Computational and Structural Biotechnology Journal 18 (2020) 1012–1019

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

Discovery of alternative polyadenylation dynamics from single cell types

Congting Ye [a], Juncheng Lin [a], Qingshun Q. Li [a,b,*]

[a] Key Laboratory of the Ministry of Education for Coastal and Wetland Ecosystems, College of the Environment and Ecology, Xiamen University, Xiamen, Fujian 361102, China
[b] Graduate College of Biomedical Sciences, Western University of Health Sciences, Pomona, CA 91766, USA

ARTICLE INFO

Article history:
Received 20 February 2020
Received in revised form 12 April 2020
Accepted 14 April 2020
Available online 20 April 2020

Keywords:
Alternative polyadenylation
3′ end sequencing
Cell type
Single-cell RNA-seq
Computational analysis

ABSTRACT

Alternative polyadenylation (APA) occurs in the process of mRNA maturation by adding a poly(A) tail at different locations, resulting increased diversity of mRNA isoforms and contributing to the complexity of gene regulatory network. Benefit from the development of high-throughput sequencing technologies, we could now delineate APA profiles of transcriptomes at an unprecedented pace. Especially the single cell RNA sequencing (scRNA-seq) technologies provide us opportunities to interrogate biological details of diverse and rare cell types. Despite increasing evidence showing that APA is involved in the cell type-specific regulation and function, efficient and specific laboratory methods for capturing poly(A) sites at single cell resolution are underdeveloped to date. In this review, we summarize existing experimental and computational methods for the identification of APA dynamics from diverse single cell types. A future perspective is also provided.

Contents

1. Introduction

In eukaryotic cells, mature mRNA is generated from precursors that undergo multiple molecular processing, including 5′ end capping, splicing, and 3′ end cleavage and polyadenylation [1]. The location of the cleavage on the nascent RNA, followed by the synthesis of a poly(A) tail, is termed polyadenylation site or poly(A) site. It is well known that most protein-coding genes are transcribed into diverse mRNA isoforms through differential usage of different poly(A) sites. This phenomenon is known as alternative polyadenylation (APA) [2]. More than 50% of genes in animals and plants contain ≥2 poly(A) sites [3–5]. Most APAs occur in 3′

* Correspondence author at: Graduate College of Biomedical Sciences, Western University of Health Sciences, 309 E. 2nd Street, Pomona, CA 91769, USA.
E-mail addresses: yec@xmu.edu.cn (C. Ye), qqli@westernu.edu (Q.Q. Li).

https://doi.org/10.1016/j.csbj.2020.04.009

untranslated regions (UTR-APA), resulting in an inclusion or exclusion of specific sequences in 3′ UTRs. The 3′ UTRs tend to harbor many *cis*-elements, including poly(A) signals, mRNA stability regulatory elements, RNA export and localization signals, as well as miRNA targets [6,7]. Thus, the dynamic usages of UTR-APA could influence mRNA stability, translational efficiency, nuclear export and cytoplasmic localization [6,7]. In some cases, 3′ UTRs are involved in differentiation of protein functionalities through UTR-APA-mediated protein–protein interactions [8–10]. Moreover, some APAs are also found in 5′ UTR, intron and protein coding regions (CR-APA), resulting in truncated proteins or proteins with altered functions [2,6,11]. Inhibition of non-canonical poly(A) sites could protect mRNA integrity and globally regulate APA profiles [12,13], and inactivation of tumor-suppressor-genes were widely observed in tumors through intronic polyadenylation [14].

APA is regulated by the polyadenylation machinery and associated proteins [15–18], and is increasingly found to be tissue- and cell type-specific [19–25]. Through single-molecule sequencing, abundant tissue- and species-specific poly(A) sites were observed in maize and sorghum [19]. While APA in 3′ UTR of Pax3 was proved to control the fate of muscle stem cell and muscle function under homeostatic conditions [20], some differential APA events were found having less impact on muscle-specific expression profiles comparing to noncoding RNAs [21]. Differential APA responses to Cadmium toxicity were monitored across different root cell types in Arabidopsis [22]. The mouse brain cells showed substantial 3′ UTR dynamics across diverse cell types, especially, neurons globally preferred using distal, while microglia exhibited preferences of proximal, poly(A) sites [5,23,24]. Cell type specificities of APA profiles discovered from multiple tumor types show promising prognostic potential in certain cancers and cardiac diseases [26–28]. However, our understandings of APA regulation and function in different cell types is far from clear [2,7]. Interrogation of APA profiles from different cells is the first step towards revealing such a mechanism.

To profile the cell type-specific APA or differentially expressed APA among diverse cell types, several kinds of wet lab approaches have been developed, and could be classified into three categories: (1) Sort and purify cells into different cell types before library construction and followed by a normal 3′ end sequencing [22,29]; (2) Extract poly(A) profile of specific cell types from intact tissues based on co-immunoprecipitation (co-IP) of mRNA 3′ ends [23,24]; (3) Integration of conventional bulk 3′ end sequencing and single cell RNA-seq protocols to interrogate the cell types and poly(A) profile simultaneously [30]. Meanwhile, in order to take full advantage of massive volume of existing single cell RNA-seq (scRNA-seq) data, most of which adopted a 3′ end-enriched sequencing strategy and were mainly used to delineate the cell heterogeneity based on gene expression, several computational tools (scDAPA [31], scAPA [32], and Sierra [33]) were recently developed to decipher the cell heterogeneity at the APA level and demonstrated promising efficacy.

In this review, we summarize these experimental methods detecting APAs from individual cell types, and computational programs delineating APA sites and dynamics from 3′ end-enriched scRNA-seq data.

2. Experimental methods for cell type-specific APA identification

To comprehensively and precisely obtain APA profiles of different cell types, the best way is designing appropriately experimental approaches to differentiate cell populations and capture mRNA fragments with poly(A) tails directly, followed by library construction and high-throughput sequencing. The currently available experimental methods devoting in cell type APA identification could be classified into the following three categories as shown in Fig. 1.

