. (**G**) Comparison of SCAPE predicted weights (blue) and *in situ* hybridization signals (red) of *Gtf2a1* isoforms in mouse bone marrow. The dots and error bars represent mean and standard deviation of SCAPE estimations of 4 biological replicates from microwell-based mouse cell atlas dataset.

the three annotated pA sites in *Cdk19* were identified by SCAPE using data from both 10x and Microwell-Seq, and further validated using 3'-seq, ONT and PacBio datasets. More importantly, SCAPE identified one and four novel pA sites in *Fbxo42* and *Kdm5a* from the MBM data, respectively (Figure 2D), which were supported by the ONT or PacBio derived from the same sample as well as public Microwell, and the 3'-seq data validated 4 pA sites except the most proximal one. In contrast, bulk RNA-seq data did not show clear change points at these sites, suggesting that APA inference directly based on bulk RNA-seq might be challenging. Next, we validated a distal novel pA site (Figure 2E) predicted by SCAPE in a transcription factor *Gtf2a1*, which was associated with white blood cell counts in genome-wide association studies (56,57). We performed *in situ* hybridization (Figure 2F) for different APA isoforms of *Gtf2a1* and their expression level were consistent with the SCAPE predicted isoform weights (Figure 2G). In summary, our results suggested that SCAPE outperformed other methods in terms of precision and recall at different cutoffs, providing the highest sensitivity and accuracy in both the simulated and real datasets. Moreover, SCAPE enabled *de novo* identification of APA sites with high accuracy and robustness in mouse bone marrow at single-cell level.

APA provides an extra layer of information for cell clustering

Next, we applied SCAPE on the mouse cell atlas (MCA) dataset to chart pA sites across 36 mouse organs at single-cell level (12). A total of 31 558 pA sites were identified from 119 921 cells, 17 751 (56.2%) pA sites of which were annotated in GENCODE v13 (54) or polyA_DB (55), while 13 807 (43.8%) were unannotated (Supplementary Figure S3A). 65% of them had canonical or one-edited poly(A) signals within their 50 bp upstream regions (Supplementary Figure S3B).

To further dissect the contributions of gene expression and APA isoform to the cell clustering, we clustered the cells based on the expression of APA isoforms (read counts assigned to a pA site were used as the proxy of APA isoform). We treated the gene expression and APA isoform as 'different modals' and performed the multimodal analysis from Seurat, which uses weighted nearest-neighbor (WNN), an unsupervised method to learn the relative utility of the gene expression and APA isoform in each cell (39). For example, we extracted endothelia and epithelial cells from all organs and clustered them using WNN, where gene expression and APA isoform were used and their relative contributions to the clustering could be quantified as weights. Note that the cell type annotation is obtained by clustering each organ independently based on gene expression in the original publication (12). Independent unsupervised analysis of the gene and APA expression revealed largely consistent cell classifications in endothelia (Endo) and epithelial (Epi) cells from MCA (Figure 3A, left and middle). However, they indeed exhibited some differences. For example, Endo *Maged2*^{hi} cells, which have been identified as an subcluster from the previous study (12), were partially blended with other cells in gene expression results but separated clearly in the APA data (red circle in Figure 3A), potentially due to the cell-specific APA isoforms such as *Clic4* (Supplementary Figure S3C-D). We thus used the WNN graph to derive an integrated UMAP with both gene and APA expression measured within a cell and to obtain a joint definition of cellular state for Endo and Epi cells (Figure 3A, right). Moreover, when comparing the unsupervised weights calculated by WNN from Seurat (Figure 3B), we noticed that cells classified as Endo *Maged2*^{hi} cells were assigned higher APA modality weights, while Epi (Fetal kidney) cells were assigned higher gene expression modality weights, suggesting that both gene and APA level contribute to the improvement of cell clustering in different cell types.

Next, we applied this multimodal integration of gene expression and APA to derive WNN UMAPs for 36 mouse organs (Supplementary Figure S3E, F). In total, 67 clusters were identified based on APA isoform expression (Figure 3C). Next, we identified 1608 tissue-specific APA isoforms (Figure 3D, Supplementary Table S4). Moreover, we noticed that tissues from fetal and neonatal stages tend to be clustered together, suggesting that there may be some potentially development-related APA isoforms (Figure 3D, E). To compare the levels of APA across various organs, we quantified the number of genes with multiple pA sites (Figure 3F) in each organ by randomly selecting 3000 UMI per cell, mitigating the biases brought by varying sequencing depth. Interestingly, testis showed one of the highest numbers of multi-pA genes with the most unannotated pA sites and was clustered as an outlier (Figure 3D–F), consistent with the 'out-of-testis' hypothesis that testis had permissive transcription, allowing novel isoforms (e.g. Supplementary Figure S3G) and a higher transcriptional diversity to be selected when beneficial (58,59). Taken together, these results suggest that APA isoform-level quantification provides extra information and improves the accuracy of cell clustering. One possible explanation was that certain content of the cell specificity was regulated at the APA isoform level but was hidden at the gene level.

Differential APA analysis reveals coordinated regulation between APA and miRNA

To explore the extent of tissue- and cell-type specific regulation of APA transcripts, we analyzed the gene expression and APA isoforms in the brain and bone marrow cells (n = 9264, Figure 4A) from MCA dataset. In total 15 clusters were formed, and 1348 APA isoforms were differentially expressed compared between two tissues (Supplementary Table S6). These isoforms belonged to genes that were enriched in tissue-relevant pathways such as 'ensheathment of neurons' in the brain and 'defense response to bacterium' in the blood (Figure 4B, Supplementary Table S5). Among 161 highly tissue-specific pA sites (P-value < 0.01 & $|\log_2(\text{fold change})| > 1$), we presented two genes *Mbp* and *Bzw1* as examples, where differential expression was observed for APA isoforms but not for the whole gene. *Mbp*, a gene related to neuro and IL-1 pathway (60), was expressed both in the bone marrow and brain (Figure 4C). In contrast, the proximal pA site (18:82558711:+) was specifically expressed in blood cells, whereas the distal pA site of Mbp (18:82585640:+) was preferentially expressed in the brain cells (Figure 4D). A similar pattern was also found in Bzw1,

