



Single-Cell RNA-Seq Technologies and Related Computational Data Analysis

Geng Chen^{1*}, Baitang Ning² and Tielu Shi^{1*}

¹ Center for Bioinformatics and Computational Biology, and Shanghai Key Laboratory of Regulatory Biology, Institute of Biomedical Sciences, School of Life Sciences, East China Normal University, Shanghai, China, ² National Center for Toxicological Research, United States Food and Drug Administration, Jefferson, AR, United States

OPEN ACCESS

Edited by:

Filippo Geraci,
Italian National Research Council
(CNR), Italy

Reviewed by:

Vsevolod Jurievich Makeev,
Vavilov Institute of General Genetics
(RAS), Russia
Iros Barozzi,
Imperial College London,
United Kingdom

*Correspondence:

Geng Chen
gchen@bio.ecnu.edu.cn
Tielu Shi
tieliushi@yahoo.com

Specialty section:

This article was submitted to
Bioinformatics and Computational
Biology,
a section of the journal
Frontiers in Genetics

Received: 05 December 2018

Accepted: 21 March 2019

Published: 05 April 2019

Citation:

Chen G, Ning B and Shi T (2019)
Single-Cell RNA-Seq Technologies
and Related Computational Data
Analysis. *Front. Genet.* 10:317.
doi: 10.3389/fgene.2019.00317

Single-cell RNA sequencing (scRNA-seq) technologies allow the dissection of gene expression at single-cell resolution, which greatly revolutionizes transcriptomic studies. A number of scRNA-seq protocols have been developed, and these methods possess their unique features with distinct advantages and disadvantages. Due to technical limitations and biological factors, scRNA-seq data are noisier and more complex than bulk RNA-seq data. The high variability of scRNA-seq data raises computational challenges in data analysis. Although an increasing number of bioinformatics methods are proposed for analyzing and interpreting scRNA-seq data, novel algorithms are required to ensure the accuracy and reproducibility of results. In this review, we provide an overview of currently available single-cell isolation protocols and scRNA-seq technologies, and discuss the methods for diverse scRNA-seq data analyses including quality control, read mapping, gene expression quantification, batch effect correction, normalization, imputation, dimensionality reduction, feature selection, cell clustering, trajectory inference, differential expression calling, alternative splicing, allelic expression, and gene regulatory network reconstruction. Further, we outline the prospective development and applications of scRNA-seq technologies.

Keywords: single-cell RNA-seq, cell clustering, cell trajectory, alternative splicing, allelic expression

INTRODUCTION

Bulk RNA-seq technologies have been widely used to study gene expression patterns at population level in the past decade. The advent of single-cell RNA sequencing (scRNA-seq) provides unprecedented opportunities for exploring gene expression profile at the single-cell level. Currently, scRNA-seq has become a favorable choice for studying the key biological questions of cell heterogeneity and the development of early embryos (only include a few number of cells), since bulk RNA-seq mainly reflects the averaged gene expression across thousands of cells. In recent years, scRNA-seq has been applied to various species, especially to diverse human tissues (including normal and cancer), and these studies revealed meaningful cell-to-cell gene expression variability (Jaitin et al., 2014; Grun et al., 2015; Chen et al., 2016a; Cao et al., 2017; Rosenberg et al., 2018). With the innovation of sequencing technologies, some different scRNA-seq protocols have been proposed in the past few years, which largely facilitated the understanding of dynamic gene

expression at single-cell resolution (Kolodziejczyk et al., 2015; Haque et al., 2017; Picelli, 2017; Chen et al., 2018). One of them is the highly efficient strategy LCM-seq (Nichterwitz et al., 2016) which combines laser capture microscopy (LCM) and Smart-seq2 (Picelli et al., 2013) for single-cell transcriptomics without tissue dissociation. Currently available scRNA-seq protocols can be mainly split into two categories based on the captured transcript coverage: (i) full-length transcript sequencing approaches [such as Smart-seq2 (Picelli et al., 2013), MATQ-seq (Sheng et al., 2017) and SUPeR-seq (Fan X. et al., 2015)]; and (ii) 3'-end [e.g., Drop-seq (Macosko et al., 2015), Seq-Well (Gierahn et al., 2017), Chromium (Zheng et al., 2017), and DroNC-seq (Habib et al., 2017)] or 5'-end [such as STRT-seq (Islam et al., 2011, 2012)] transcript sequencing technologies. Each scRNA-seq protocol has its own benefits and drawbacks, resulting in that different scRNA-seq approaches have distinct features and disparate performances (Ziegenhain et al., 2017). In conducting single-cell transcriptomic study, specific scRNA-seq technology may need to be employed in consideration of the balance between research goal and sequencing cost.

Owing to the low amount of starting material, scRNA-seq has limitations of low capture efficiency and high dropouts (Haque et al., 2017). Compared to bulk RNA-seq, scRNA-seq produces noisier and more variable data. The technical noise and biological variation (e.g., stochastic transcription) raise substantial challenges for computational analysis of scRNA-seq data. A variety of tools have been designed to conducting diverse bulk RNA-seq data analyses, but many of those methods cannot be directly applied to scRNA-seq data (Stegle et al., 2015). Except short-read mapping, almost all data analyses (such as differential expression, cell clustering, and gene regulatory network inference) have certain disparities in methods between scRNA-seq and bulk RNA-seq. Due to the high technical noise, quality control (QC) is crucial for identifying and removing the low-quality scRNA-seq data to get robust and reproducible results. Furthermore, some analyses including alternative splicing (AS) detection, allelic expression exploration and RNA-editing identification are not suitable for the 3' or 5'-tag sequencing protocols of scRNA-seq, but these analyses could be applicable to the data generated by whole-transcript scRNA-seq. On the other hand, an increasing number of tools are specially proposed for analyzing scRNA-seq data, and each method has its own pros and cons (Stegle et al., 2015; Bacher and Kendzierski, 2016). Therefore, to effectively handle the high variability of scRNA-seq data, attention should be paid to choosing appropriately analytical approaches.

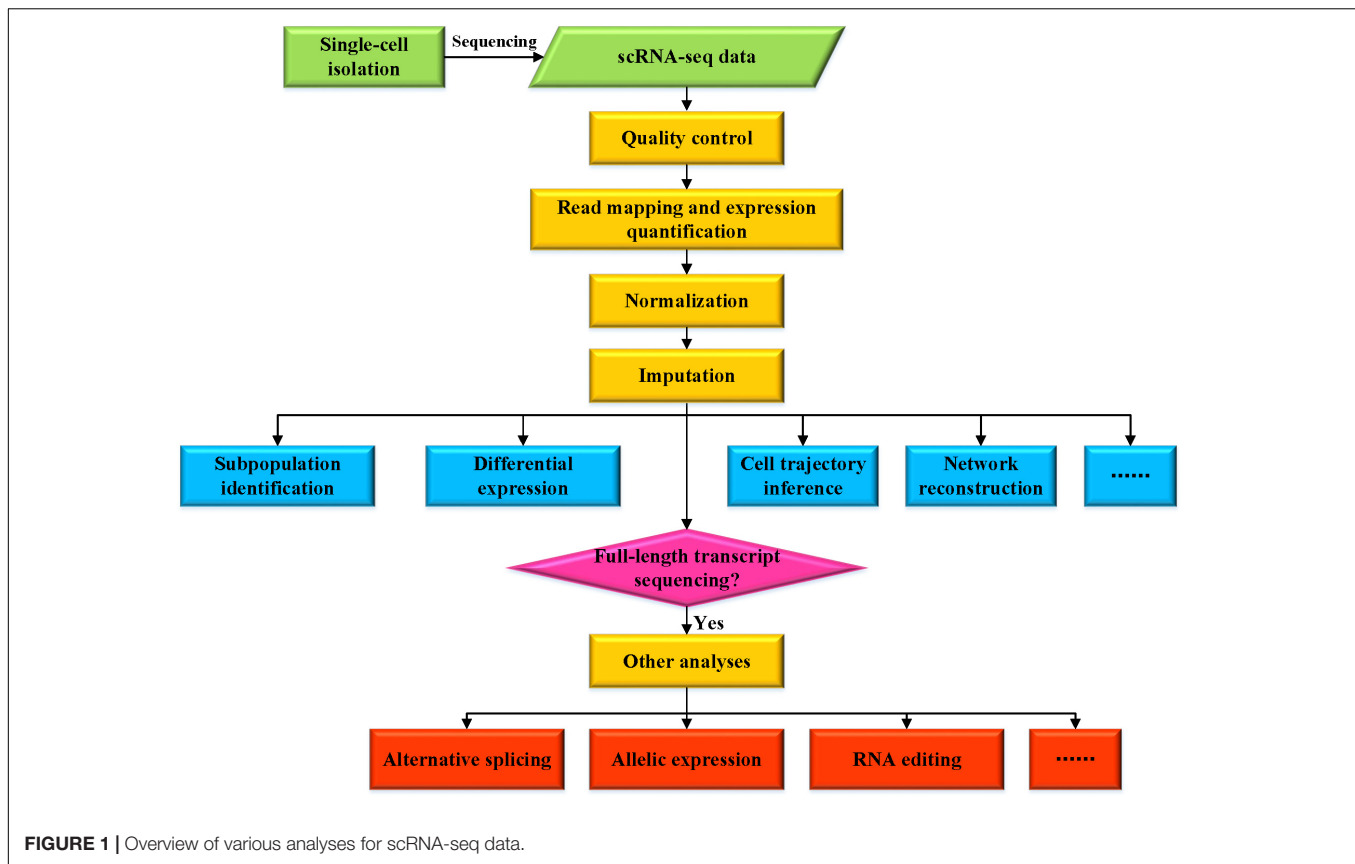
This Review aims to summarize and discuss currently available scRNA-seq technologies and various data analysis methods. We first introduce distinct single-cell isolation protocols and various scRNA-seq technologies developed in recent years. Then we focus on the analyses of scRNA-seq data and highlight the analytical differences between bulk RNA-seq and scRNA-seq data. Considering the high technical noise and complexity of scRNA-seq data, we also provide recommendations on the selection of suitable tools to analyze scRNA-seq data and ensure the reproducibility of results.