2.1. Cell sorting methods

These methods first dissociate complex tissues into individual cells which is then sorted into pure cell populations before performing 3′ end sequencing [22,29]. Tissues mincing could be finished by mechanical disaggregation (scalpels or tweezers) with or without enzymatic dissociation by collagenase (Col II, Col IV, Col V, or Col XI) plus DNase [34]. For plants, it requires chopping of tissues into small pieces, followed by a cellulase cocktail digestion [22]. After that, individual cells could be sorted into pure cell populations using fluorescence-activated cell sorting (FACS) [35,36], micromanipulation (single cells are individually aspirated from a cell population under a microscope using glass micropipettes) [37], optical tweezers (integration of imaging based cell selection and laser beam to manipulate single cells) [38], or microfluidics [39] etc. Total mRNAs of individual cell types are then extracted for library construction and sequencing using a NGS platform, *e.g.* Illumina or Ion Torrent.

Based on the adopted library construction strategies, 3′ end sequencing protocols could be classified into two categories: oligo(dT) primer-based protocols and RNA manipulation-based protocols. Oligo(dT) primer-based protocols mainly utilize oligo (dT) primer to directly capture poly(A) fragments of mRNA molecules, followed by reverse transcription to produce cDNA for sequencing, such as 3′-seq [29], A-seq [40], PAS-seq [41], and PAT-seq [42] (Fig. 1A). These protocols are easy to implement and time saving, but typically accompany with internal priming (annealing to internal adenine-rich sequences), sequencing desynchronization, and bias arising from enzymatic digestions. To overcome the internal priming problem, several RNA manipulation-based protocols were developed by incorporating multiple RNA manipulation steps. For example, 3P-seq adopts a splint adaptor to ligate complete mRNAs, and RNase T1/H digestions to obtain target fragments for sequencing [43]; 3′READS employs chimeric U5 and T45 and a stringent primer washing in library construction [44]. However, they are generally time consuming, involving multiple steps of RNA manipulation, and prone to bias of the cleavage efficiency of the enzymes.

Limitations of cell sorting methods include: (1) potential introduction of unexpected cellular stress during tissue dissociation, cell type purification and laser illumination, thus potentially alter the gene expression and polyadenylation profiles; (2) the need of a high volume sample; (3) not applicable to tissues hard to be dissociated or containing rare cells that are not efficiently collected during sorting.

2.2. Crosslinking immunoprecipitation and GFP tagging methods

To address the impact of manipulation and/or cell sorting on gene expression, Hwang et al., proposed cTag-PAPERCLIP based on crosslinking immunoprecipitation (CLIP) and green fluorescent protein (GFP) tagging to extract poly(A) profile of pure cell populations from intact tissues [23,24]. Taking advantage that poly(A)-binding protein (PABP) has high affinity to poly(A) tails (not internal adenine-rich sequences) [45], PABP was used to pull down the poly(A) tail contained mRNA fragments before performing reverse transcription reactions with oligo(dT) [46] (Fig. 1B). This approach also eliminated potential internal priming issues resulting from oligo(dT) primers annealing to internal adenine-rich sequences in mRNAs. In order to profile specific cell type in mixed tissues, mouse models conditionally and selectively expressing GFP-