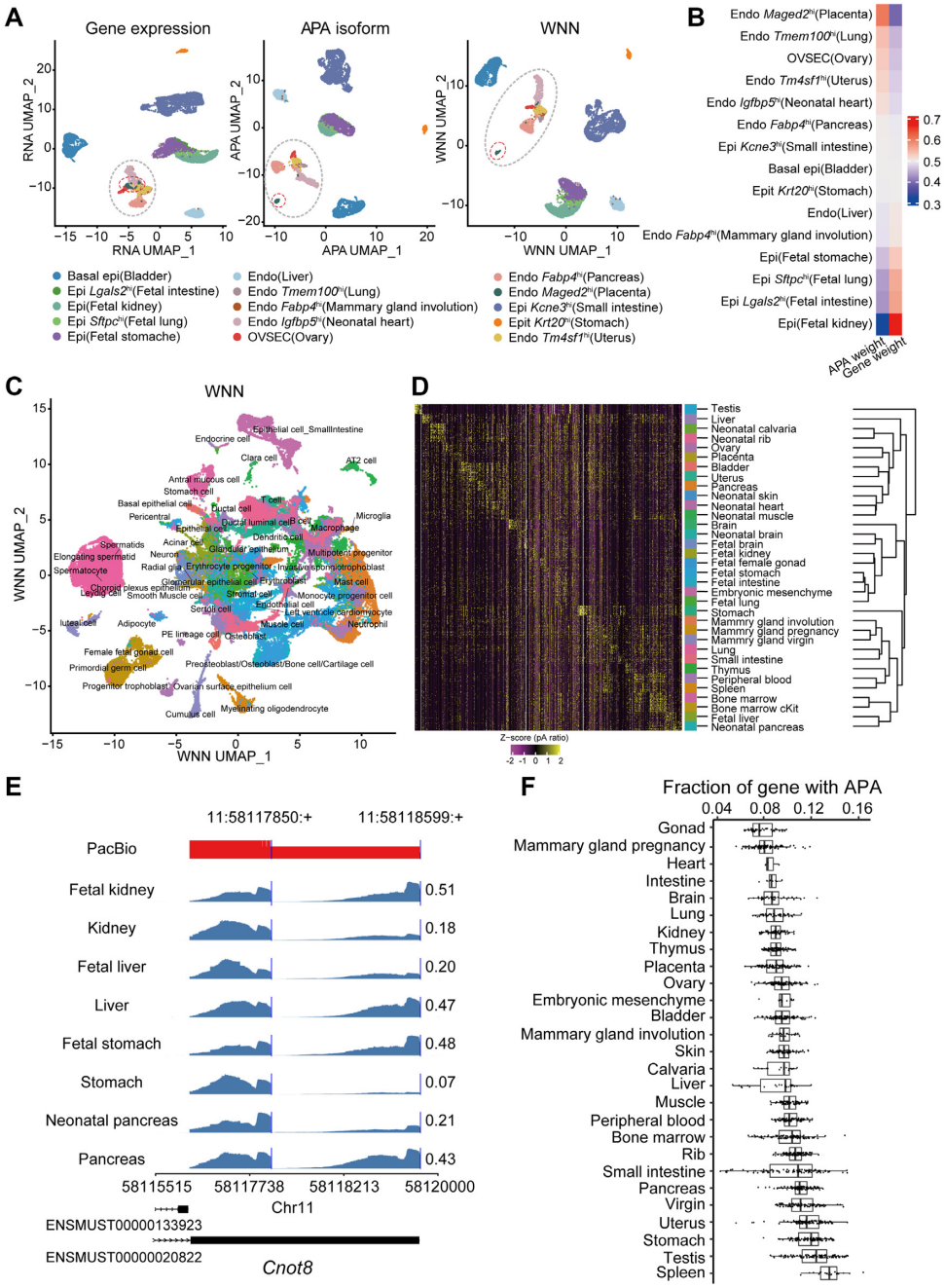

Figure 3. Tissue- and cell-type specific APA isoforms improve cell clustering. (**A**) UMAP plot of gene expression (left), APA isoforms (middle) and weighted combination of gene and APA isoforms (right) similarities (WNN) for endothelia and epithelial cells from MCA. Cell types are indicated by different colors. Endo *Maged*^{hi} (red dotted circle) do not separate in gene expression analysis, but form distinct cluster in APA and WNN analysis. Epi, epithelial cell; Endo, Endothelia cell; OVSEC, Ovarian vascular surface endothelium cell. (**B**) Mean gene and APA modality weights for cells in (A). Modality weights were calculated for each cell without knowledge of cell type labels. (**C**) UMAP plot using WNN analysis based on gene expression and APA isoforms of all cells from 36 mouse organs. Organs are indicated by different colors. Cell types of clusters are labeled. (**D**) A hierarchical clustering heatmap showing differentially expressed pA (columns) across mouse organs (rows). Yellow indicates to a high expression level, and purple corresponds to low expression levels. (**E**) Sashimi plot showing differential isoform usage of *Cnot8* across development stages. Proximal pA site 11:58117850:+ (left) and distal pA site 11:58118599:+ (right) exhibit a variable expression in fetal, neonatal and adult organs. PacBio transcripts are provided at the top and the annotations of *Cnot8* transcripts are provided at the bottom. (**F**) Fraction of multi-pA genes (≥2 pA sites), in each organ after randomly selecting 3000 UMI per cell. Each dot represents a cell. Spleen and testis have the highest proportions of multi-pA genes.