ISOLATION OF SINGLE CELLS

The first step of scRNA-seq is isolation of individual cells (**Figure 1**), although the capture efficiency is a big challenge for scRNA-seq. Currently, several different approaches are available for isolating single cells, including limiting dilution, micromanipulation, flow-activated cell sorting (FACS), laser capture microdissection (LCM), and microfluidics (Gross et al., 2015; Kolodziejczyk et al., 2015; Hwang et al., 2018). Limiting dilution technique uses pipettes to isolate cells by dilution, the main limitation of this method is inefficient. Micromanipulation is a classical approach used to retrieve cells from samples with a small number of cells, such as early embryos or uncultivated microorganisms, while this technique is time-consuming and low throughput. FACS has been widely used for isolating single cells, which requires large starting volumes (>10,000 cells) in suspension. LCM is an advanced strategy used for isolating individual cells from solid tissues by using a laser system aided by computer. Microfluidics is increasingly popular due to its property of low sample consumption, precise fluid control and low analysis cost. These single-cell isolation protocols have their own advantages and show distinct performances in terms of capture efficiency and purity of the target cells (Gross et al., 2015; Hu et al., 2016).

CURRENTLY AVAILABLE SCRNA-SEQ TECHNOLOGIES

To date, a number of scRNA-seq technologies have been proposed for single-cell transcriptomic studies (**Table 1**). The first scRNA-seq method was published by Tang et al. (2009), and then many other scRNA-seq approaches were subsequently developed. Those scRNA-seq technologies differ in at least one of the following aspects: (i) cell isolation; (ii) cell lysis; (iii) reverse transcription; (iv) amplification; (v) transcript coverage; (vi) strand specificity; and (vii) UMI (unique molecular identifiers, molecular tags that can be applied to detect and quantify the unique transcripts) availability. One conspicuous difference among these scRNA-seq methods is that some of them can produce full-length (or nearly full-length) transcript sequencing data (e.g., Smart-seq2, SUPeR-seq, and MATQ-seq), whereas others only capture and sequence the 3'-end [such as Drop-seq, Seq-Well and DroNC-seq, SPLiT-seq (Rosenberg et al., 2018)] or 5'-end (e.g., STRT-seq) of the transcripts (**Table 1**). Distinct scRNA-seq protocols may possess disparate strengths and weaknesses, and several published reviews have compared a portion of them in detail (Kolodziejczyk et al., 2015; Haque et al., 2017; Picelli, 2017; Ziegenhain et al., 2017). A previous study demonstrated that Smart-seq2 can detect a bigger number of expressed genes than other scRNA-seq technologies including CEL-seq2 (Hashimshony et al., 2016), MARS-seq (Jaitin et al., 2014), Smart-seq (Ramskold et al., 2012), and Drop-seq protocols (Ziegenhain et al., 2017). Recently, Sheng et al. (2017) showed that another full-length transcript sequencing approach MATQ-seq could outperform Smart-seq2 in detecting low-abundance genes.



Compared to 3'-end or 5'-end counting protocols, full-length scRNA-seq methods have incomparable advantages in isoform usage analysis, allelic expression detection, and RNA editing identification owing to their superiority of transcript coverage. Moreover, for detecting certain lowly expressed genes/transcripts, full-length scRNA-seq approaches could be better than 3' sequencing methods (Ziegenhain et al., 2017). Notably, droplet-based technologies [e.g., Drop-seq (Macosko et al., 2015), InDrop (Klein et al., 2015), and Chromium (Zheng et al., 2017)] can generally provide a larger throughput of cells and a lower sequencing cost per cell compared to whole-transcript scRNA-seq. Thus, droplet-based protocols are suitable for generating huge amounts of cells to identify the cell subpopulations of complex tissues or tumor samples.

Strikingly, several scRNA-seq technologies can capture both polyA+ and polyA- RNAs, such as SUPeR-seq (Fan X. et al., 2015) and MATQ-seq (Sheng et al., 2017). These protocols are extremely useful for sequencing long noncoding RNAs (lncRNAs) and circular RNAs (circRNAs). Lots of studies have demonstrated that lncRNAs and circRNAs play important roles in diverse biological processes of cells and may serve as crucial biomarkers for cancers (Barrett and Salzman, 2016; Chen et al., 2016b; Quinn and Chang, 2016; Kristensen et al., 2018); therefore, such scRNA-seq methods can provide unprecedented opportunities to comprehensively explore the expression dynamics of both protein-coding and noncoding RNAs at the single-cell level.

Compared to traditional bulk RNA-seq technologies, scRNA-seq protocols suffer higher technical variations. In order to estimate the technical variances among different cells, spike-ins [such as External RNA Control Consortium (ERCC) controls (External, 2005)] and UMIs have been widely used in corresponding scRNA-seq methods. The RNA spike-ins are RNA transcripts (with known sequences and quantity) that are applied to calibrate the measurements of RNA hybridization assays, such as RNA-Seq, and UMIs can theoretically enable the estimation of absolute molecular counts. It is worth noting that ERCC and UMIs are not applicable to all scRNA-seq technologies due to the inherent protocol differences. Spike-ins are used in approaches like Smart-seq2 and SUPeR-seq but are not compatible with droplet-based methods, whereas UMIs are typically applied to 3'-end sequencing technologies [such as Drop-seq (Macosko et al., 2015), InDrop (Klein et al., 2015), and MARS-seq (Jaitin et al., 2014)]. Consequently, users can select the suitable scRNA-seq method according to the technical properties and advantages, number of cells to be sequenced and cost considerations.