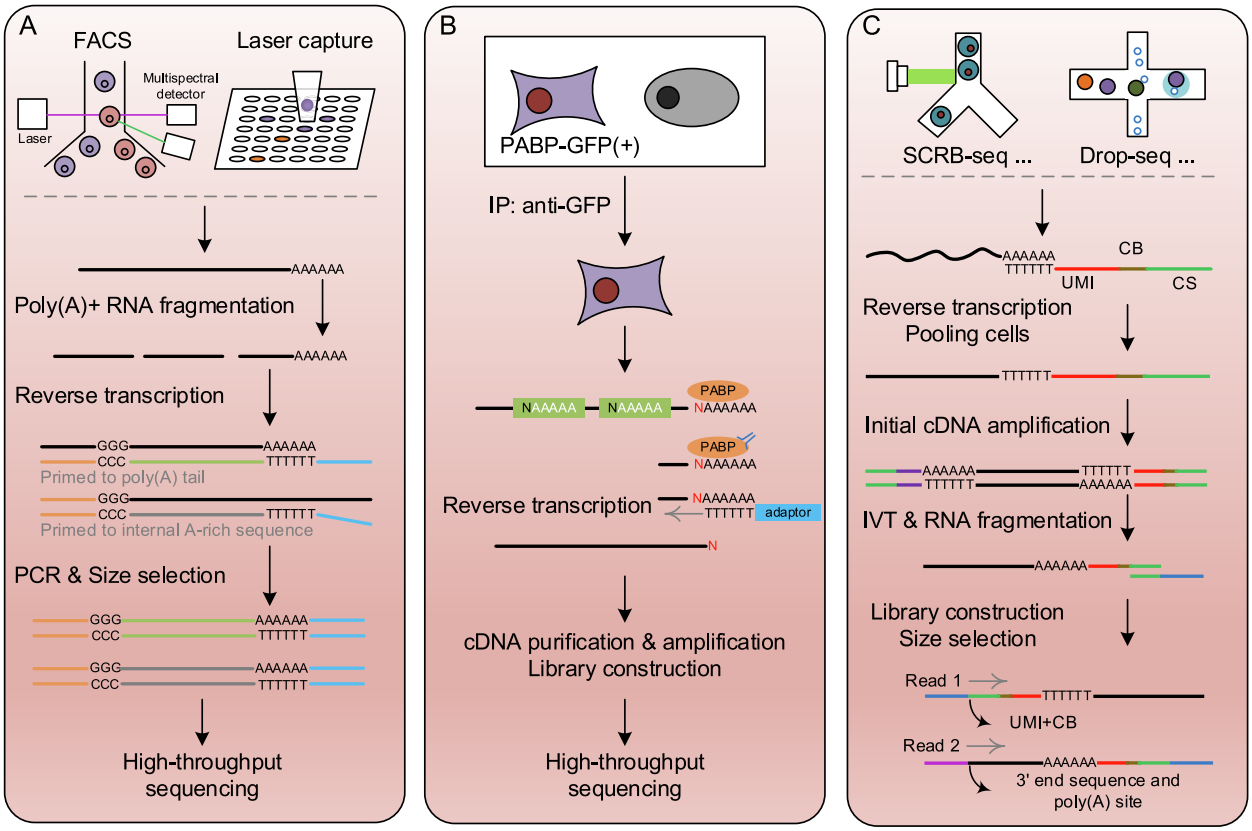

Fig. 1. Experimental protocols for profiling APA in cell type-specific manner. (A) A cell sorting method using PAT-seq; (B) The cTag-PAPERCLIP using crosslinking immunoprecipitation and GFP tagging; (C) The BAT-seq using cellular and molecular barcodes. See more details in the text.

tagged poly(A)-binding proteins were generated using the Cre/Lox technology [47] (Fig. 1B).

The technical advantages of cTag-PAPERCLIP include less extracellular stress in tissue dissociation compared to manual manipulation and FACS, and the elimination of potential internal priming events without *in silico* filtering. Limitations include that cTag-PAPERCLIP cannot handle with unknown or rare cell populations, and the organism must express tagged-PABP in a cell-type specific manner. Since PABP plays multiple key roles in regulating mRNA localization, turnover and translation, it is unclear how GFP-tag may negatively impact the folding structures and many essential functions of PABP, thus altering cellular physiology and state [48]. Additionally, some RNA-binding proteins are aggregation-prone (like PABPN1), overexpression may result in aggregation and depletion of themselves (hence bias for APA), and the aggregation is also cell-type specific [17]. Moreover, it is time consuming for plants and/or those organisms without an efficient transformation protocol. Similar to cell sorting methods, it is also not applicable to cases where sample amounts are limited.

2.3. Cellular and molecular barcoding methods

The rapid growth of single cell RNA-seq (scRNA-seq) technology provides us unprecedented opportunity to delineate the cell state and diversity, and changes our understanding of organisms and organs by tracing the widespread cell heterogeneity [49,50]. It's conceivable to incorporate the scRNA-seq protocol into conventional 3′ end sequencing protocols. Velten et al. proposed a method BAT-seq [30] by combining a bulk 3′ end sequencing protocol TIF-seq [51] and a cellular and molecular barcode-based scRNA-seq protocol [52]. In these protocols, cellular barcodes were introduced to differentiate cell identities of mRNA molecules, and unique molecular identifiers (UMI) were added for absolute quantification of gene expression [52,53]. Cellular and UMI barcoding could be accomplished by reverse transcription through plate-based (STRT, CEL-seq2) or droplet-based protocols (Drop-seq, MARS-seq) [54,55]. Reverse transcription took place inside each droplet/well with a single cell, after which cells were pooled in bulk for initial amplification of cDNAs. Fragments of mRNA with poly(A) tails were captured using a biotinylated tag, followed by 3′ end library construction and Illumina paired-end sequencing (Fig. 1C).

Undoubtedly, the marriage of scRNA-seq and 3′ end sequencing is the ideal way to delineate APA dynamics from diverse cell types. Unknown and rare cells or sub-type cell populations could be detected based on gene expression profiles and cell-type level gene markers. Furthermore, RNA molecules and usages of distinct poly (A) sites could be precisely quantified by the embedded cellular and UMI barcodes, eliminating potential biases introduced by unbalanced PCR amplification efficiencies and/or diverse mRNA fragment lengths [52,53]. However, the single cell-based method BAT-seq encompasses several limitations, including a low sensitivity (limited detection of ~5% of RNA molecules), complicated manipulation and computation steps [6].

3. Computational methods for cell type-specific APA identification

Although a serial of experimental methods were developed to directly profile the APA of multiple species and tissues in the past decade, the availability of 3′ end sequencing data and their coverage of different cell types are still limited and incomparable to the conventional bulk RNA-seq data [6,56]. To remedy the lag of APA

Table 1
Characteristics of currently available computational methods for detecting APA dynamics from scRNA-seq data.

	Strategy	Quantification method	Statistical method	scRNA-seq data type
scAPA	Peak calling-based	$\log_2\left(\frac{C_1+1}{(C+1)}\right)$	Chi-squared test	3′ end
Sierra	Peak calling-based	$\log_2\left(\frac{E\times(p_{11}\cdots p_{n1})}{(g_1+1)}+1\right)$	Wilcoxon rank-sum test	3′ end
scDAPA	Density distribution-based	$\sum_{n=1}^{N}\frac{\lvert p_n^A-p_n^B\rvert}{2}$	Wilcoxon rank-sum test	3′ end, full-length