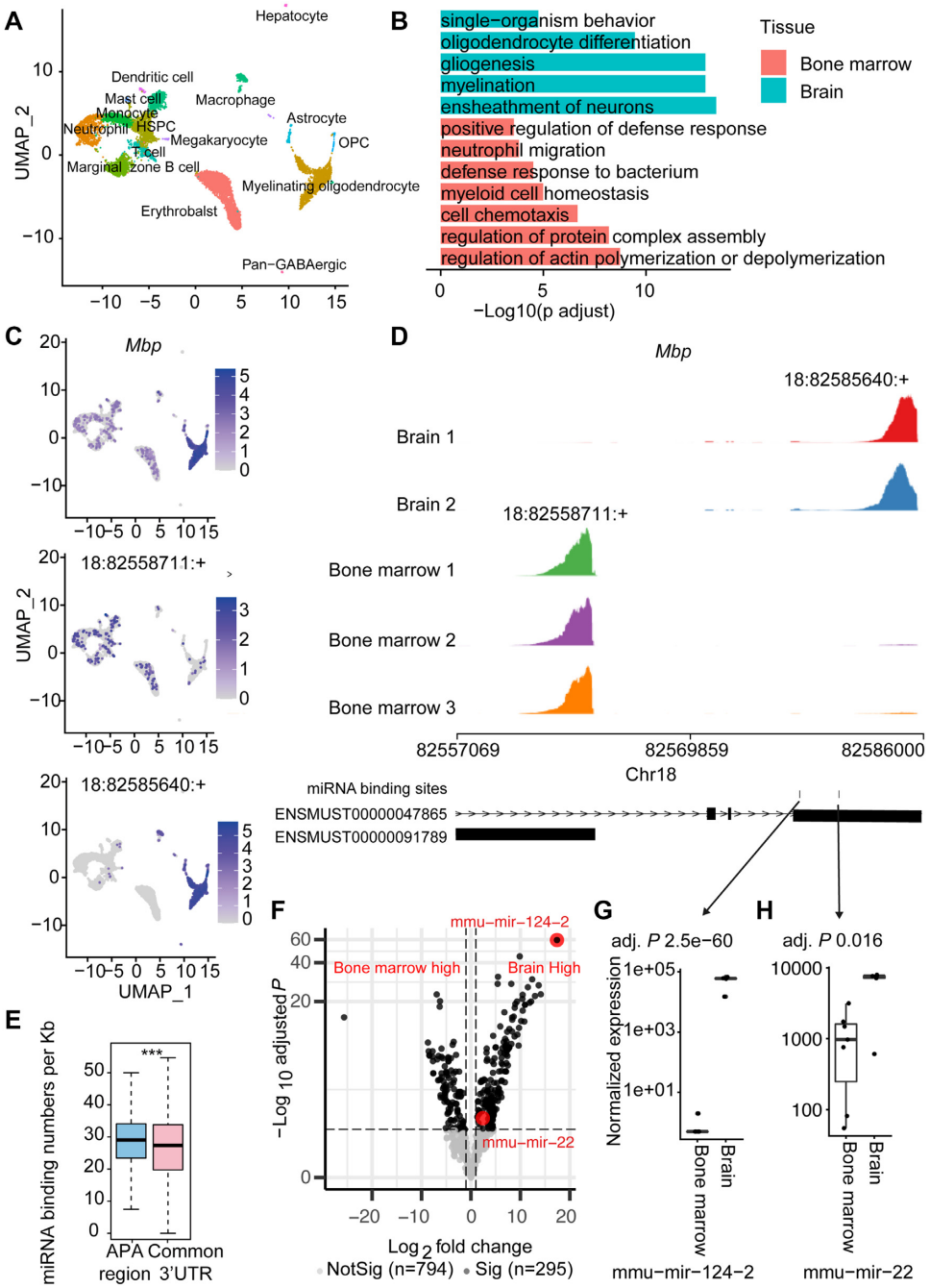

Figure 4. Differential pA sites between mouse bone marrow and brain. (**A**) UMAP plot of cells from mouse bone marrow and brain based on gene expression from MCA dataset. Cell types are indicated by colours and texts. The left side contains mainly bone marrow cells and the right side are brain cells. (**B**) GO enrichment of 1044 differentially expressed APA isoforms (correspond to multi-pA genes) in bone marrow and brain. Genes expressed in the target organ were used as the background. (**C**) *Mbp* gene (top) and pA sites expression (middle and bottom) projected on UMAP plot (A). Each dot represents a cell and the color represents the normalized expression using the 'NormalizeData' function of Seurat. *Mbp* showed expression in both bone marrow and brain. pA site 18:82558711:+ is preferentially used in the bone marrow, while pA site 18:82585640:+ is preferentially used in the brain. (**D**) Sashimi plot showing differential isoform usage of *Mbp* between brain and bone marrow. Annotations of miRNA binding sites and *Mbp* transcripts are provided at the bottom. Proximal pA site 18:82558711:+ (left) exhibits a higher expression in three biological replicates of bone marrow, which has no miRNA binding sites. Distal pA site 18:82585640:+ (right) exhibits a higher expression in two biological replicates of brain, which has two miRNA binding sites. (**E**) Box plot of miRNA binding numbers per kilobase among APA region and common 3' UTR region (Wilcox test $P = 8.727e^{-9}$). APA region represented variable 3' UTR region among two transcript which caused by alternative polyadenylation. (**F**) Differentially expressed miRNAs targeting *Mbp*. The Volcano plot shows all miRNA expressions in the brain versus bone marrow. X-axis and y-axis are the log2 fold change (significance cutoff = ±1) and negative log10 adjusted p-value (significant cutoff = 0.05) of miRNA expression comparison between brain and bone marrow. Each dot represents a miRNA with gray dots meaning no significance. The left and right plates contain miRNAs highly expressed in bone marrow and brain, respectively. 2 (red dots) out of 2 miRNAs targeting the 3' UTR of *Mbp* passed statistical significance. Particularly, mmu-mir-124–2 is the top differentially expressed miRNA between the brain and bone marrow, among all miRNAs. (**G**, **H**) Two differentially expressed miRNAs targeting *Mbp*: mmu-mir-124-2 (G) and mmu-mir-22 (H). The arrows indicate their genomic locations in panel (D). Y-axis is the normalized expression in log scale and x-axis indicates the organs. Both mmu-mir-124-2 and mmu-mir-22 are highly expressed in the brain.

coding for an RNA binding protein, whose APA isoforms were switched between brain and blood cells (Supplementary Figure S4B–E).

The 3' UTRs often contain binding sites of miRNAs and RNA binding proteins involved in the regulation of transcripts. We predicted miRNA binding sites in 3' UTR of each gene using TargetScan (61) and calculated the miRNA binding numbers per kilobase (kb) of the APA shortening and inclusion regions in the mouse genome. A significantly higher number of miRNA binding was observed in the APA shortening region than that of in the common 3' UTR region (Wilcox test $P = 8.727e^{-9}$, Figure 4E), consistent with the roles of miRNA in regulation APA isoforms (62). To investigate the potential links between the tissue-specific APA isoforms and miRNAs, we compared the miRNA expression between the mouse bone marrow and brain based on data from miRbase (63). Among the 295 differentially expressed miRNA (Supplementary Table S7), we highlighted two miRNAs that predicted binding to the upstream region of the distal pA sites of *Mbp* (Figure D). Both miRNAs (Figure F–H) exhibited differential expression for *Mbp*, where mmu-mir-124-2 exhibited the highest significance and expression fold-change (adjusted $P = 2.5e^{-60}$) across all extracted miRNAs, the co-expression of the *Mbp* long isoform and mmu-mir-124-2. Notably, SCAPE enables the investigation of the heterogeneity of this tissue-specific APA of *Mbp* at the single-cell level, we found that the megakaryocyte and macrophage, even though they were from bone marrow, use the distal pA sites, resemble the APA usage of cells from the brain (Supplementary Figure S4A), suggesting that the APA heterogeneity exists within a tissue and the necessity of studying APA at the single-cell level. For *Bzw1*, we observed that the short isoform was co-expressed with mmu-mir-142a in the bone marrow (Supplementary Figure S4F–H), whereas the brain-specific long isoform contains two mmu-mir-144 binding sites (Supplementary Figure S4F, G, I), which showed a 320 fold up-regulation in the brain. These results suggested that tissue-specific APA usage might be associated with expression of miRNAs that bind to their 3' UTR, indicating the coordinated regulation between APA isoform and miRNA.

Tumor-specific APA expression and regulation in human glioblastoma

To characterize tumour-specific APA events, we compared the APA expression of neuronal cells between tumour cells and normal cells from human GBM (HGB) scRNA-seq data (44) using SCAPE. Both immune and neuronal cells were presented in the samples, according to the clustering results of all cells (Supplementary Figure S5A). Here, we only kept the neuronal cells ($n = 6513$) for comparison, which were further clustered into five cell types (Figure A), including neurons, astrocytes, oligodendrocyte progenitor cells (OPC), oligodendrocytes and radial glia. *KIF1B*, involved in neuronal apoptosis and transport of synaptic vesicles (64), was expressed in all five cell types. The proximal pA site (1:103066788:+, ENST00000377093) was highly expressed in the radial glia, astrocytes and OPCs, whereas the distal pA site (1:10381671:+, ENST00000622724) was the major isoform in the oligodendrocytes and neurons with 26 additional exons (Figure 5B, C), suggesting an isoform switch in different cell types.