READ ALIGNMENT AND EXPRESSION QUANTIFICATION OF SCRNA-SEQ DATA

The mapping ratio of reads is an important indicator of the overall quality of scRNA-seq data. Since both scRNA-seq and

TABLE 1 | Summary of widely used scRNA-seq technologies.

| Methods | Transcript coverage | UMI possibility | Strand specific | References |
|----------------------|---------------------|-----------------|-----------------|--------------------------|
| Tang method | Nearly full-length | No | No | Tang et al., 2009 |
| Quartz-Seq | Full-length | No | No | Sasagawa et al., 2013 |
| SUPeR-seq | Full-length | No | No | Fan X. et al., 2015 |
| Smart-seq | Full-length | No | No | Ramskold et al., 2012 |
| Smart-seq2 | Full-length | No | No | Picelli et al., 2013 |
| MATQ-seq | Full-length | Yes | Yes | Sheng et al., 2017 |
| STRT-seq and STRT/C1 | 5'-only | Yes | Yes | Islam et al., 2011, 2012 |
| CEL-seq | 3'-only | Yes | Yes | Hashimshony et al., 2012 |
| CEL-seq2 | 3'-only | Yes | Yes | Hashimshony et al., 2016 |
| MARS-seq | 3'-only | Yes | Yes | Jaitin et al., 2014 |
| CytoSeq | 3'-only | Yes | Yes | Fan H.C. et al., 2015 |
| Drop-seq | 3'-only | Yes | Yes | Macosko et al., 2015 |
| InDrop | 3'-only | Yes | Yes | Klein et al., 2015 |
| Chromium | 3'-only | Yes | Yes | Zheng et al., 2017 |
| SPLIT-seq | 3'-only | Yes | Yes | Rosenberg et al., 2018 |
| sci-RNA-seq | 3'-only | Yes | Yes | Cao et al., 2017 |
| Seq-Well | 3'-only | Yes | Yes | Gierahn et al., 2017 |
| DroNC-seq | 3'-only | Yes | Yes | Habib et al., 2017 |
| Quartz-Seq2 | 3'-only | Yes | Yes | Sasagawa et al., 2018 |

bulk RNA-seq technologies generally sequence transcripts into reads to generate the raw data in fastq format, no differences exist between these two types of RNA-seq data in read alignment. The mapping tools originally developed for bulk RNA-seq are also applicable to scRNA-seq data. Numerous spliced alignment programs have been designed for mapping RNA-seq data, which was extensively discussed previously (Li and Homer, 2010; Chen et al., 2011). Generally, the read mapping algorithms mainly fall into two categories: spaced-seed indexing based and Burrows-Wheeler transform (BWT) based (Li and Homer, 2010). Currently popular aligners like TopHat2 (Kim et al., 2013), STAR (Dobin and Gingeras, 2015), and HISAT (Kim et al., 2015) perform well in mapping speed and accuracy, and they can efficiently map billions of reads to the reference genome or transcriptome (Table 2). STAR is a suffix-array based method and is faster than TopHat2, but it requires a huge memory size (28 gigabytes for human genome) for read mapping (Dobin and Gingeras, 2015). Engstrom et al. systematically evaluated 26 read alignment protocols (did not include HISAT) and found that different mapping tools exhibit distinct strengths and weakness, where some programs are with a faster mapping speed but a lower accuracy in splice junction detection (Engstrom et al., 2013). HISAT is developed based on BWT and Ferragina-Manzini (FM) methods. Kim et al. (2015) showed that HISAT is currently the fastest tool that can achieve equal or better accuracy than other available aligners.

For gene/transcript expression quantification, distinct approaches are needed, based on the range of transcript sequence captured by scRNA-seq. The data generated by whole-transcript scRNA-seq methods (such as Smart-seq2 and MATQ-seq) can

TABLE 2 | Tools for read mapping and expression quantification of scRNA-seq data.

| Tools | Category | URL | References |
|-----------|---------------------------|---|--------------------------|
| TopHat2 | Read mapping | https://ccb.jhu.edu/software/tophat/index.shtml | Kim et al., 2013 |
| STAR | Read mapping | https://github.com/alexdobin/STAR | Dobin and Gingeras, 2015 |
| HISAT2 | Read mapping | https://ccb.jhu.edu/software/hisat2/index.shtml | Kim et al., 2015 |
| Cufflinks | Expression quantification | https://github.com/cole-trapnell-lab/cufflinks | Trapnell et al., 2010 |
| RSEM | Expression quantification | https://github.com/deweylab/RSEM | Li and Dewey, 2011 |
| StringTie | Expression quantification | https://github.com/gpertea/stringtie | Pertea et al., 2015 |

be analyzed with the software developed for bulk RNA-seq to quantify gene/transcript expression. Two main approaches are available for transcriptome reconstruction: *de novo* assembly (does not need a reference genome) and reference-based or genome-guided assembly (Chen et al., 2017b). *De novo* transcriptome assembly methods are primarily applied to the organisms that lack a reference genome, and are generally with a lower accuracy than that of genome-guided assembly (Garber et al., 2011). The popular genome-guided assembly tools including Cufflinks (Trapnell et al., 2010), RSEM (Li and Dewey, 2011), and StringTie (Pertea et al., 2015) have been broadly used in many scRNA-seq studies to get relative gene/transcript expression estimation in reads or fragments per kilobase per million mapped reads (RPKM or FPKM) or transcripts per million mapped reads (TPM) (Table 2). Pertea et al. (2015) stated that StringTie outperforms other genome-guided approaches in gene/transcript reconstruction and expression quantification. On the other hand, for the 3'-end scRNA-seq protocols (e.g., CEL-seq2, MARS-seq, Drop-seq, and InDrop), specific algorithms are required to calculate gene/transcript expression based on UMIs. SAVER (single-cell analysis via expression recovery) is an efficient UMI-based tool recently proposed for accurately estimating gene expression of single cells (Huang et al., 2018). In theory, UMI-based scRNA-seq can largely reduce the technical noise, which remarkably benefits the estimation of absolute transcript counts (Islam et al., 2014).

QUALITY CONTROL OF SCRNA-SEQ DATA

The limitations in scRNA-seq including bias of transcript coverage, low capture efficiency, and sequencing coverage result in that scRNA-seq data are with a higher level of technical noise than bulk RNA-seq data (Kolodziejczyk et al., 2015). Even for the most sensitive scRNA-seq protocol, it is a frequent phenomenon that some specific transcripts cannot be detected (termed dropout events) (Haque et al., 2017). Generally, scRNA-seq experiments

can generate a portion of low-quality data from the cells that are broken or dead or mixed with multiple cells (Ilicic et al., 2016). These low-quality cells will hinder the downstream analysis and may lead to misinterpretation of the data. Accordingly, QC of scRNA-seq data is crucial to identify and remove the low-quality cells.

To exclude the low-quality cells from scRNA-seq, close attention should be paid to avoid multi-cells or dead cells in the cell capture step. After sequencing, a series of QC analyses are required to eliminate the data from low-quality cells. Those samples contain only a few number of reads should be discarded first since insufficient sequencing depth may lead to the loss of a large portion of lowly and moderately expressed genes. Then tools initially developed for QC of bulk RNA-seq data, such as FastQC¹, can be employed to check the sequencing quality of scRNA-seq data. Moreover, after read alignment, samples with very low mapping ratio should be eliminated because they contain massively unmappable reads that might be resulted from RNA degradation. If extrinsic spike-ins (such ERCC) were used in scRNA-seq, technical noise could be estimated. The cells with an extremely high portion of reads mapped to the spike-ins indicate that they were probably broken during cell capture process and should be removed (Ilicic et al., 2016). Cytoplasmic RNAs are usually lost but mitochondrial RNAs are retained for broken cells, thus the ratio of reads mapped to mitochondrial genome is also informative for identifying low-quality cells (Bacher and Kendzierski, 2016). Additionally, the number of expressed genes/transcripts can be detected in each cell is also suggestive. If only a small number of genes can be detected in a cell, this cell is probably damaged or dead or suffered from RNA degradation. Considering the high noise of scRNA-seq data, a threshold of 1 FPKM/RPKM was usually applied to define the expressed genes. Some QC methods for scRNA-seq have been proposed (Stegle et al., 2015; Ilicic et al., 2016), including SinQC (Jiang et al., 2016) and Scater (McCarthy et al., 2017), these tools are useful for QC of scRNA-seq data.

BATCH EFFECT CORRECTION

Batch effect is a common source of technical variation in high-throughput sequencing experiments. The innovation and decreasing cost of scRNA-seq enable many studies to profile the transcriptomes of a huge amount of cells. The large scale scRNA-seq data sets might be separately generated with distinct operators at different times, and could also be produced in multiple laboratories using disparate cell dissociation protocols, library preparation approaches and/or sequencing platforms. These factors would introduce systematic error and confound the technical and biological variability, leading to that the gene expression profile in one batch systematically differs from that in another (Leek et al., 2010; Hicks et al., 2018). Therefore, batch effect is a major challenge in scRNA-seq data analysis, which may mask the underlying biology and cause spurious results. To avoid incorrect data integration and interpretation, batch effects must

be corrected before the downstream analysis. Because of the data feature differences between scRNA-seq and bulk RNA-seq, batch-correction approaches specially proposed for bulk RNA-seq [e.g., RUVseq (Risso et al., 2014) and svaseq (Leek, 2014)] may not be suitable for scRNA-seq. Several methods have been recently designed to mitigate the batch effects in scRNA-seq data, such as MNN (mutual nearest neighbor) (Haghverdi et al., 2018) and kBET (k-nearest neighbor batch effect test) (Buttner et al., 2019). MNN corrects the batch effects using the data from the most similar cells in different batches. kBET is a χ^2 -based method for quantifying batch effects in scRNA-seq data. These specific batch-correction approaches for scRNA-seq data can perform better than the methods developed for bulk RNA-seq (Haghverdi et al., 2018; Buttner et al., 2019).

