Molecules and Cells

Minireview

Dissecting Cellular Heterogeneity Using Single-Cell RNA Sequencing

Yoon Ha Choi, and Jong Kyoung Kim*

Department of New Biology, DGIST, Daegu 42988, Korea
*Correspondence: jkkim@dgist.ac.kr
http://dx.doi.org/10.14348/molcells.2019.2446
www.molcells.org

Cell-to-cell variability in gene expression exists even in a homogeneous population of cells. Dissecting such cellular heterogeneity within a biological system is a prerequisite for understanding how a biological system is developed, homeostatically regulated, and responds to external perturbations. Single-cell RNA sequencing (scRNA-seq) allows the quantitative and unbiased characterization of cellular heterogeneity by providing genome-wide molecular profiles from tens of thousands of individual cells. A major question in analyzing scRNA-seq data is how to account for the observed cell-to-cell variability. In this review, we provide an overview of scRNA-seq protocols, computational approaches for dissecting cellular heterogeneity, and future directions of single-cell transcriptomic analysis.

Keywords: cellular heterogeneity, RNA sequencing, single-cell, single-cell genomics, single-cell transcriptomics

INTRODUCTION

A single fertilized egg gives rise to all cell types in the human body. Despite carrying the same genetic information, every cell in our body is unique and shows substantial variability in cellular phenotype compared with other cells (Eldar and Elowitz, 2010; Raj and van Oudenaarden, 2008). A central challenge in biology is to understand how such cellular diversity is generated from a single cell, how it is regulated for tissue homeostasis, and how it is exploited for mounting appropriate responses to external perturbations in normal and diseased tissues. Answering these questions requires single-cell measurements of molecular and cellular features.

Over the past decade, single-cell RNA sequencing (scRNA-seq) technologies have been developed that provide an unbiased view of cell-to-cell variability in gene expression within a population of cells (Chen et al., 2018; Kolodziejczyk et al., 2015a; Tanay and Regev, 2017; Wagner et al., 2016). Recent technological developments in both microfluidic and barcoding approaches allow the transcriptomes of tens of thousands of single cells to be assayed. Coupled with the exponential increase in the amount of single-cell transcriptomic data, computational tools necessary to achieve robust biological findings are being actively developed (Stegle et al., 2015; Zappia et al., 2018). In this review, we provide an overview of scRNA-seq protocols and existing computational methods for dissecting cellular heterogeneity from scRNA-seq data, and discuss their assumptions and limitations. We also examine potential future developments in the field of single-cell genomics.

TECHNOLOGIES OF SCRNA-SEQ

The first paper demonstrating the feasibility of profiling the transcriptomes of individual mouse blastomeres and oocytes captured by micromanipulation was published in 2009 (Tang et al., 2009)—1 year after the introduction of bulk RNA-seq (Lister et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008). The early protocols for scRNA-seq were applied only

Received 11 December, 2018; revised 9 January, 2019; accepted 9 January, 2019; published online 12 February, 2019

eISSN: 0219-1032

to a small number of cells and suffered from a high level of technical noise resulting from inefficient reverse transcription (RT) and amplification (Ramskold et al., 2012; Sasagawa et al., 2013; Tang et al., 2009). These limitations of early protocols have been mitigated by two innovative barcoding approaches.

Cellular and molecular barcoding

The cell barcoding approach integrates a short cell barcode (CB) into cDNA at the early step of RT, first introduced in the single-cell tagged reverse transcription sequencing (STRT-seq) protocol (Islam et al., 2011). All cDNAs from cells are pooled for multiplexing, and downstream steps are carried out in a single tube, reducing reagent and labor costs. The cell barcoding approach was adopted to increase the number of cells in a plate-based or droplet-based platform. Early protocols relied on the plate-based platform, in which each cell is sorted into individual wells of a microplate, such as a 96- or 384-well plate, using fluorescence-activated cell sorting (FACS) or micropipettes (Hashimshony et al., 2012; Islam et al., 2011; Jaitin et al., 2014). Each well contains well-specific barcoded RT primers (Hashimshony et al., 2012; Jaitin et al., 2014) or barcoded oligonucleotides for template-switching PCR (Islam et al., 2011), and subsequent steps after RT are performed on pooled samples. In the droplet-based platform, encapsulating single cells in a nanoliter emulsion droplet containing lysis buffer and beads coated with barcoded RT primers was found to markedly increase the number of cells to tens of thousands in a single run (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017a).