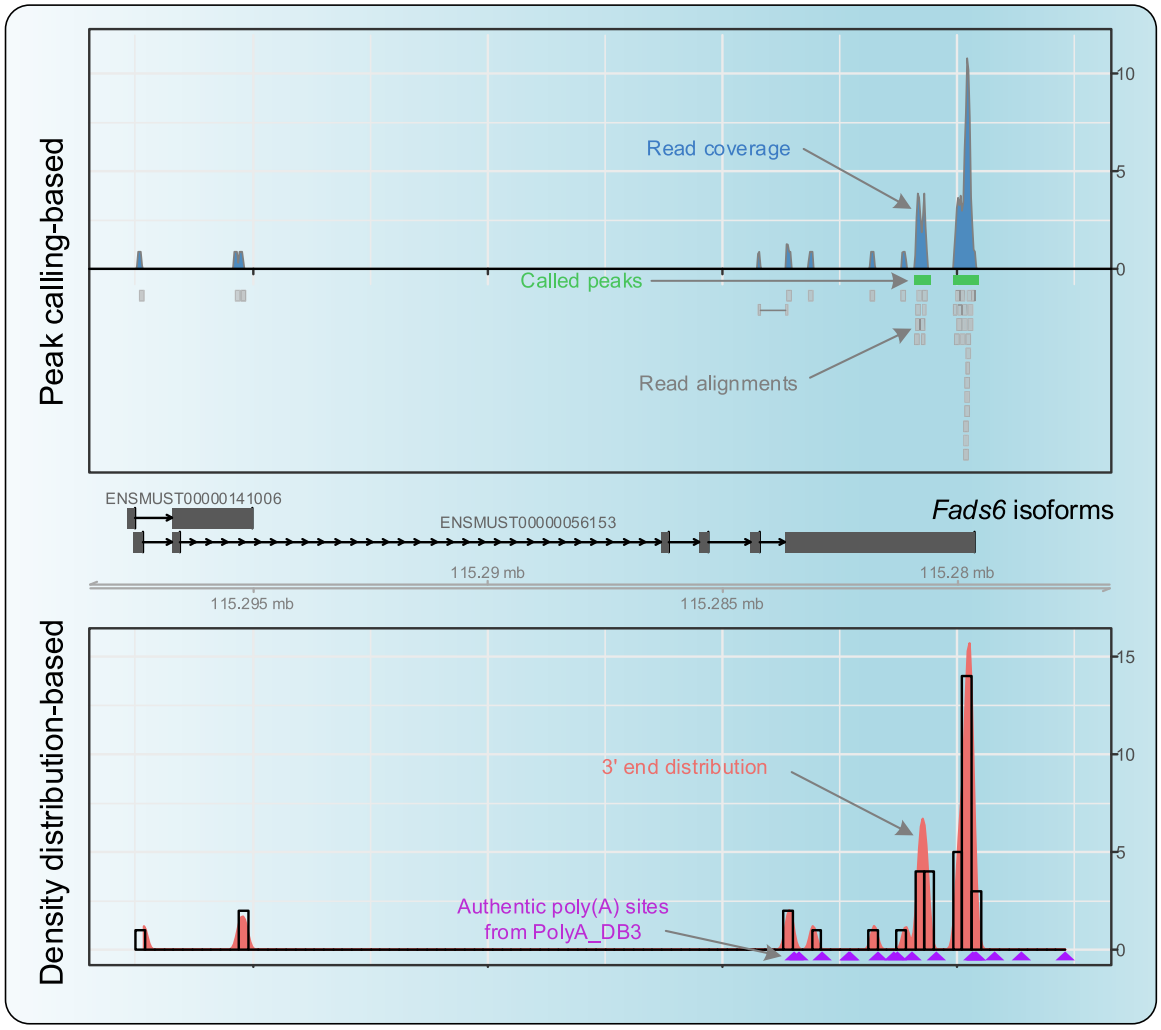

Fig. 2. Illustration of the peak calling-based and density distribution-based methods in APA dynamics identification using 3′ enriched scRNA-seq data. Low coverage peaks are missed by peak calling-based methods, and overlapping peaks resulted from usage of adjacent poly(A) sites cannot be separated (top panel); usage of adjacent poly(A) sites are divided into separate bins by density distribution-based methods, while concrete number of poly(A) sites cannot be determined (bottom panel).

studies and take advantage of the massive volume of existing RNA-seq data, a number of computational methods have been developed to identify and quantify APA from conventional bulk RNA-seq data (reviewed in [56,57]). In terms of sequencing strategies, reads from conventional bulk RNA-seq data would be distributed evenly across individual transcripts, fluctuations of read coverages resulted from diverse isoforms (including APA) could be detected by specific computational approaches, such as DaPars [27], APA-trap[58], ChangePoint [59], IntMAP [60], 3USS [61], and TECtool [62].

Similar to the bulk RNA-seq, massive scRNA-seq data are continuously produced in an unprecedented speed and coverage. Since efficacious single cell-based 3′ end sequencing protocol is still lacking, in order to explore the APA dynamics between different cell types or biological conditions, several computational methods were proposed using scRNA-seq data. According to their adopted quantification strategies, these methods could be clustered into two categories: peak calling-based methods and density distribution-based methods (Table 1).

3.1. Peak calling-based methods

Unlike conventional bulk RNA-seq, most of scRNA-seq methods introduces cellular and molecular barcodes into library constructions, reads from a specific cell and molecule will be uniquely barcoded and be unbiasedly quantified [54,55]. Thus, typically only

the 3′ end or 5′ end most fragments will be sequenced in scRNA-seq protocols, yielding sequenced reads that are enriched in locations around the transcription start or polyadenylation sites of individual isoforms in high probability. Taking it into account, scAPA [32] and Sierra [33] were proposed recently to quantify the cell type-specific APA regulation from 3′ enriched scRNA-seq data by a peak calling strategy (Table 1, Fig. 2).

These methods generally take the aligned-BAM file and the cell cluster annotation file as inputs. Aligned BAM file and cell cluster annotation are obtained through routine scRNA-seq analysis pipelines and tools (*e.g.* 10× Cell Ranger [53], STAR [63], Seurat [64], and SCran [65]). Key steps of these methods include:

(I) Divide the mixed BAM files into individual BAM files of reads from a single cell type. Assignment and deduplication of reads could be done according to their attached cellular and UMI barcodes [66], and cell type classification results obtained from analyzing the homogeneity of gene expression profiles of cells [64,65].
(II) Identify alignment peaks from aligned-BAM files. Since the result of read alignments of scRNA-seq data is similar to that of ChIP-seq data, the enriched and accumulated alignments are shown as lots of individual peaks (Fig. 2). Generally, an individual peak presents a preference of a specific poly(A) site of the gene, multiple peaks of a gene indicates the occurrence of APA in the corresponding gene. To conduct the peak identification, scAPA utilizes Homer *fndPeaks* [67], a popular peak calling tool of ChIP-seq data analysis, to locate the enriched regions. On the other hand, Sierra employs a splice-aware peak caller to overcome the potential spliced-peaks caused by introns [33].