NORMALIZATION OF SCRNA-SEQ DATA

To correctly interpret the results from scRNA-seq data, normalization is an essential step to get the signal of interest by adjusting unwanted biases resulted from capture efficiency, sequencing depth, dropouts, and other technical effects. Technical noise of scRNA-seq is an obvious problem due to the low starting material and challenging experimental protocols. Normalization of scRNA-seq data will benefit the downstream analyses including cell subpopulation identification and differential expression calling. In general, normalization can be divided into two different types: within-sample normalization and between-sample normalization (Vallejos et al., 2017). Within-sample normalization aims to remove the gene-specific biases (e.g., GC content and gene length), which makes gene expression comparable within one sample (such as RPKM/FPKM and TPM). In contrast, between-sample normalization is to adjust sample-specific differences (e.g., sequencing depth and capture efficiency) to enable the comparison of gene expression between samples. Generally, those simple normalization strategies are based on sequencing depth or upper quartile. If spike-ins or UMIs are used in scRNA-seq protocol, normalization can be refined based on the performance of spike-ins/UMIs (Bacher and Kendzierski, 2016).

A number of approaches have been developed for between-sample normalization of bulk RNA-seq data, such as DESeq2 (Love et al., 2014) and trimmed mean of M values (TMM) (Robinson and Oshlack, 2010). DESeq2 calculates scaling factor based on the read counts across different samples, while TMM removes the extreme log fold changes (Vallejos et al., 2017). However, bulk-based normalization approaches may be not suitable for the data of single-cell transcriptomics. Because scRNA-seq generates abundant zero-expression values and has a higher level of technical variation than bulk RNA-seq, using bulk RNA-seq normalization approaches may cause overcorrection in scRNA-seq for lowly expressed genes (Vallejos et al., 2017). Several normalization methods have been proposed for scRNA-seq data, such as SCnorm (Bacher et al., 2017), SAMstr (Katayama et al., 2013) and a recently introduced deconvolution approach that uses the summed expression values across pools of cells to conduct normalization (Lun et al., 2016). SCnorm is based on quantile regression, while

¹<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

SAMstrt relies on spike-ins. Bacher et al. (2017) believed that traditional normalization methods developed for bulk RNA-seq may introduce artifacts for normalizing scRNA-seq data, while SCnorm can effectively normalize scRNA-seq data and improve principal component analysis (PCA) and the identification of differentially expressed genes.

IMPUTATION OF SCRNA-SEQ DATA

Single-cell RNA sequencing data generally contain many missing values or dropouts that were caused by failed amplification of the original RNAs. The frequency of dropout events for scRNA-seq is protocol-dependent, and is closely associated with the number of sequencing reads generated for each cell (Svensson et al., 2017). The dropout events increase the cell-to-cell variability, leading to signal influence on every gene, and obscuration of gene-gene relationship detection. Therefore, dropouts can largely affect the downstream analyses since a significant portion of truly expressed transcripts may not be detectable in scRNA-seq. Imputation is a useful strategy to replace the missing data (dropouts) with substituted values. Although some methods have been proposed for imputation of bulk RNA-seq data, they are not directly applicable to scRNA-seq data (Zhang and Zhang, 2018). Several imputation methods have been recently developed for scRNA-seq, including SAVER (Huang et al., 2018), MAGIC (van Dijk et al., 2018), ScImpute (Li and Li, 2018), DrImpute (Gong et al., 2018), and AutoImpute (Talwar et al., 2018). SAVER is a Bayesian-based model designed for UMI-based scRNA-seq data to recover the true expression level of all genes. MAGIC imputes the gene expression by building Markov affinity-based graph. The developers of ScImpute suggested that SAVER and MAGIC may lead to expression changes of the genes unaffected by dropouts, while ScImpute can impute the dropout values without introducing new biases through using the information from the same genes unlikely affected by dropouts in other similar cells. DrImpute is a clustering-based approach and can effectively separate the dropout zeros from true zeros. AutoImpute is an autoencoder-based method that learns the inherent distribution of scRNA-seq data to impute the missing values. Recently, Zhang et al. evaluated different imputation methods and found that the performances of these approaches are correlated with their model hypothesis and scalability (Zhang and Zhang, 2018).

DIMENSIONALITY REDUCTION AND FEATURE SELECTION

Single-cell RNA sequencing data are with a high dimensionality, which may involve thousands of genes and a large number of cells. Dimensionality reduction and feature selection are two main strategies for dealing with high dimensional data (Andrews and Hemberg, 2018a). Dimensionality reduction methods generally project the data into a lower dimensional space by optimally preserving some key properties of the original data. PCA is a linear dimensional reduction algorithm, which assumes

that the data is approximately normally distributed. T-distributed stochastic neighbor embedding (t-SNE) is a non-linear approach mainly designed for visualizing high dimensional data (van der Maaten and Hinton, 2008). Both PCA and t-SNE have been broadly used in diverse scRNA-seq studies to reduce the data dimension and visualize the cells discriminated into distinct subpopulations (Chen et al., 2016a; Rosenberg et al., 2018). It is worth noting that PCA cannot effectively represent the complex structure of scRNA-seq data and t-SNE has limitations of slow computation time and different embeddings for processing the same dataset multiple times. Recently, UMAP (uniform manifold approximation and projection) (Becht et al., 2018), and scvis (Ding et al., 2018) were specially developed for reducing the dimensions of scRNA-seq data. Becht et al. showed that UMAP provides the fastest run times, the highest reproducibility and the most meaningful organization of cell clusters than other dimensionality reduction approaches (Becht et al., 2018).

Feature selection removes the uninformative genes and identifies the most relevant features to reduce the number of dimensions used in downstream analysis. Reducing the number of genes by performing feature selection can largely speed up the calculations of large-scale scRNA-seq data (Andrews and Hemberg, 2018b). Differential expression is a widely used method for feature selection in bulk RNA-seq experiments, but it is hard to apply to scRNA-seq data since the information of predetermined and/or homogeneous subpopulations needed for differential expression calling of scRNA-seq data [e.g., SCDE (Kharchenko et al., 2014)] is often unavailable. Unsupervised feature selection algorithms specially designed for scRNA-seq data can be divided into the following groups: (i) highly variable genes (HVG) based; (ii) spike-in based; and (iii) dropout-based (Andrews and Hemberg, 2018a). HVG methods rely on the assumption that the genes with highly variable expression across cells are resulted from biological effects rather than technical noise. The HVG approaches include algorithms proposed by Brennecke et al. (2013), and FindVariableGenes (FVG) implemented in Seurat (Satija et al., 2015). Spike-in based approaches identify the genes showing significant higher variance than those of spike-ins with similar expression levels [e.g., scLVM (Buettner et al., 2015) and BASiCS (Vallejos et al., 2015)], which shares similar idea of HVG. Dropout based methods take advantage of the dropout distribution of scRNA-seq data to perform feature selection, like M3Drop (Andrews and Hemberg, 2018b). Andrews and Hemberg showed that their M3Drop tool outperforms existing variance-based feature selection approaches.