Differential APA usage was also observed in tumors compared to healthy tissues. First, we performed CNA analysis for detecting the presence and/or absence of somatic CNA indicated malignant or non-malignant cells, respectively (Online Methods). CNA analysis revealed large-scale amplifications and deletions in most cells, including the glioblastoma hallmarks of chromosome 7 gain and chromosome 10 loss (Supplementary Figure S6). Next, since astrocyte was the major cell type in HGB samples from all three patients (Supplementary Figure S5B), we only kept astrocytes from patient SF11159 (with the highest sequencing depth) for the comparison to exclude differences in patients and cell types. Clustering of the selected astrocytes (Figure 5D) showed two separate clusters for tumor and normal cells, where tumor cells were identified by CONICSmat (46). *CALM3*, coding for a protein that binds calcium and functions as an enzymatic cofactor (65), had an APA isoform switch (Figure 5E-F), where distal pA site (19:46610849:+) was more prevalent in the malignant cells. Expressions of miRNAs targeting 3' UTR of *CALM3*, in normal and tumor cells were obtained from microRNAome (50) and TCGA (66), respectively. We found that 9 out of 11 miRNAs exhibited significant differences (5 up-regulation and 4 down-regulation) between normal and tumor cells (Figure 5G, Supplementary Figure S5C), such as hsa-mir-532 (Figure 5I) and hsa-mir-122 (Figure 5H).

Dynamic APA usage analysis reveals APA preference in cell differentiation

The observations of widespread changes of 3' UTR length during stem cell differentiation, early embryonic development, and somatic reprogramming (7,8,67) at the bulk level, suggested that APA might be tightly regulated during cell fate determination. To assess whether SCAPE can identify these dynamic changes of APA, we generated the scRNA-seq data from mouse bone marrow. Using the 10x Genomics platform, we sequenced and filtered cells based on stringent criteria (Online Methods), resulting in 4301 high-quality single cells from adult mice. We clustered them using unbiased graph-based clustering (68) and classified 11 cell types based on canonical markers (Supplementary Figure S7A, B). Next, we focused on 1,823 cells involved in erythropoiesis including hematopoietic stem and progenitor cells (HSPC), pro-erythroblast (ProE), basophilic-erythroblast (BasoE), early- and late- polychromatic erythroblast (PolyE), and orthochromatic erythroblast (OrthoE) (Figure 6A, Supplementary Figure S7C).

Next, we reconstructed the cell differentiation trajectory based on RNA velocity using scVelo (41). Indeed, the RNA velocity showed a trend from HSPC to OrthoE, reflecting the erythrocyte differentiation (Figure 6B). A shortening trend was observed after averaging pA length over all genes in each cell type during erythropoiesis (Figure 6C). We then calculated an APA usage preference index, ranging from 0 (the proximal pA site) to 1 (the distal pA site) by grouping adjacent 10 cells in the pseudotime space into one to reduce random noises (Online Methods). Interestingly, we observed that the APA usage index of 89% genes (245 out of 275 genes with multiple pA sites) decreased over the pseu-