After peak identification, scAPA utilizes the BEDTools *merge* [68] to combine overlapping peaks, followed by a Gaussian finite mixture modeling [69] to separate potential adjacent peaks caused by the fuzzy distribution of aligned-reads. In contrast to scAPA, Sierra takes a similarity score, considering both the distance and lengths of two peaks, to determine whether the two peaks should be merged or not.

(III) Annotate and quantify individual peaks. All peaks called by scAPA are annotated with a predefined 3′ UTR extracted from human and mouse genomes. Numbers of unique reads located in annotated peaks of each cell type are then counted by *featureCounts* of the Rsubread package [70]. While Sierra takes the GTF annotation file and genomic coordinates of peaks as inputs to annotate the genomic features of identified peaks, and quantify the UMI number of each peak using the GenomicAlignments package [71].
(IV) Detect APA dynamics from cell type-specific APA profiles. Several distinct approaches were proposed in detecting the differential usage of APA in genes between samples in cell type-specific manner. In scAPA, the chi-squared test is used to evaluate the significance of difference, and a proximal poly(A) site usage index (*proximal PUI*, Table 1) was proposed to quantify the relative usage of the most proximal poly(A) site within a gene. Sierra employs the Wilcoxon rank-sum test to detect the 3′ UTR lengthening or shortening events, and calculates a relative expression level (*R*, Table 1) for each peak in each cell.

Peak calling method is not sensitive to overlapping peaks resulting from reads of isoforms with adjacent poly(A) sites (Fig. 2) [72,73], only ~5% of poly(A) sites adjacent to each other within 200 bp could be differentiated by scAPA. For distance ranges in 200–300 bp, the recognition rate is ~30% [32]. Reads spreading in a wide range and consequently a low coverage will be missed by peak calling methods in high probability. Moreover, the high dropout rate and low sequencing throughput in single cells of currently available scRNA-seq technologies also hamper the discovery precision and depth in identification of APA dynamics. Thus, not all valid reads could be recovered or accurately split by the peak calling methods, and not all APA events could be discovered from these scRNA-seq data.

3.2. Density distribution-based methods

Through the reads generated by 3′ enriched scRNA-seq enrich in the upstream regions of poly(A) sites, their distributions generally cover a rather wide region and in a fuzzy mode [33,74]. To overcome the fuzzy distribution, scDAPA utilizes a non-parametric density distribution method to quantify the difference between two compared groups, rather than calling the precise locations and boundaries of peaks [31]. The scDAPA also takes aligned-BAM/SAM file and cell cluster annotation file as inputs, and consists of three core modules:

(I) Extract and divide reads by cell type classifications. The mapped reads should be in BAM/SAM format, and reads are well tagged with cell barcodes and UMIs. The scDAPA employs an automatic shell script to deduplicate and extract valid reads based on the 10× Cell Ranger BAM Tag system (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/output/bam) and the default alignment Tag system [75]. Valid reads will be separated into individual SAM files according to the cell type classification result including cell barcodes and their corresponding cell types/groups.
(II) Extract and annotate 3′ ends from aforementioned reads. The coordinates of the 3′ most ends of these retrieved reads are extracted using BEDTools [68] and annotated by genes with a provided GTF file. All the reads overlapping with a genic region (including both exonic and intronic regions) will be reserved, others located in intergenic regions will be discarded.
(III) Identify APA dynamics between two different cell types or the same cell type in different conditions. Instead of assigning 3′ enriched reads into separate peaks, scDAPA divides the genic region into equal-size bins (histograms), calculate the density distribution of each bins by gene and by cell type, and directly quantify the difference using a site distribution difference (*SDD*, Table 1) index (Fig. 2). The significance of difference is measured by the Wilcoxon rank-sum test, and adjusted by the Benjamini Hochberg method if multiple statistical tests were performed [76].

Since the performance of histogram quantification is sensitive to the bin size [77], scDAPA chooses the 100 bp as default bin-size through evaluating a sequential of bin-sizes and inspecting the distances between 3′ ends of scRNA-seq data and authentic poly(A) site annotations [31,74]. Furthermore, the scDAPA is also applicable to scRNA-seq data sequencing full-length of mRNA molecules, in terms of only 3′ most ends are used in density quantification [54]. The limitation of density distribution-based method is that it quantifies relative difference between groups, rather than quantifies the usage of individual poly(A) sites. Besides, it also faces the innate drawback of currently available scRNA-seq technologies, a low gene capturing ratio, and many APA events cannot be detected. Thus, the identified APA events cannot fully represent the APA landscape of the cell. Before better technologies become available, other means to improve data processing (e.g. [79]) could be implemented to reach this goal.

4. Summary and outlook

APA is increasingly recognized as an important regulator in many processes, including mRNA translation, stability, nuclear export and localization [2,7]. Tissue- and cell type- specific regulation of APA were also widely observed and involved in cell activation, proliferation, development and oncogenesis [23,24,56,78], implying a complex role of APA. To decipher and understand the regulatory mechanism of APA, it's necessary to investigate the APA atlas at a cellular resolution.

Here, we summarize the recent development of experimental methods in discovery of APA dynamics in cell type-specific manner, as well as computation method utilizing the scRNA-seq data. The available experimental protocols have largely advanced the cell type-specific study of APA, however, these methods still contain many inherent drawbacks as aforementioned. The cell sorting method may introduce unexpected stress in sample preparation, and affect the cellular states and APA profiles consequently; the crosslinking immunoprecipitation and GFP tagging method requires designing specify transgene model expressing GFP-tagged PABP in specific cell types, that limits its wide application in various cell populations; the cellular and molecular barcoding method could quantify gene expression and APA profiles from single cells by employing a scRNA-seq protocol, allowing discerning rare and unknown cell populations possible. Collectively, the single cell-based 3′ end sequencing method is of the most promising application prospect. However, technical ameliorations of this type of combinatory protocol, such as consideration of a more powerful and stable single cell platform 10× Chromium, to generate an easy-to-implement protocol of high accuracy and robustness are pressing needs.