CELL SUBPOPULATION IDENTIFICATION

A key goal of scRNA-seq data analysis is to identify cell subpopulations (different populations are often distinct cell types) within a certain condition or tissue to unravel the heterogeneity of cells. Notably, cell subpopulation identification should be carried out after QC and normalization of scRNA-seq data, otherwise artifacts could be introduced. Approaches for

clustering cells can be mainly grouped into two categories based on whether prior information is used. If a set of known markers was used in clustering, the methods are prior information based. Alternatively, unsupervised clustering methods can be used for *de novo* identification of cell populations with scRNA-seq data. The algorithms for unsupervised clustering can be primarily divided into the following types: (i) k-means; (ii) hierarchical clustering; (iii) density-based clustering; and (iv) graph-based clustering (Andrews and Hemberg, 2018a). K-means is a fast approach that assigns cells to the nearest cluster center, and it requires the predetermined number of clusters. Hierarchical clustering can determine the relationships between clusters, but it generally works slower than k-means. Density-based clustering methods need a large number of samples to accurately calculate densities and usually assume that all clusters have equal density. Graph-based clustering can be considered as an extension of density-based clustering, and it can be applied to millions of cells. Some clustering methods have been specially designed for scRNA-seq data, such as single-cell consensus clustering (SC3) (Kiselev et al., 2017) and the clustering approach implemented in Seurat (Satija et al., 2015), which can facilitate the identification of cell subpopulations (Table 3). SC3 is an unsupervised approach that combines multiple clustering approaches, which has a high accuracy and robustness in single-cell clustering. Seurat identifies the cell clusters mainly based on a shared nearest neighbor (SNN) clustering algorithm. Once the subpopulations are determined, the markers that can best discriminate distinct subpopulations are usually identified through differential expression calling or analysis of variance (ANOVA).

DIFFERENTIAL EXPRESSION ANALYSIS OF SCRNA-SEQ DATA

Differential expression analysis is very useful to find the significantly differentially expressed genes (DEGs) between distinct subpopulations or groups of cells. The DEGs are crucial for interpreting the biological difference between two compared

conditions. The technical variability, high noise (e.g., dropouts) and massive sample size of scRNA-seq data raise challenges in differential expression calling (McDavid et al., 2013). Moreover, multiple possible cell states can exist within a population of cells, leading to the multimodality of gene expression in cells (Vallejos et al., 2016). The tools originally developed for bulk RNA-seq data have been used in many single-cell studies to identify the DEGs, but the applicability of these methods for scRNA-seq data is still unclear. In recent years, some specific methods have been proposed for conducting differential expression calling based on scRNA-seq data, such as MAST (Finak et al., 2015), SCDE (Kharchenko et al., 2014), DEsingle (Miao et al., 2018), Census (Qiu et al., 2017), and BCseq (Chen and Zheng, 2018) (Table 4). MAST is based on linear model fitting and likelihood ratio testing. SCDE is a Bayesian approach using a low-magnitude Poisson process to account for dropouts. DEsingle employs Zero-Inflated Negative Binomial model to estimate the dropouts and real zeros. BCseq mitigates the technical noise in a data-adaptive manner. Soneson and Robinson recently assessed 36 differential expression methods (including the tools designed for scRNA-seq and bulk RNA-seq data) and revealed significant differences among these approaches in the characteristics and number of DEGs (Soneson and Robinson, 2018). An increasing number of tools for differential expression analysis of scRNA-seq data will be developed, and users are encouraged to choose the tools specially

TABLE 4 | Differential expression analysis tools for RNA-seq data.

| Methods | Category | URL | References |
|----------|-------------|---|-----------------------------|
| ROTS | Single cell | https://bioconductor.org/packages/release/bioc/html/ROTS.html | Seyednasrollah et al., 2016 |
| MAST | Single cell | https://github.com/RGLab/MAST | Finak et al., 2015 |
| BCseq | Single cell | https://bioconductor.org/packages/devel/bioc/html/bcSeq.html | Chen and Zheng, 2018 |
| SCDE | Single cell | http://hms-dbmi.github.io/scde/ | Kharchenko et al., 2014 |
| DEsingle | Single cell | https://bioconductor.org/packages/DEsingle | Miao et al., 2018 |
| Census | Single cell | http://cole-trapnell-lab.github.io/monocle-release/ | Qiu et al., 2017 |
| D3E | Single cell | https://github.com/hemberg-lab/D3E | Delmans and Hemberg, 2016 |
| BPSC | Single cell | https://github.com/nghiavtr/BPSC | Vu et al., 2016 |
| DESeq2 | Bulk | https://bioconductor.org/packages/release/bioc/html/DESeq2.html | Love et al., 2014 |
| edgeR | Bulk | https://bioconductor.org/packages/release/bioc/html/edgeR.html | Robinson et al., 2010 |
| Limma | Bulk | https://bioconductor.org/packages/release/bioc/html/limma.html | Ritchie et al., 2015 |
| Ballgown | Bulk | http://www.bioconductor.org/packages/release/bioc/html/ballgown.html | Frazee et al., 2015 |

TABLE 3 | Subpopulation identification methods for scRNA-seq data.

| Methods | URL | References |
|-----------|---|---------------------------|
| SC3 | http://bioconductor.org/packages/SC3 | Kiselev et al., 2017 |
| ZIFA | https://github.com/epierson9/ZIFA | Pierson and Yau, 2015 |
| Destiny | https://github.com/theislab/destiny | Angerer et al., 2016 |
| SNN-Cliq | http://bioinfo.uncc.edu/SNNCliq/ | Xu and Su, 2015 |
| RaceID | https://github.com/dgrun/RaceID | Grun et al., 2015 |
| SCUBA | https://github.com/gcyuan/SCUBA | Marco et al., 2014 |
| BackSPIN | https://github.com/linnarsson-lab/BackSPIN | Zeisel et al., 2015 |
| PAGODA | http://hms-dbmi.github.io/scde/ | Fan et al., 2016 |
| CIDR | https://github.com/VCCRI/CIDR | Lin et al., 2017 |
| pcaReduce | https://github.com/JustinaZ/pcaReduce | Zurauskiene and Yau, 2016 |
| Seurat | https://github.com/satijalab/seurat | Satija et al., 2015 |
| TSCAN | https://github.com/zij90/TSCAN | Ji and Ji, 2016 |

designed for scRNA-seq to identify DEGs in consideration of the complex features of scRNA-seq data.

CELL LINEAGE AND PSEUDOTIME RECONSTRUCTION

The cells in many biological systems exhibit a continuous spectrum of states and involve transitions between different cellular states. Such dynamic processes within a portion of cells can be computationally modeled by reconstructing the cell trajectory and pseudotime based on scRNA-seq data. Pseudotime is an ordering of cells along the trajectory of a continuously developmental process in a system, which allows the identification of the cell types at the beginning, intermediate, and end states of the trajectory (Griffiths et al., 2018). Besides revealing the gene expression dynamics across cells, single-cell trajectory inference can also benefit the identification of the factors triggering state transitions. A number of tools have been proposed for trajectory inference, e.g., Monocle (Trapnell et al., 2014), Waterfall (Shin et al., 2015), Wishbone (Setty et al., 2016), TSCAN (Ji and Ji, 2016), Monocle2 (Qiu et al., 2017), Slingshot (Street et al., 2018), and CellRouter (Lummertz da Rocha et al., 2018) (Table 5). The resulting trajectory topology can be linear, bifurcating, or a tree/graph structure. Monocle builds a minimum spanning tree (MST) for cells to search for the longest backbone based on independent component analysis (ICA). Monocle2 uses a distinct approach that incorporates unsupervised data-driven methods with reversed graph embedding (RGE), which is more robust and much faster than Monocle. Slingshot is a cluster-based approach for identifying multiple trajectories with varying levels of supervision. CellRouter utilizes flow networks to identify cell-state transition trajectories. Recently, Saelens et al. (2018) evaluated a number of single-cell trajectory inference approaches (did not include CellRouter), and found that Slingshot, TSCAN and Monocle2 outperform other methods.

ALTERNATIVE SPLICING AND RNA EDITING ANALYSIS OF SCRNA-SEQ DATA

Most of published single-cell studies mainly explored the transcriptome variation between individual cells at gene level. In eukaryotic genome, AS allows multi-exon genes to generate different isoforms, which can largely increase the diversity of both protein-coding and noncoding RNAs. Five basic modes are generally recognized for AS, including exon-skipping (cassette exon), mutually exclusive exons, alternative donor site, alternative acceptor site, and intron retention. Lots of studies have shown that AS is very common in mammals and over 90% of human genes could undergo AS based on bulk RNA-seq data (Wang et al., 2008; Chen et al., 2017a). Moreover, AS play crucial roles in a variety of biological processes and abnormal AS may be correlated with cancers (Sveen et al., 2016). The findings revealed by bulk RNA-seq data can only reflect the averaged AS patterns of numerous cells at population level.