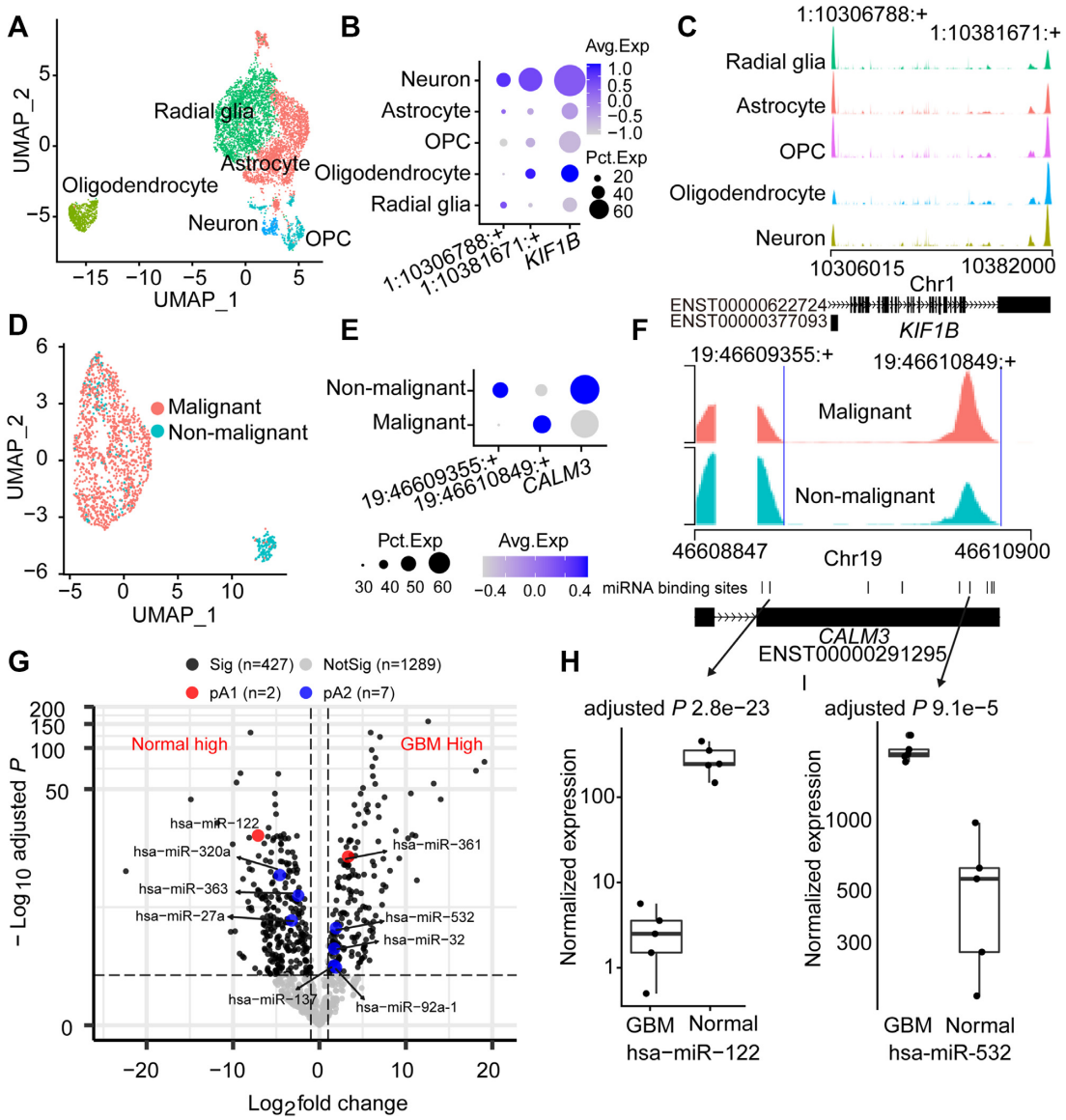

Figure 5. Differential APA analysis of cell types and cell states in human GBM. (**A**) UMAP plot of non-myeloid cells based on gene expression from three Glioblastoma patients. Cell types are indicated by different colours and texts. OPC, Oligodendrocyte progenitor cells. (**B**) Relative expression of *KIF1B* and its two pA sites in different cell types. The size of the circle represents the proportion of expressing cells. The colour indicates the average expression level over expressing cells of a given cell type. (**C**) Sashimi plot of *KIF1B* in different cell types. Oligodendrocytes and neuron preferentially express the distal pA site 1:10381671:+. (**D**) UMAP plot of non-malignant and malignant astrocyte cells based on gene expression from patients SF11159. Cells are classified into malignant (red) and non-malignant (blue) cells using CONICSmat. (**E**) Relative expression of *CALM3* and its two pA sites in malignant and non-malignant astrocyte cells, with circle size and colour indicating expressing cell proportion and average expression level, respectively. pA site 19:46609355:+ is mainly expressed in non-malignant cells, while 19:46610849:+ is mainly expressed in malignant cells. (**F**) Sashimi plot of *CALM3* in malignant and non-malignant astrocyte cells. Blue lines represent the pA sites. Malignant cells exhibit a preference on pA site 19:46610849:+. (**G**) Differentially expressed miRNAs targeting *CALM3*. The Volcano plot shows miRNA expressions in normal brain versus GBM (Glioblastoma) tissues. X-axis and y-axis are the log2 fold change (significance cutoff = ±1) and negative $\log_{10}$ adjusted *P*-value (significant cutoff = 0.05) of miRNA expression comparison between normal and GBM tissues. Each dot represents a miRNA with gray dots meaning no significance. The left and right plates contain miRNAs highly expressed in normal or GBM tissues, respectively. Nine (red and blue dots) out of 10 miRNAs targeting the 3' UTR of *CALM3* show significance. Red dots are miRNA targeting the proximal pA site 19:46609355:+, while blue dots target the distal pA site 19:46610849:+. (H-I) Two example differentially expressed miRNAs targeting *CALM3*: hsa-miR-122 (**H**) and hsa-miR-532 (**I**). The arrows indicate their genomic locations in panel (F). Y-axis is the normalized expression in log scale.

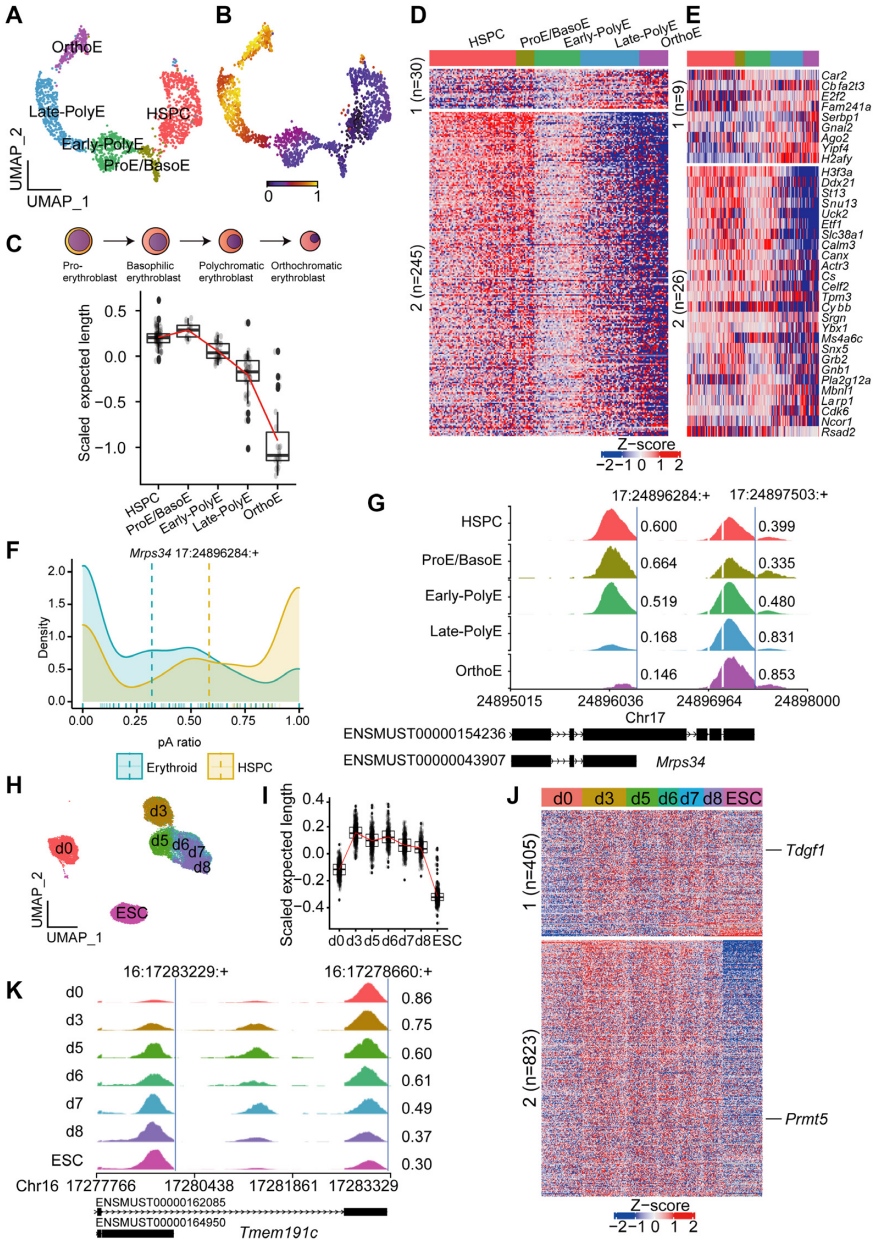

Figure 6. Dynamic APA usage in erythropoiesis and somatic cell reprogramming. (**A**) UMAP plot (gene expression) of cells in erythropoiesis of mouse. Cell types are indicated by colors and texts. (**B**) Pseudotime of each cell estimated by scVelo. Cells are projected to the same UMAP plot. The color indicates the estimated pseudotime. (**C**) Boxplot of the mean of the expected pA lengths of all genes in each cell type in the erythropoiesis process. Each dot represents agglomerated 10 neighbouring cells in pseudotime space. The expected pA length is decreasing from HSPC to OrthoE in the erythropoiesis process. The y-axis shows the scaled value (*z*-score) across all data points. (**D**) Dynamics of expected pA length of all genes in erythropoiesis. The color reflects expected pA length (standardized). Each row represents a gene. Each column represents 10 grouped neighboring cells in pseudotime space. Colours in the top row indicate differentiating cell types from HSPC to OrthoE, with the same colour scheme in (a). The upper and lower parts are genes with increasing and decreasing trends, respectively. (**E**) Dynamics of expected pA length of 35 driver genes estimated by scVelo (erythropoiesis). Genes are classified into increasing (upper) and decreasing (lower) trends. Known driver gene such as *Cdk6* show a decreasing trend. (**F**) Density plot of pA usage with different pA category of *Mrps34* (17:24896284:+) between HSPC (J shape) and erythroid (L shape). (**G**) Sashimi plot of *Mrps34* in different cell types during erythropoiesis. Blues lines represent pA sites. Numbers in the coverage peaks represent the corresponding proportion of the pA sites. The usage of proximal pA site 17:24896284:+ gradually decreases, while the usage of distal pA site 17:24897503:+ gradually increases. (**H**) UMAP plot (gene expression) of iPSC. Cells from different stage are indicated by colors and texts. (I) Boxplot of the mean of the expected pA lengths of all genes in each stage in the iPSC process. Each dot represents agglomerated 10 neighboring cells in pseudotime space provided by local embedding. The expected pA length is first increasing from day 0 (d0) and then decreasing from day 3 (d3) to ESC. Note that the y-axis shows the scaled value (*z*-score) across all data points. (**J**) Dynamics of expected pA length of all genes (iPSC). The color reflects expected pA length (standardized). Each row represents a gene. Each column represents 10 grouped neighboring cells in the pseudotime space, with pseudotime estimated using local embedding. Colours in the top row indicate differentiating cell types from day 0 to ESC, with the same colour scheme as panel (H). (**K**) Sashimi plot of *Tmem191c* in different iPSC differentiation stages. Blues lines represent pA sites. Numbers in the coverage peaks represent the corresponding pA site's proportion. In the iPSC differentiation process, the usage of the proximal pA site 16:17283229:+ gradually increases, while the usage of distal pA site 16:17278660:+ gradually decreases. Note that the central unannotated peak is from the anti-sense strand, which is not relevant to *Tmem191c*.