In the past decade, a massive volume of scRNA-seq data, covering a wide variety of cell types and biological conditions, were generated. Computational approaches mining 3′ end enriched scRNA-seq data expand their application scope, and largely fill the gap in study of cell type-specific APA. Theses computational methods have been successfully applied on several 3′ enriched scRNA-seq data sets, *e.g.*, a global 3′ UTR shortening were observed in activated T cells compared to naïve T cells using scAPA [32], and aberrant APA dynamics were observed in lung cancer and acute myeloid leukemia by scAPA and scDAPA [32,74], respectively. However, the low resolution of 3′ enriched scRNA-seq in depicting poly(A) sites hampers the peak calling-based methods in accurate separation of peaks generated by close poly(A) sites; density distribution-based method cannot quantify usages of individual poly(A) site. The inherent low gene capturing ratio of scRNA-seq protocols also limits the observation depth of these computational methods. To improve the discovery accuracy and reduce the false discovery rate, potential ways include integrating the reference annotation of poly(A) sites, such as PolyA_DB 3 [3] and PlantAPAdb [4], to guide the identification and quantification of poly(A) site usages from scRNA-seq data, and improving the sequencing throughput of single cells and/or the number of cell inputs. All in all, more innovatively and efficaciously computational approaches for estimating cell type-specific APA usage from scRNA-seq data sets are still in high demand.

Declaration of competing interest

The authors report no conflict of interest.

Author contribution statement

Congting Ye: Conceptualization, Visualization, Writing - original draft, Funding acquisition. **Juncheng Lin:** Writing - review & editing. **Qingshun Quinn Li:** Conceptualization, Funding acquisition, Writing - review & editing.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (61802323), and a grant from the National Key R&D Project of China (2016YFE0108800).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.04.009.

References

[1] Desterro J, Bak-Gordon P, Carmo-Fonseca M. Targeting mRNA processing as an anticancer strategy. Nat Rev Drug Discov 2019:1–18. https://doi.org/10.1038/s41573-019-0042-3.
[2] Tian B, Manley JL. Alternative polyadenylation of mRNA precursors. Nat Rev Mol Cell Biol 2017;18:18–30. https://doi.org/10.1038/nrm.2016.116.
[3] Wang R, Nambiar R, Zheng D, Tian B. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. Nucleic Acids Res 2018;46:D315–9. https://doi.org/10.1093/nar/gkx1000.
[4] Zhu S, Ye W, Ye L, Fu H, Ye C, Xiao X, et al. PlantAPAdb: a comprehensive database for alternative polyadenylation sites in plants. Plant Physiol 2020;182:228–42. https://doi.org/10.1104/pp.19.00943.
[5] Guvenek A, Tian B. Analysis of alternative cleavage and polyadenylation in mature and differentiating neurons using RNA-seq data. Quant Biol 2018;6:253–66. https://doi.org/10.1007/s40484-018-0148-3.
[6] Chen W, Jia Q, Song Y, Fu H, Wei G, Ni T. Alternative polyadenylation: methods, findings, and impacts. Genomics Proteomics Bioinformatics 2017;15:287–300. https://doi.org/10.1016/j.gpb.2017.06.001.
[7] Mayr C. What are 3′ UTRs doing?. Cold Spring Harb Perspect Biol 2018;11:. https://doi.org/10.1101/cshperspect.a034728a034728.
[8] Berkovits BD, Mayr C. Alternative 3'UTRs act as scaffolds to regulate membrane protein localization. Nature 2015;522:363–7. https://doi.org/10.1038/nature14321.
[9] Chartron JW, Hunt KCL, Frydman J. Cotranslational signal independent SRP preloading during membrane targeting. Nature 2016;536:224–8. https://doi.org/10.1038/nature19309.
[10] Ma W, Mayr C. A membraneless organelle associated with the endoplasmic reticulum enables 3′utr-mediated protein-protein interactions. Cell 2018;175:1492–506. https://doi.org/10.1016/j.cell.2018.10.007.
[11] Guo C, Spinelli M, Liu M, Li QQ, Liang C. A genome-wide study of "non-3UTR" polyadenylation sites in Arabidopsis thaliana. Sci Rep 2016;6:1–10. https://doi.org/10.1038/srep28060.
[12] Shi J, Deng Y, Huang S, Huang C, Wang J, Xiang AP, et al. Suboptimal RNA–RNA interaction limits U1 snRNP inhibition of canonical mRNA 3' processing. RNA Biol 2019;16(10):1448–60. https://doi.org/10.1080/15476286.2019.1636596.
[13] Wang R, Zheng D, Wei L, Ding Q, Tian B. Regulation of intronic polyadenylation by PCF11 impacts mRNA expression of long genes. Cell Rep 2019;26:2766–78. https://doi.org/10.1016/j.celrep.2019.02.049.
[14] Lee S-H, Singh I, Tisdale S, Abdel-Wahab O, Leslie CS, Mayr C. Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. Nature 2018;561:127–31. https://doi.org/10.1038/s41586-018-0465-8.
[15] Elkon R, Ugalde AP, Agami R. Alternative cleavage and polyadenylation: extent, regulation and function. Nat Rev Genet 2013;14:496–506. https://doi.org/10.1038/nrg3482.
[16] de Klerk E, Venema A, Anvar SY, Goeman JJ, Hu O, Trollet C, et al. Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation. Nucleic Acids Res 2012;40:9089–101. https://doi.org/10.1093/nar/gks655.
[17] Jenal M, Elkon R, Loayza-Puch F, van Haaften G, Kühn U, Menzies FM, et al. The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. Cell 2012;149:538–53. https://doi.org/10.1016/j.cell.2012.03.022.
[18] Abbassi-Daloii T, Yousefi S, de Klerk E, Grossouw L, Riaz M, Hoen PAC, et al. An alanine expanded PABPN1 causes increased utilization of intronic polyadenylation sites. NPJ Aging Mech Dis 2017;3:1–8. https://doi.org/10.1038/s41514-017-0007-x.
[19] Wang B, Regulski M, Tseng E, Olson A, Goodwin S, McCombie WR, et al. A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing. Genome Res 2018;28:921–32. https://doi.org/10.1101/gr.227462.117.
[20] de Morree A, Klein JDD, Gan Q, Farup J, Urtasun A, Kanugovi A, et al. Alternative polyadenylation of Pax3 controls muscle stem cell fate and muscle function. Science 2019;366:734–8. https://doi.org/10.1126/science.aax1694.
[21] Raz V, Riaz M, Tatum Z, Kielbasa SM, Hoen PAC't. The distinct transcriptomes of slow and fast adult muscles are delineated by noncoding RNAs. FASEB J 2017;32:1579–90. https://doi.org/10.1096/fj.201700861R.