TABLE 5 | Methods for single-cell trajectory inference.

| Tools | Dimensionality reduction | URL | References |
|------------|--|---|--------------------------------|
| Monocle | ICA | http://cole-trapnell-lab.github.io/monocle-release/ | Trapnell et al., 2014 |
| Waterfall | PCA | https://www.cell.com/cms/10.1016/j.stem.2015.07.013/attachment/3e966901-034f-418a-a439-996c50292a11/mmc9.zip | Shin et al., 2015 |
| Wishbone | Diffusion maps | https://github.com/ManuSetty/wishbone | Setty et al., 2016 |
| GrandPrix | Gaussian Process Latent Variable Model | https://github.com/ManchesterBioinference/GrandPrix | Ahmed et al., 2019 |
| SCUBA | t-SNE | https://github.com/gcyuan/SCUBA | Marco et al., 2014 |
| DPT | Diffusion maps | https://media.nature.com/original/nature-assets/nmeth/journal/v13/n10/extref/nmeth.3971-S3.zip | Haghighverdi et al., 2016 |
| TSCAN | PCA | https://github.com/zji90/TSCAN | Ji and Ji, 2016 |
| Monocle2 | RGE | http://cole-trapnell-lab.github.io/monocle-release/ | Qiu et al., 2017 |
| Slingshot | Any | https://github.com/kstreet13/slingshot | Street et al., 2018 |
| CellRouter | Any | https://github.com/edroaldo/cellrouter | Lummertz da Rocha et al., 2018 |

Due to the high noise (e.g., dropouts and uneven transcript coverage) and low sequencing coverage of scRNA-seq data, the splicing quantification methods initially developed for bulk RNA-seq data are not suitable for scRNA-seq data. Since expression dynamics is a key aspect of cell populations, it is promising to study AS at single-cell resolution to gain insights into cell-level isoform usage. To date, only a few number of AS detection approaches were devised for scRNA-seq data, such as SingleSplice (Welch et al., 2016), Census (Qiu et al., 2017), BRIE (Huang and Sanguinetti, 2017), and Expedition (Song et al., 2017) (Table 6). SingleSplice uses a statistical model to detect the genes with a significant isoform usage without estimating the expression levels of full-length transcripts. Census models the isoform counts of each gene with a linear model as a Dirichlet-multinomial distribution. BRIE is a Bayesian hierarchical model for differential isoform quantification. Expedition contains a suite of algorithms for identifying AS, assigning splicing modalities and visualize modality changes. The AS detection approaches specially designed for scRNA-seq data are just emerging, thus the innovation and improvement of such methods will largely facilitate AS exploration at the single-cell level.

On the other hand, RNA-editing is an important post-transcriptional processing event that leads to sequence changes on RNA molecules (Gott and Emeson, 2000). Similarly, RNA-editing is mainly studied using bulk RNA-seq technologies but rarely explored at the single-cell level. Currently, the limitations of scRNA-seq largely prevented the application of RNA-editing

TABLE 6 | Alternative splicing detection tools for scRNA-seq data.

| Tools | URL | References |
|--------------|---|-----------------------------|
| SingleSplice | https://github.com/jw156605/SingleSplice | Welch et al., 2016 |
| Expedition | https://github.com/YeoLab/Expedition | Song et al., 2017 |
| BRIE | https://github.com/huangyh09/brie | Huang and Sanguinetti, 2017 |
| Census | http://cole-trapnell-lab.github.io/monocle-release/ | Qiu et al., 2017 |

detection to individual cells. Accordingly, with the development of both scRNA-seq technologies and single-cell editing detection algorithms, exploration of RNA-editing dynamics among single cells will be feasible. Notably, both AS and RNA-editing are mainly suitable for the data generated by scRNA-seq protocols that can sequence full-length transcripts such as Smart-seq2 and MATQ-seq rather than 3'-end scRNA-seq approaches.

ALLELIC EXPRESSION EXPLORATION WITH SCRNA-SEQ DATA

Diploid species contain two sets of chromosomes that are separately obtained from their parents. Allelic expression analysis can reveal whether genes are equally expressed between parental and maternal genomes. For autosomes, the parental and maternal expression are generally expressed equally, and aberrant expression of parental or maternal genome may cause certain diseases (McKean et al., 2016). Up to now, few methods were developed to detect the genome-wide allelic expression profile of genes based on scRNA-seq data. One main caution of allelic expression calling is that the high dropouts of scRNA-seq data may introduce many false positives. Deng et al. (2014) used a series of stringent criteria to filter the potentially false allelic calls resulted from the technical variability of scRNA-seq in studying allelic expression profile of mouse preimplantation embryos. The robustness of this strategy was further demonstrated in analyzing the dynamics of X chromosome inactivation along developmental progression using mouse embryonic stem cells (Chen et al., 2016a). SCALE was recently proposed to classify the gene expression into silent, monoallelic and biallelic, states by adopting an empirical Bayes approach (Jiang et al., 2017). We believe that allelic expression analysis at single-cell level can largely facilitate the understanding of the underlying mechanisms of dosage compensation and related diseases. It is worth noting that allelic expression investigation at single-cell level also needs the whole-transcript scRNA-seq and is mainly applicable to the organism that has available paternal and maternal single nucleotide polymorphism (SNP) information.

GENE REGULATORY NETWORK RECONSTRUCTION

Gene regulatory network inference has been widely conducted in numerous bulk RNA-seq studies, while scRNA-seq also

provides great potential for such analysis. For bulk RNA-seq data, networks are usually constructed from a number of samples using the tools like weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008; Chen et al., 2017a). A basic assumption is that the genes highly correlated in expression could be co-regulated. Because such an analysis is unable to determine the regulatory relationship, the resulting networks are typically undirected. Theoretically, the cells of scRNA-seq can be treated as the samples of bulk RNA-seq, then similar approaches are applicable to scRNA-seq data for constructing gene regulatory network.

Network inference of scRNA-seq data may reveal meaningful gene correlations and provide biologically important insights that could not be uncovered by population-level data of bulk RNA-seq. However, due to the technical noise of scRNA-seq and different subpopulations or states of cells, attention should be paid to network reconstruction. To reduce spurious results, network inference should be carried out on each subpopulation or the cells with the same stage. Recently, Aibar et al. (2017) developed SCENIC method to reconstruct the gene regulatory network from scRNA-seq data and they showed that SCENIC can robustly predict the interactions between transcription factors and target genes. PIDC is another software designed to infer gene regulatory network from single-cell data using multivariate information theory (Chan et al., 2017). Such network inference tools facilitate the identification of expression regulatory network from single-cell transcriptomic data and provide critically biological insights into the regulatory relationships between genes.

CONCLUSION

In the past 10 years, a great advancement has been achieved in scRNA-seq and a variety of scRNA-seq protocols have been developed. The development and innovation of scRNA-seq largely facilitated single-cell transcriptomic studies, leading to insightful findings in cell expression variability and dynamics. Moreover, the throughput of scRNA-seq has significantly increased with the exciting progress in cellular barcoding and microfluidics. Meanwhile, scRNA-seq methods that can be used for fixation and frozen samples have also been proposed recently, which will greatly benefit the study of highly heterogeneous clinical samples. However, currently available scRNA-seq approaches still have a high dropout problem, in which weakly expressed genes would be missed. The improvement of RNA capture efficiency and transcript coverage will definitely reduce the technical noise of scRNA-seq. Moreover, since most of current scRNA-seq methods mainly capture polyA+ RNAs, the development of protocols that can capture both polyA+ and polyA- RNAs (such as MATQ-seq) will enable comprehensive investigation of both protein-coding and non-coding gene expression dynamics at single-cell resolution.

Since the noise of scRNA-seq data is high, it is crucial to use appropriate methods to overcome the problem in analyzing scRNA-seq data. QC is necessary to exclude those low-quality cells to avoid involving artifacts in data interpretation. Furthermore, batch effect correction (if need), between sample

normalization and imputation are also important and should be conducted before cell subpopulation identification, differential expression calling, and other downstream analyses. Additionally, factors such as cell size and cell cycle state could play important roles in cell variability for certain types of cells, such biases are also need to be considered. Although an increasing number of methods have been specially designed to interpret scRNA-seq data, advances of novel methods that can effectively handle the technical noise and expression variability of cells are still required. Specifically, the approaches that can accurately analyze AS and RNA-editing with scRNA-seq data are highly useful to unravel post-transcriptional mechanisms in individual cells. Overall, bioinformatics analysis of scRNA-seq data is still challenging, special attention should be paid in data interpretation, and more efficient tools are in urgent need.