dotime in our linear regression analysis (Figure 6D, Supplementary Table S8). Next, we inferred the driver genes in erythrocyte differentiation using scVelo. Importantly, driver genes tended to be with dynamic APA usage (Figure 6E, chi-square test, $P = 1.61e^{-4}$). The distribution of APA usage can be further classified, indicating the minor (L shape) or major (J shape) in each cell type (Supplementary Figure S7D, E). We found that genes having a switch of APA distribution were enriched with GO terms (Supplementary Figure S8A) including structural constituent of ribosome (e.g. mitochondrial ribosomal protein *Mrps34*). The major isoform of *Mrps34* switched from HSPC to erythroid, elongating the 3' UTR during the process (Figure 6F, G). Taken together, SCAPE enabled the dissection of dynamic APA usages during *in vivo* cell differentiation. Also, we discovered that the majority of genes, including driver genes, showed a shortening trend during erythropoiesis at the single-cell level, suggesting that APA regulation may be of importance during the red blood cell differentiation.

Next, we investigated how APA changed when somatic cells were reprogrammed into induced pluripotent stem cells (iPSCs). We analysed 34 174 cells from a time-course Oct4/Sox2/Klf4 (OSK) reprogramming process in mouse (43). UMAP based on pA counts showed that cells at day 0 (d0) were well-separated from other cells, but the cells from d3 to d8 exhibited a contiguous trajectory (Figure 6H). The dynamic changes of APA usage were also observed when comparing the averaged pA length (Figure 6I), exhibiting a sharp increase from d0 to d3 followed by a gradually decreasing pattern towards d8. Next, we estimated the pseudotime of each cell using local embedding (69) and merged 10 adjacent cells in the pseudotime space for further analysis. 405 (33%) multi-pA genes, including gene (e.g. *Tdgf1*) that is important for mature iPSC formation (70), lengthened their 3'UTR during the reprogramming, whereas the majority (823) of genes such as *Prmt5* that enhances the generation of iPSCs (71) showed a decreasing trend (Figure 6J). Notably, in addition to the known regulatory genes, we identified that *Tmem191c*, a gene coding for a transmembrane protein, presented a dynamic change to the proximal pA site during the iPSC reprogramming process (Figure 6K). Differentially expressed pA sites at d0 and d3 were enriched for epithelial cell proliferation and cytokine-mediated of cell migration, in accordance with previous studies (43,72) showing that stem cell fate and immune responses were interconnected. We also found that differentially expressed pA sites between d8 and embryonic stem cells (ESC) were enriched for functions related to transitioning from the naïve to the primed pluripotent state, such as cell cycle phase transition, embryonic development pathway and pyruvate metabolic process (Supplementary Figure S8B).

Taken together, SCAPE revealed the APA usage dynamics during cell differentiation and iPSC reprogramming at single-cell level. We identified potential driver genes that utilized different APA isoforms in erythropoiesis and a dramatic change of APA usage from d0 to d3 during iPSC reprogramming. Future experiments are warranted to explore of the potential roles these dynamic APA usages of driver genes.

DISCUSSION

We have developed a powerful probabilistic model named SCAPE, facilitating the *de novo* identification and quantification of pA sites from poly(A)-enriched scRNA-seq data. SCAPE used the cDNA insert size to further improve the accuracy of APA calling. Note that the accuracy of SCAPE relies on the prior information of (a) insert size, (b) poly(A) length and (c) poly(A)-enrichment process. In addition, the poly(T) oligonucleotide may bind to A-rich region in the gene body rather than the poly(A) tail in scRNA-seq library preparation, which may affect the accuracy of predicted pA sites. Nonetheless, we reasoned that these factors can be addressed to some extent. First, a size-selection step is commonly used in paired-end RNA-seq to control the mean length of inserted cDNA fragments, which may be acquired from the experiment protocol or can be measured precisely from read pairs that mapped to large constitutive regions such as intronless 3' UTRs (73). Second, the poly(A) tail length may be different between transcripts/cell types. We could estimate its distribution from cleavage reads in the read 1 of scRNA-seq data (Methods). ONT direct RNA sequencing and other methods such as TAIL-seq have reported that poly(A) tail length ranges mostly from 50 to 100 nt) in mammalians (52,53,74,75), In SCAPE, we used 20–150 nt as the range which should be modified accordingly if the data is generated from other organisms. Lastly, most common single-cell RNAs-seq protocols rely on oligo-DT primers to enrich polyadenylated mRNA molecules. scRNA-seq sometimes contains 15–25% unspliced intronic sequences, mostly originating from secondary priming positions within the intronic regions (76). To avoid capturing these internal priming polyT and calling them poly(A) sites, SCAPE separated the regions into 3' UTR and intronic regions. When a potential poly(A) site was identified, it was compared with the internal priming regions and was removed when this site is proximal to these priming regions.

In summary, we have demonstrated that APA isoforms provide an additional level of information that improves the cell clustering. Our method enables investigation of the novel APA signals and cell type-specific isoforms in different cell types as well as exploration of the heterogeneity and differential regulation of APA in various tissues. The tissue- and cell type- specificity regulated at the post-transcriptional level, which sometimes are latent at gene expression level, could be used to infer cell identity in the task of cell type classification. Furthermore, our method raises the exciting prospect of identifying new cell subsets using known or novel pA sites as markers. Given a large amount of differentially expressed APA have been identified. It is worth noting that some may be originated from biological and technical noises. Our method and results will facilitate the identification of the functional APA isoforms. Future experiments to explore the role of dynamic APA usage and its regulation related to miRNA will further our understanding the regulation and biological function of APA during cell differentiation, reprogramming and tumorigenesis.

DATA AVAILABILITY

SCAPE is available at https://github.com/LuChenLab/SCAPE. We deposited the mouse bone marrow dataset in the NCBI Sequence Read Archive (SRA) under accession No. PRJNA706066. All other data are available from the corresponding authors upon reasonable request.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Yu Zhou and Sisi Li for technical support, and other members from Chen's lab for constructive comments and critically reading the manuscript. L. Cheng acknowledges the computational resources provided by the Aalto Science-IT project. We also thank the anonymous reviewers' comments and suggestions.

Author contributions: L.C., L. Cheng. and J-w.L. designed the experiments. L.C. oversaw the whole project. L.C and L. Cheng supervised the bioinformatics, L. Cheng undertook and supervised the statistical modelling, J-w.L. supervised the experiments. R.Z. implemented SCAPE and performed methods comparison. R.Z. assisted by X.X., P.H., M-y.X., D.Z., Q-x.Y. and C.T., conducted the bioinformatics, processed the scRNA-seq. X-r.Z. and J.-w.S. conducted the scRNA-seq, PacBio and ONT of mouse bone marrow. Y-c.Z., L.-f.Z. and S.-s.C. performed *in situ* validation. All authors read and approved the contents of the manuscript.