[22] Cao J, Ye C, Hao G, Dabney-Smith C, Hunt AG, Li QQ. Root hair single cell type specific profiles of gene expression and alternative polyadenylation under cadmium stress. Front Plant Sci 2019;10:589. https://doi.org/10.3389/fpls.2019.00589.

[23] Hwang HW, Saito Y, Park CY, Blachère NE, Tajima Y, Fak JJ, et al. cTag-PAPERCLIP reveals alternative polyadenylation promotes cell-type specific protein diversity and shifts Araf isoforms with microglia activation. Neuron 2017;95:1334–49. https://doi.org/10.1016/j.neuron.2017.08.024.

[24] Jereb S, Hwang HW, Van Otterloo E, Govek EE, Fak JJ, Yuan Y, et al. Differential 3' processing of specific transcripts expands regulatory and protein diversity across neuronal cell types. ELife 2018;7:. https://doi.org/10.7554/eLife.34042e34042.

[25] Kim N, Chung W, Eum HH, Lee HO, Park WY. Alternative polyadenylation of single cells delineates cell types and serves as a prognostic marker in early stage breast cancer. PLoS ONE 2019;14:. https://doi.org/10.1371/journal.pone.0217196e0217196.

[26] Singh P, Alley TL, Wright SM, Kamdar S, Schott W, Wilpan RY, et al. Global changes in processing of mRNA 3′ untranslated regions characterize clinically distinct cancer subtypes. Cancer Res 2009;69:9422–30. https://doi.org/10.1158/0008-5472.CAN-09-2236.

[27] Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. Nat Commun 2014;5:5274. https://doi.org/10.1038/ncomms6274.

[28] Creemers EE, Amira B, Ugalde AP, van Deutekom HWM, van der Made I, de Groot Nina E, et al. Genome-wide polyadenylation maps reveal dynamic mRNA 3′-end formation in the failing human heart. Circ Res 2016;118:433–8. https://doi.org/10.1161/CIRCRESAHA.115.307082.

[29] Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. Genes Dev 2013;27:2380–96. https://doi.org/10.1101/gad.229328.113.

[30] Velten L, Anders S, Pekowska A, Järvelin AI, Huber W, Pelechano V, et al. Single-cell polyadenylation site mapping reveals 3′ isoform choice variability. Mol Syst Biol 2015;11:812. https://doi.org/10.15252/msb.20156198.

[31] Ye C, Zhou Q, Wu X, Yu C, Ji G, Saban DR, et al. scDAPA: detection and visualization of dynamic alternative polyadenylation from single cell RNA-seq data. Bioinformatics 2020;36:1262–4. https://doi.org/10.1093/bioinformatics/btz701.

[32] Shulman ED, Elkon R. Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. Nucleic Acids Res 2019;47(19):10027–39. https://doi.org/10.1093/nar/gkz781.

[33] Patrick R, Humphreys DT, Janbandhu V, Oshlack A, Ho JWK, Harvey RP, et al. Discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. BioRxiv 2020:1–47. https://doi.org/10.1101/867309.

[34] Leelatian N, Doxie DB, Greenplate AR, Mobley BC, Lehman JM, Sinnaeve J, et al. Single cell analysis of human tissues and solid tumors with mass cytometry. Cytometry B Clin Cytom 2017;92:68–78. https://doi.org/10.1002/cyto.b.21481.

[35] Chattopadhyay PK, Price DA, Harper TF, Betts MR, Yu J, Gostick E, et al. Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. Nat Med 2006;12:972–7. https://doi.org/10.1038/nm1371.

[36] Chattopadhyay PK, Roederer M. Cytometry: Today's technology and tomorrow's horizons. Methods 2012;57:251–8. https://doi.org/10.1016/j.ymeth.2012.02.009.

[37] Citri A, Pang ZP, Südhof TC, Wernig M, Malenka RC. Comprehensive qPCR profiling of gene expression in single neuronal cells. Nat Protoc 2012;7:118–27. https://doi.org/10.1038/nprot.2011.430.

[38] Yoshimoto N, Kida A, Jie X, Kurokawa M, Iijima M, Niimi T, et al. An automated system for high-throughput single cell-based breeding. Sci Rep 2013;3:1–9. https://doi.org/10.1038/srep01191.

[39] Sackmann EK, Fulton AL, Beebe DJ. The present and future role of microfluidics in biomedical research. Nature 2014;507:181–9. https://doi.org/10.1038/nature13118.

[40] Martin G, Gruber AR, Keller W, Zavolan M. Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′ UTR length. Cell Rep 2012;1:753–63. https://doi.org/10.1016/j.celrep.2012.05.003.

[41] Shepard PJ, Choi E-A, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA 2011;17:761–72. https://doi.org/10.1261/rna.2581711.

[42] Wu X, Liu M, Downie B, Liang C, Ji G, Li QQ, et al. Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. Proc Natl Acad Sci 2011;108:12533–8. https://doi.org/10.1073/pnas.1019732108.

[43] Jan CH, Friedman RC, Ruby JG, Bartel DP. Formation, regulation and evolution of Caenorhabditis elegans 3′UTRs. Nature 2011;469:97–101. https://doi.org/10.1038/nature09616.