Collectively, scRNA-seq and its related computational methods largely promote the development of single-cell

transcriptomics. The continuous innovation of scRNA-seq technologies and concomitant advances in bioinformatics approaches will greatly facilitate biological and clinical researches, and provide deep insights into the gene expression heterogeneity and dynamics of cells.

AUTHOR CONTRIBUTIONS

GC and TS designed the study and wrote the manuscript. BN edited the manuscript and provided constructive comments.

FUNDING

This work was supported by the National Science Foundation of China (31771460, 91629103 and 31671377), National Key Research and Development Program of China (2016YFC0902100).

REFERENCES

- Ahmed, S., Rattray, M., and Boukouvelas, A. (2019). GrandPrix: scaling up the Bayesian GPLVM for single-cell data. *Bioinformatics* 35, 47–54. doi: 10.1093/bioinformatics/bty533
- Aibar, S., Gonzalez-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086. doi: 10.1038/nmeth.4463
- Andrews, T. S., and Hemberg, M. (2018a). Identifying cell populations with scRNASeq. *Mol. Aspects Med.* 59, 114–122. doi: 10.1016/j.mam.2017.07.002
- Andrews, T. S., and Hemberg, M. (2018b). M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics* doi: 10.1093/bioinformatics/bty1044 [Epub ahead of print].
- Angerer, P., Haghighi, L., Buttner, M., Theis, F. J., Marr, C., and Buettner, F. (2016). destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* 32, 1241–1243. doi: 10.1093/bioinformatics/btv715
- Bacher, R., Chu, L. F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., et al. (2017). SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* 14, 584–586. doi: 10.1038/nmeth.4263
- Bacher, R., and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* 17:63. doi: 10.1186/s13059-016-0927-y
- Barrett, S. P., and Salzman, J. (2016). Circular RNAs: analysis, expression and potential functions. *Development* 143, 1838–1847. doi: 10.1242/dev.128074
- Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., et al. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37, 38–44. doi: 10.1038/nbt.4314
- Brennecke, P., Anders, S., Kim, J. K., Kolodziejczyk, A. A., Zhang, X., Proserpio, V., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* 10, 1093–1095. doi: 10.1038/nmeth.2645
- Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., et al. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* 33, 155–160. doi: 10.1038/nbt.3102
- Buttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., and Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* 16, 43–49. doi: 10.1038/s41592-018-0254-1
- Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357, 661–667. doi: 10.1126/science.aam8940
- Chan, T. E., Stumpf, M. P. H., and Babbie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* 5, 251–267.e3. doi: 10.1016/j.cels.2017.08.014
- Chen, G., Chen, J., Yang, J., Chen, L., Qu, X., Shi, C., et al. (2017a). Significant variations in alternative splicing patterns and expression profiles between human-mouse orthologs in early embryos. *Sci. China Life Sci.* 60, 178–188. doi: 10.1007/s11427-015-0348-5
- Chen, G., Shi, T. L., and Shi, L. M. (2017b). Characterizing and annotating the genome using RNA-seq data. *Sci. China Life Sci.* 60, 116–125. doi: 10.1007/s11427-015-0349-4
- Chen, G., Schell, J. P., Benitez, J. A., Petropoulos, S., Yilmaz, M., Reinius, B., et al. (2016a). Single-cell analyses of X Chromosome inactivation dynamics and pluripotency during differentiation. *Genome Res.* 26, 1342–1354. doi: 10.1101/gr.201954.115
- Chen, G., Yang, J., Chen, J., Song, Y., Cao, R., Shi, T., et al. (2016b). Identifying and annotating human bifunctional RNAs reveals their versatile functions. *Sci. China Life Sci.* 59, 981–992. doi: 10.1007/s11427-016-0054-1
- Chen, G., Wang, C., and Shi, T. (2011). Overview of available methods for diverse RNA-Seq data analyses. *Sci. China Life Sci.* 54, 1121–1128. doi: 10.1007/s11427-011-4255-x
- Chen, L., and Zheng, S. (2018). BCseq: accurate single cell RNA-seq quantification with bias correction. *Nucleic Acids Res.* 46:e82. doi: 10.1093/nar/gky308
- Chen, X., Teichmann, S. A., and Meyer, K. B. (2018). From tissues to cell types and back: single-cell gene expression analysis of tissue architecture. *Annu. Rev. Biomed. Data Sci.* 1, 29–51. doi: 10.1146/annurev-biodatasci-080917-013452
- Delmans, M., and Hemberg, M. (2016). Discrete distributional differential expression (D3E)—a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics* 17:110. doi: 10.1186/s12859-016-0944-6
- Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343, 193–196. doi: 10.1126/science.1245316
- Ding, J., Condon, A., and Shah, S. P. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* 9:2002. doi: 10.1038/s41467-018-04368-5
- Dobin, A., and Gingeras, T. R. (2015). Mapping RNA-seq reads with STAR. *Curr. Protoc. Bioinformatics* 51, 11.14.1–11.14.19. doi: 10.1002/0471250953.bi114551
- Engstrom, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Ratsch, G., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* 10, 1185–1191. doi: 10.1038/nmeth.2722
- External, R. N. A. C. C. (2005). Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics* 6:150. doi: 10.1186/1471-2164-6-150
- Fan, H. C., Fu, G. K., and Fodor, S. P. (2015). Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science* 347:1258367. doi: 10.1126/science.1258367
- Fan, X., Zhang, X., Wu, X., Guo, H., Hu, Y., Tang, F., et al. (2015). Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse

- preimplantation embryos. *Genome Biol.* 16:148. doi: 10.1186/s13059-015-0706-1
- Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., et al. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* 13, 241–244. doi: 10.1038/nmeth.3734
- Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16:278. doi: 10.1186/s13059-015-0844-5
- Frazee, A. C., Pertea, G., Jaffe, A. E., Langmead, B., Salzberg, S. L., and Leek, J. T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat. Biotechnol.* 33, 243–246. doi: 10.1038/nbt.3172
- Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* 8, 469–477. doi: 10.1038/nmeth.1613
- Gierahn, T. M., Wadsworth, M. H. II, Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., et al. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* 14, 395–398. doi: 10.1038/nmeth.4179
- Gong, W., Kwak, I. Y., Pota, P., Koyano-Nakagawa, N., and Garry, D. J. (2018). DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* 19:220. doi: 10.1186/s12859-018-2226-y
- Gott, J. M., and Emeson, R. B. (2000). Functions and mechanisms of RNA editing. *Annu. Rev. Genet.* 34, 499–531. doi: 10.1146/annurev.genet.34.1.499
- Griffiths, J. A., Scialdone, A., and Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* 14:e8046. doi: 10.1525/msb.20178046
- Gross, A., Schoendube, J., Zimmermann, S., Steeb, M., Zengerle, R., and Koltay, P. (2015). Technologies for single-cell isolation. *Int. J. Mol. Sci.* 16, 16897–16919. doi: 10.3390/ijms160816897
- Grun, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., et al. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525, 251–255. doi: 10.1038/nature14966
- Habib, N., Avraham-David, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., et al. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* 14, 955–958. doi: 10.1038/nmeth.4407
- Haghverdi, L., Buttner, M., Wolf, F. A., Büttner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* 13, 845–848. doi: 10.1038/nmeth.3971
- Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36, 421–427. doi: 10.1038/nbt.4091
- Haue, A., Engel, J., Teichmann, S. A., and Lonnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* 9:75. doi: 10.1186/s13073-017-0467-4
- Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., et al. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* 17:77. doi: 10.1186/s13059-016-0938-8
- 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. doi: 10.1016/j.celrep.2012.08.003
- Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19, 562–578. doi: 10.1093/biostatistics/kxx053
- Hu, P., Zhang, W., Xin, H., and Deng, G. (2016). Single cell isolation and analysis. *Front. Cell. Dev. Biol.* 4:116. doi: 10.3389/fcell.2016.00116
- Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., et al. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* 15, 539–542. doi: 10.1038/s41592-018-0033-z
- Huang, Y., and Sanguinetti, G. (2017). BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biol.* 18:123. doi: 10.1186/s13059-017-1248-5
- Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* 50:96. doi: 10.1038/s12276-018-0071-8
- Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., et al. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* 17:29. doi: 10.1186/s13059-016-0888-1
- Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J. B., Lonnberg, P., et al. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 21, 1160–1167. doi: 10.1101/gr.110882.110
- Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J. B., Lonnberg, P., et al. (2012). Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nat. Protoc.* 7, 813–828. doi: 10.1038/nprot.2012.022
- Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., et al. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* 11, 163–166. doi: 10.1038/nmeth.2772
- Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343, 776–779. doi: 10.1126/science.1247651
- Ji, Z., and Ji, H. (2016). TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* 44:e117. doi: 10.1093/nar/gkw430
- Jiang, P., Thomson, J. A., and Stewart, R. (2016). Quality control of single-cell RNA-seq by SinQC. *Bioinformatics* 32, 2514–2516. doi: 10.1093/bioinformatics/btw176
- Jiang, Y., Zhang, N. R., and Li, M. (2017). SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol.* 18:74. doi: 10.1186/s13059-017-1200-8
- Katayama, S., Tohonen, V., Linnarsson, S., and Kere, J. (2013). SAMstrat: statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics* 29, 2943–2945. doi: 10.1093/bioinformatics/btt511
- Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11, 740–742. doi: 10.1038/nmeth.2967
- Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36
- Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14, 483–486. doi: 10.1038/nmeth.4236
- Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201. doi: 10.1016/j.cell.2015.04.044
- Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58, 610–620. doi: 10.1016/j.molcel.2015.04.005
- Kristensen, L. S., Hansen, T. B., Veno, M. T., and Kjems, J. (2018). Circular RNAs in cancer: opportunities and challenges in the field. *Oncogene* 37, 555–565. doi: 10.1038/onc.2017.361
- Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559
- Leek, J. T. (2014). svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* 42:e161. doi: 10.1093/nar/gku864
- Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11, 733–739. doi: 10.1038/nrg2825
- Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323
- Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.* 11, 473–483. doi: 10.1093/bib/bbq015
- Li, W. V., and Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* 9:997. doi: 10.1038/s41467-018-03405-7
- Lin, P., Troup, M., and Ho, J. W. (2017). CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* 18:59. doi: 10.1186/s13059-017-1188-0