FUNDING

National Key Research and Development Program of China, Stem Cell and Translational Research [2017YFA0106800 to L.C. and J-w.L.]; National Science Fund for Excellent Young Scholars [81722004 to L.C.]; Marie Curie Individual fellowship [663830 to L.Cheng]; Wellcome Trust Institutional Strategic Support Fund (Cardiff University) and Academy of Finland [335858 to L.Cheng]. Funding for open access charge: Core funding of West China Second University Hospital, Sichuan University.

Conflict of interest statement. None declared

REFERENCES

1. Singh,I., Lee,S.H., Sperling,A.S., Samur,M.K., Tai,Y.T., Fulciniti,M., Munshi,N.C., Mayr,C. and Leslie,C.S. (2018) Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nat Commun*, **9**, 1716.
2. Passacantilli,I., Panzeri,V., Bielli,P., Farini,D., Pilozzi,E., Fave,G.D., Capurso,G. and Sette,C. (2017) Alternative polyadenylation of ZEB1 promotes its translation during genotoxic stress in pancreatic cancer cells. *Cell Death Dis*, **8**, e3168.
3. Thivierge,C., Tseng,H.W., Mayya,V.K., Lussier,C., Gravel,S.P. and Duchaine,T.F. (2018) Alternative polyadenylation confers pten mRNAs stability and resistance to microRNAs. *Nucleic Acids Res*, **46**, 10340–10352.
4. Berkovits,B.D. and Mayr,C. (2015) Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature*, **522**, 363–367.
5. Chen,M. and Manley,J.L. (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell. Biol.*, **10**, 741–754.
6. Mayr,C. (2016) Evolution and biological roles of alternative 3'UTRs. *Trends Cell Biol.*, **26**, 227–237.
7. Ji,Z., Lee,J.Y., Pan,Z., Jiang,B. and 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.
8. Ji,Z. and 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.
9. Venkat,S., Tisdale,A.A., Schwarz,J.R., Alahmari,A.A., Maurer,H.C., Olive,K.P., Eng,K.H. and Feigin,M.E. (2020) Alternative polyadenylation drives oncogenic gene expression in pancreatic ductal adenocarcinoma. *Genome Res*, **30**, 347–360.
10. Hashimshony,T., Wagner,F., Sher,N. and Yanai,I. (2012) CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep*, **2**, 666–673.
11. Zheng,G.X., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat Commun*, **8**, 14049.
12. Han,X., Wang,R., Zhou,Y., Fei,L., Sun,H., Lai,S., Saadatpour,A., Zhou,Z., Chen,H., Ye,F. *et al.* (2018) Mapping the mouse cell atlas by microwell-seq. *Cell*, **173**, 1091-1107.
13. Shulman,E.D. and Elkon,R. (2019) Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic Acids Res.*, **47**, 10027–10039.
14. Wu,X., Liu,T., Ye,C., Ye,W. and Ji,G. (2021) scAPAtrap: identification and quantification of alternative polyadenylation sites from single-cell RNA-seq data. *Brief Bioinform*, **22**, bbaa273.
15. Patrick,R., Humphreys,D.T., Janbandhu,V., Oshlack,A., Ho,J.W.K., Harvey,R.P. and Lo,K.K. (2020) Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome Biol*, **21**, 167.
16. Gao,Y., Li,L., Amos,C.I. and Li,W. (2021) Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression. *Genome Res*, **22**, 222.
17. Li,G.W., Nan,F., Yuan,G.H., Liu,C.X., Liu,X., Chen,L.L., Tian,B. and Yang,L. (2021) SCAPTURE: a deep learning-embedded pipeline that captures polyadenylation information from 3' tag-based RNA-seq of single cells. *Genome Biol*, **22**, 221.
18. Li,W.V., Zheng,D., Wang,R. and Tian,B. (2021) MAAPER: model-based analysis of alternative polyadenylation using 3' end-linked reads. *Genome Biol*, **22**, 222.
19. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M. *et al.* (2015) Highly parallel Genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.
20. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
21. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M. 3rd, Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.
22. Anders,S., Reyes,A. and Huber,W. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res*, **22**, 2008–2017.
23. Morey,R.D. and Rouder,J.N. (2018) *BayesFactor: Computation of Bayes Factors for Common Designs*.
24. Song,Y., Botvinnik,O.B., Lovci,M.T., Kakaradov,B., Liu,P., Xu,J.L. and Yeo,G.W. (2017) Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Mol Cell*, **67**, 148–161.
25. Linker,S.M., Urban,L., Clark,S.J., Chhatriwala,M., Amatya,S., McCarthy,D.J., Ebersberger,I., Vallier,L., Reik,W., Stegle,O. *et al.* (2019) Combined single-cell profiling of expression and DNA methylation reveals splicing regulation and heterogeneity. *Genome Biol*, **20**, 30.
26. Ha,K.C.H., Blencowe,B.J. and Morris,Q. (2018) QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol*, **19**, 45.

27. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
28. Gordon,S.P., Tseng,E., Salamov,A., Zhang,J., Meng,X., Zhao,Z., Kang,D., Underwood,J., Grigoriev,I.V., Figueroa,M. *et al.* (2015) Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS One*, **10**, e0132628.
29. Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.
30. De Coster,W., D'Hert,S., Schultz,D.T., Cruts,M. and Van Broeckhoven,C. (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, **34**, 2666–2669.
31. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
32. Tang,A.D., Soulette,C.M., van Baren,M.J., Hart,K., Hrabeta-Robinson,E., Wu,C.J. and Brooks,A.N. (2020) Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun*, **11**, 1438.
33. Ke,R., Mignardi,M., Pacureanu,A., Svedlund,J., Botling,J., Wahlby,C. and Nilsson,M. (2013) In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods*, **10**, 857–860.
34. Liu,S., Punthambaker,S., Iyer,E.P.R., Ferrante,T., Goodwin,D., Furth,D., Pawlowski,A.C., Jindal,K., Tam,J.M., Mifflin,L. *et al.* (2021) Barcoded oligonucleotides ligated on RNA amplified for multiplexed and parallel in situ analyses. *Nucleic Acids Res.*, **49**, e58.
35. Ren,X., Deng,R., Zhang,K., Sun,Y., Teng,X. and Li,J. (2018) SpliceRCA: in situ single-cell analysis of mRNA splicing variants. *ACS Cent. Sci.*, **4**, 680–687.
36. Codeluppi,S., Borm,L.E., Zeisel,A., La Manno,G., van Lunteren,J.A., Svensson,C.I. and Linnarsson,S. (2018) Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods*, **15**, 932–935.
37. Lamprecht,M.R., Sabatini,D.M. and Carpenter,A.E. (2007) CellProfiler: free, versatile software for automated biological image analysis. *Biotechniques*, **42**, 71–75.
38. McCarthy,D.J., Campbell,K.R., Lun,A.T. and Wills,Q.F. (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, **33**, 1179–1186.
39. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M. 3rd, Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
40. Kaminow,B., Yunusov,D. and Dobin,A. (2021) STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. bioRxiv doi: https://doi.org/10.1101/2021.05.05.442755, 05 May 2021, preprint: not peer reviewed.
41. Bergen,V., Lange,M., Peidli,S., Wolf,F.A. and Theis,F.J. (2020) Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol*, **38**, 1408–1414.
42. Yu,G., Wang,L.G., Han,Y. and He,Q.Y. (2012) clusterProfiler: an r package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287.
43. Guo,L, Lin,L, Wang,X, Gao,M, Cao,S, Mai,Y., Wu,F., Kuang,J., Liu,H., Yang,J. *et al.* (2019) Resolving cell fate decisions during somatic cell reprogramming by single-cell RNA-Seq. *Mol Cell*, **73**, 815–829.
44. Bhaduri,A., Di Lullo,E., Jung,D., Muller,S., Crouch,E.E., Espinosa,C.S., Ozawa,T., Alvarado,B., Spatazza,J., Cadwell,C.R. *et al.* (2020) Outer radial Glia-like cancer stem cells contribute to heterogeneity of glioblastoma. *Cell Stem Cell*, **26**, 48–63.
45. Dennis,G. Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*, **4**, P3.
46. Muller,S., Cho,A., Liu,S.J., Lim,D.A. and Diaz,A. (2018) CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. *Bioinformatics*, **34**, 3217–3219.
47. Sommerkamp,P., Altamura,S., Renders,S., Narr,A., Ladel,L., Zeisberger,P., Eiben,P.L., Fawaz,M., Rieger,M.A., Cabezas-Wallscheid,N. *et al.* (2020) Differential alternative polyadenylation landscapes mediate hematopoietic stem cell activation and regulate glutamine metabolism. *Cell Stem Cell*, **26**, 722–738.
48. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nat Methods*, **9**, 357–359.
49. Colaprico,A., Silva,T.C., Olsen,C., Garofano,L., Cava,C., Garolini,D., Sabedot,T.S., Malta,T.M., Pagnotta,S.M., Castiglioni,I. *et al.* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71.
50. McCall,M.N., Kim,M.S., Adil,M., Patil,A.H., Lu,Y., Mitchell,C.J., Leal-Rojas,P., Xu,J., Kumar,M., Dawson,V.L. *et al.* (2017) Toward the human cellular microRNAome. *Genome Res*, **27**, 1769–1781.
51. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
52. Liu,Y., Nie,H., Liu,H. and Lu,F. (2019) Poly(A) inclusive RNA isoform sequencing (PAIso-seq) reveals wide-spread non-adenosine residues within RNA poly(A) tails. *Nat. Commun.*, **10**, 5292.
53. Legnini,I., Alles,J., Karaiskos,N., Ayoub,S. and Rajewsky,N. (2019) FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nat. Methods*, **16**, 879–886.
54. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.*, **22**, 1760–1774.
55. Wang,R., Nambiar,R., Zheng,D. and Tian,B. (2018) PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res.*, **46**, D315–D319.
56. Chen,M.H., Raffield,L.M., Mousas,A., Sakaue,S., Huffman,J.E., Moscati,A., Trivedi,B., Jiang,T., Akbari,P., Vuckovic,D. *et al.* (2020) Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell*, **182**, 1198–1213.
57. Vuckovic,D., Bao,E.L., Akbari,P., Lareau,C.A., Mousas,A., Jiang,T., Chen,M.H., Raffield,L.M., Tardaguila,M., Huffman,J.E. *et al.* (2020) The polygenic and monogenic basis of blood traits and diseases. *Cell*, **182**, 1214–1231.
58. Kaessmann,H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res.*, **20**, 1313–1326.
59. Soumillon,M., Necsulea,A., Weier,M., Brawand,D., Zhang,X., Gu,H., Barthes,P., Kokkinaki,M., Nef,S., Gnirke,A. *et al.* (2013) Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.*, **3**, 2179–2190.
60. Campagnoni,A.T. and Skoff,R.P. (2001) The pathobiology of myelin mutants reveal novel biological functions of the MBP and PLP genes. *Brain Pathol*, **11**, 74–91.
61. Agarwal,V., Bell,G.W., Nam,J.W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.
62. Mayr,C. and Bartel,D.P. (2009) Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, **138**, 673–684.
63. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, **34**, D140–D144.
64. Schlisio,S., Kenchappa,R.S., Vredeveld,L.C., George,R.E., Stewart,R., Greulich,H., Shahriari,K., Nguyen,N.V., Pigny,P., Dahia,P.L. *et al.* (2008) The kinesin KIF1Bbeta acts downstream from egln3 to induce apoptosis and is a potential 1p36 tumor suppressor. *Genes Dev*, **22**, 884–893.
65. Berchtold,M.W. and Villalobo,A. (2014) The many faces of calmodulin in cell proliferation, programmed cell death, autophagy, and cancer. *Biochim Biophys Acta*, **1843**, 398–435.
66. Wong,H.A., Fatimy,R.E., Onodera,C., Wei,Z., Yi,M., Mohan,A., Gowrisankaran,S., Karmali,P., Marcusson,E., Wakimoto,H. *et al.* (2015) The cancer genome atlas analysis predicts MicroRNA for targeting cancer growth and vascularization in glioblastoma. *Mol. Ther.*, **23**, 1234–1247.
67. Boutet,S.C., Cheung,T.H., Quach,N.L., Liu,L., Prescott,S.L., Edalati,A., Iori,K. and Rando,T.A. (2012) Alternative polyadenylation mediates microRNA regulation of muscle stem cell function. *Cell Stem Cell*, **10**, 327–336.
68. Becht,E., McInnes,L., Healy,J., Dutertre,C.A., Kwok,I.W.H., Ng,L.G., Ginhoux,F. and Newell,E.W. (2018) Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.***37**, 38–44.

69. Roweis,S.T. and Saul,L.K. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, 2323–2326.
70. Yang,C.S., Chang,K.Y. and Rana,T.M. (2014) Genome-wide functional analysis reveals factors needed at the transition steps of induced reprogramming. *Cell Rep*, **8**, 327–337.
71. Chu,Z., Niu,B., Zhu,H., He,X., Bai,C., Li,G. and Hua,J. (2015) PRMT5 enhances generation of induced pluripotent stem cells from dairy goat embryonic fibroblasts via down-regulation of p53. *Cell Prolif*, **48**, 29–38.
72. Mosteiro,L., Pantoja,C., Alcazar,N., Marion,R.M., Chondronasiou,D., Rovira,M., Fernandez-Marcos,P.J., Munoz-Martin,M., Blanco-Aparicio,C., Pastor,J. *et al.* (2016) Tissue damage and senescence provide critical signals for cellular reprogramming in vivo. *Science*, **354**, aaf4445.
73. Katz,Y., Wang,E.T., Airoldi,E.M. and Burge,C.B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*, **7**, 1009–1015.
74. Chang,H., Lim,J., Ha,M. and Kim,V.N. (2014) TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications. *Mol. Cell*, **53**, 1044–1052.
75. Soneson,C., Yao,Y., Bratus-Neuenschwander,A., Patrignani,A., Robinson,M.D. and Hussain,S. (2019) A comprehensive examination of nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.*, **10**, 3359.
76. La Manno,G., Soldatov,R., Zeisel,A., Braun,E., Hochgerner,H., Petukhov,V., Lidschreiber,K., Kastriti,M.E., Lonnerberg,P., Furlan,A. *et al.* (2018) RNA velocity of single cells. *Nature*, **560**, 494–498.