[44] Hoque M, Ji Z, Zheng D, Luo W, Li W, You B, et al. Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. Nat Methods 2013;10:133. https://doi.org/10.1038/nmeth.2288.

[45] Kahvejian A, Roy G, Sonenberg N. The mRNA closed-loop model: the function of PABP and PABP-interacting proteins in mRNA translation. Cold Spring Harb Symp Quant Biol 2001;66:293–300. https://doi.org/10.1101/sqb.2001.66.293.

[46] Hwang H-W, Park CY, Goodarzi H, Fak JJ, Mele A, Moore MJ, et al. PAPERCLIP identifies MicroRNA targets and a role of CstF64/64tau in promoting non-canonical poly(A) site usage. Cell Rep 2016;15:423–35. https://doi.org/10.1016/j.celrep.2016.03.023.

[47] Tsien JZ. Cre-Lox neurogenetics: 20 years of versatile applications in brain research and counting…. Front Genet 2016;7:. https://doi.org/10.3389/fgene.2016.00019 19.

[48] Weill U, Krieger G, Avihou Z, Milo R, Schuldiner M, Davidi D. Assessment of GFP tag position on protein localization and growth fitness in yeast. J Mol Biol 2019;431:636–41. https://doi.org/10.1016/j.jmb.2018.12.004.

[49] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell 2019;177:1888–902. https://doi.org/10.1016/j.cell.2019.05.031.

[50] Adey AC. Integration of single-cell genomics datasets. Cell 2019;177:1677–9. https://doi.org/10.1016/j.cell.2019.05.034.

[51] Pelechano V, Wei W, Jakob P, Steinmetz LM. Genome-wide identification of transcript start and end sites by transcript isoform sequencing. Nat Protoc 2014;9:1740–59. https://doi.org/10.1038/nprot.2014.121.

[52] Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods 2014;11:163–6. https://doi.org/10.1038/nmeth.2772.

[53] Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8:14049. https://doi.org/10.1038/ncomms14049.

[54] Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative analysis of single-cell RNA sequencing methods. Mol Cell 2017;65:631–43. https://doi.org/10.1016/j.molcel.2017.01.023.

[55] Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. Nat Commun 2019;10:1–11. https://doi.org/10.1038/s41467-019-12266-7.

[56] Gruber AJ, Zavolan M. Alternative cleavage and polyadenylation in health and disease. Nat Rev Genet 2019;20:599–614. https://doi.org/10.1038/s41576-019-0145-z.

[57] Chen M, Ji G, Fu H, Lin Q, Ye C, Ye W, et al. A survey on identification and quantification of alternative polyadenylation sites from RNA-seq data. Brief Bioinform 2019. https://doi.org/10.1093/bib/bbz068.

[58] Ye C, Long Y, Ji G, Li QQ, Wu X. APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. Bioinformatics 2018;34:1841–9. https://doi.org/10.1093/bioinformatics/bty029.

[59] Wang W, Wei Z, Li H. A change-point model for identifying 3′UTR switching by next-generation RNA sequencing. Bioinformatics 2014;30:2162–70. https://doi.org/10.1093/bioinformatics/btu189.

[60] Chang JW, Zhang W, Yeh HS, Park M, Yao C, Shi Y, et al. An integrative model for alternative polyadenylation, IntMAP, delineates mTOR-modulated endoplasmic reticulum stress response. Nucleic Acids Res 2018;46:5996–6008. https://doi.org/10.1093/nar/gky340.

[61] Le Pera L, Mazzapioda M, Tramontano A. 3USS: a web server for detecting alternative 3′UTRs from RNA-seq experiments. Bioinformatics 2015;31:1845–7. https://doi.org/10.1093/bioinformatics/btv035.

[62] Gruber AJ, Gypas F, Riba A, Schmidt R, Zavolan M. Terminal exon characterization with TECtool reveals an abundance of cell-specific isoforms. Nat Methods 2018;15:832–6. https://doi.org/10.1038/s41592-018-0114-z.

[63] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21. https://doi.org/10.1093/bioinformatics/bts635.

[64] Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 2015;33:495–502. https://doi.org/10.1038/nbt.3192.

[65] Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol 2016;17:1–14. https://doi.org/10.1186/s13059-016-0947-7.

[66] Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res 2017;27:491–9. https://doi.org/10.1101/gr.209601.116.

[67] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 2010;38:576–89. https://doi.org/10.1016/j.molcel.2010.05.004.

[68] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.

[69] Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. R J 2016;8:289–317. https://doi.org/10.32614/RJ-2016-021.

[70] Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res 2019;47:1–9. https://doi.org/10.1093/nar/gkz114.

[71] Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol 2013;9:. https://doi.org/10.1371/journal.pcbi.1003118e1003118.

[72] Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation bbw023. Brief Bioinform 2017;18(2):279–90. https://doi.org/10.1093/bib/bbw023.

[73] Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis bbv110. Brief Bioinform 2016;17(6):953–66. https://doi.org/10.1093/bib/bbv110.

[74] Ye C, Zhou Q, Hong Y, Li QQ. Role of alternative polyadenylation dynamics in acute myeloid leukaemia at single-cell resolution. RNA Biol 2019;16:785–97. https://doi.org/10.1080/15476286.2019.1586139.
[75] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9. https://doi.org/10.1093/bioinformatics/btp352.
[76] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 1995;57:289–300.
[77] Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics 2008;24:2537–8. https://doi.org/10.1093/bioinformatics/btn480.
[78] Yuan F, Hankey W, Wagner EJ, Li W, Wang Q. Alternative polyadenylation of mRNA and its role in cancer. Genes Dis 2019. https://doi.org/10.1016/j.gendis.2019.10.011.
[79] Ye P, Ye W, Ye C, Li S, Ye L, Ji G, et al. scHinter: imputing dropout events for single-cell RNA-seq data with limited sample size. Bioinformatics 2020;36:789–97. https://doi.org/10.1093/bioinformatics/btz627.