- 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. doi: 10.1186/s13059-014-0550-8
- Lummertz da Rocha, E., Rowe, R. G., Lundin, V., Malleshaiah, M., Jha, D. K., Rambo, C. R., et al. (2018). Reconstruction of complex single-cell trajectories using CellRouter. *Nat. Commun.* 9:892. doi: 10.1038/s41467-018-03214-y
- Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17:75. doi: 10.1186/s13059-016-0947-7
- Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi: 10.1016/j.cell.2015.05.002
- Marco, E., Karp, R. L., Guo, G., Robson, P., Hart, A. H., Trippa, L., et al. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci. U.S.A.* 111, E5643–E5650. doi: 10.1073/pnas.1408931111
- 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. doi: 10.1093/bioinformatics/btw777
- McDavid, A., Finak, G., Chattopadhyay, P. K., Dominguez, M., Lamoreaux, L., Ma, S. S., et al. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* 29, 461–467. doi: 10.1093/bioinformatics/bts714
- McKean, D. M., Homsy, J., Wakimoto, H., Patel, N., Gorham, J., DePalma, S. R., et al. (2016). Loss of RNA expression and allele-specific expression associated with congenital heart disease. *Nat. Commun.* 7:12824. doi: 10.1038/ncomms12824
- Miao, Z., Deng, K., Wang, X., and Zhang, X. (2018). DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* 34, 3223–3224. doi: 10.1093/bioinformatics/bty332
- Nichterwitz, S., Chen, G., Aguila Benitez, J., Yilmaz, M., Storvall, H., Cao, M., et al. (2016). Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. *Nat. Commun.* 7:12139. doi: 10.1038/ncomms12139
- Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122
- Picelli, S. (2017). Single-cell RNA-sequencing: the future of genome biology is now. *RNA Biol.* 14, 637–650. doi: 10.1080/15476286.2016.1201618
- Picelli, S., Bjorklund, A. K., Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 10, 1096–1098. doi: 10.1038/nmeth.2639
- Pierson, E., and Yau, C. (2015). ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* 16:241. doi: 10.1186/s13059-015-0805-z
- Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y. A., and Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* 14, 309–315. doi: 10.1038/nmeth.4150
- Quinn, J. J., and Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* 17, 47–62. doi: 10.1038/nrg.2015.10
- Ramskold, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., Faridani, O. R., et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* 30, 777–782. doi: 10.1038/nbt.2282
- Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902. doi: 10.1038/nbt.2931
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616
- Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25. doi: 10.1186/gb-2010-11-3-r25
- Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360, 176–182. doi: 10.1126/science.aam8999
- Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2018). A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv* [Preprint]. doi: 10.1101/276907
- Sasagawa, Y., Danno, H., Takada, H., Ebisawa, M., Tanaka, K., Hayashi, T., et al. (2018). Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* 19:29. doi: 10.1186/s13059-018-1407-3
- Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K. D., Imai, T., et al. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* 14:R31. doi: 10.1186/gb-2013-14-4-r31
- Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33, 495–502. doi: 10.1038/nbt.3192
- Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., et al. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* 34, 637–645. doi: 10.1038/nbt.3569
- Seyednasrollah, F., Rantanen, K., Jaakkola, P., and Elo, L. L. (2016). ROTS: reproducible RNA-seq biomarker detector-prognostic markers for clear cell renal cell cancer. *Nucleic Acids Res.* 44:e1. doi: 10.1093/nar/gkv806
- Sheng, K., Cao, W., Niu, Y., Deng, Q., and Zong, C. (2017). Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat. Methods* 14, 267–270. doi: 10.1038/nmeth.4145
- Shin, J., Berg, D. A., Zhu, Y., Shin, J. Y., Song, J., Bonaguidi, M. A., et al. (2015). Single-cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* 17, 360–372. doi: 10.1016/j.stem.2015.07.013
- Soneson, C., and Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* 15, 255–261. doi: 10.1038/nmeth.4612
- Song, Y., Botvinnik, O. B., Lovci, M. T., Kakaradov, B., Liu, P., Xu, J. L., et al. (2017). Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Mol. Cell* 67, 148–161.e5. doi: 10.1016/j.molcel.2017.06.003
- Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16, 133–145. doi: 10.1038/nrg3833
- Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., et al. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 19:477. doi: 10.1186/s12864-018-4772-0
- Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R. A., and Skotheim, R. I. (2016). Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene* 35, 2413–2427. doi: 10.1038/nc.2015.318
- Svensson, V., Natarajan, K. N., Ly, L. H., Miragaia, R. J., Labalette, C., Macaulay, I. C., et al. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14, 381–387. doi: 10.1038/nmeth.4220
- Talwar, D., Mongia, A., Sengupta, D., and Majumdar, A. (2018). AutoImpute: autoencoder based imputation of single-cell RNA-seq data. *Sci. Rep.* 8:16329. doi: 10.1038/s41598-018-34688-x
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi: 10.1038/nmeth.1315
- Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386. doi: 10.1038/nbt.2859
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621
- Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). BASiCS: bayesian analysis of single-cell sequencing data. *PLoS Comput. Biol.* 11:e1004333. doi: 10.1371/journal.pcbi.1004333

- Vallejos, C. A., Richardson, S., and Marioni, J. C. (2016). Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biol.* 17:70. doi: 10.1186/s13059-016-0930-3
- Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* 14, 565–571. doi: 10.1038/nmeth.4292
- van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.
- van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell* 174, 716–729.e27. doi: 10.1016/j.cell.2018.05.061
- Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Rantalainen, M., et al. (2016). Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* 32, 2128–2135. doi: 10.1093/bioinformatics/btw202
- Wang, E. T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476. doi: 10.1038/nature07509
- Welch, J. D., Hu, Y., and Prins, J. F. (2016). Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Res.* 44:e73. doi: 10.1093/nar/gkv1525
- Xu, C., and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31, 1974–1980. doi: 10.1093/bioinformatics/btv088
- Zeisel, A., Munoz-Manchado, A. B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142. doi: 10.1126/science.aaa1934
- Zhang, L., and Zhang, S. (2018). Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2018.2848633 [Epub ahead of print].
- Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049. doi: 10.1038/ncomms14049
- Ziegenhain, C., Vieth, B., Parekh, S., Reinus, B., Guillaumet-Adkins, A., Smets, M., et al. (2017). Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* 65, 631–643.e4. doi: 10.1016/j.molcel.2017.01.023
- Zurauskiene, J., and Yau, C. (2016). pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* 17:140. doi: 10.1186/s12859-016-0984-y

Disclaimer: The information in these materials is not a formal dissemination of the United States Food and Drug Administration.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Chen, Ning and Shi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.