The molecular barcoding approach for reducing amplification bias in PCR or in vitro transcription introduces a randomly synthesized oligonucleotide known as a unique molecular identifier (UMI) into RT primers (Islam et al., 2014). During RT, each cDNA is labeled with a UMI; thus, the number of cDNAs of a gene before amplification can be inferred by counting the number of distinct UMIs mapped to the gene, eliminating amplification bias.

Further improvements for sensitivity and throughput

These two barcoding strategies have become the standard in recently developed methods for scRNA-seq, which had already been improved compared with early protocols in terms of sensitivity and throughput. For most protocols, the sensitivity of recovering mRNA molecules present in a single cell is ~3-20% (Papalexi and Satija, 2018). Inefficient RT is responsible for such low capture rates; therefore, considerable effort has been devoted to increasing cDNA yield through optimization of RT enzymes (Hashimshony et al., 2016), buffer conditions (Picelli et al., 2013; Sasagawa et al., 2018), primers (Hashimshony et al., 2016; Picelli et al., 2013; Sasagawa et al., 2018), the subsequent amplification step (Bagnoli et al., 2018; Picelli et al., 2013), and reaction volume (Hashimshony et al., 2016). The most effective approach for improving sensitivity is to reduce the effective reaction volume, either by implementing nanoliter reactors in a microfluidics device (Hashimshony et al., 2016) or adding macromolecular crowding agents (Bagnoli et al., 2018). For example, the molecular crowding single-cell RNA barcoding and sequencing (mcSCRB-seq) protocol achieved 2.5-fold increase in sensitivity compared with its previous version by combining macromolecular crowding and optimized amplification (Bagnoli et al., 2018).

Increasing the number of cells to be profiled is essential for the unbiased characterization of cellular heterogeneity within a population of cells. Two different approaches have been developed to improve cell throughput in plate-based methods. In the first approach, instead of sorting each cell into an individual well of a microplate by FACS or manual picking, a cell suspension is randomly loaded into an array of ~100,000 microwells that accommodate one cell and one bead coated with barcoded RT primers (Gierahn et al., 2017; Han et al., 2018), increasing throughput in each experiment to tens of thousands of cells. In contrast to these approaches, which increase the number of wells in a microplate, a new approach was developed based on combinatorial cell barcoding (Cao et al., 2017; Rosenberg et al., 2018). In this technique, a suspension of cells passes through multiple rounds of split-pool barcoding in 96- or 384-well plates containing well-specific barcodes. In each round, fixed cells or nuclei are randomly loaded into individual wells and tagged with well-specific barcodes through RT, ligation, or amplification. The split-pool barcoding approach does not require a special device for making droplets or microwells, and can multiplex multiple samples in a single experiment by loading each sample into different subsets of wells at the first round of combinatorial cell barcoding. However, this approach can only be applied to permeabilized fixed cells or nuclei. For droplet-based methods, there is no upper limit on the number of cells that can be captured, at least in theory, but typically 1,000-10,000 cells are captured in one run reducing the probability of capturing two or more cells in a droplet (called "doublets"). If multiple samples labeled with unique molecular features are pooled and doublets are demultiplexed according to their molecular features, the throughput of cells can be increased, facilitating concurrent processing of multiple samples in a single experiment and minimizing technical batch effects of droplet-based methods. Several molecular features have been developed for demultiplexing doublets, including natural genetic variation of individuals (Kang et al., 2018) and lipid-modified oligonucleotides targeted to the plasma membrane (McGinnis et al., 2018).

Integration

To define the detailed molecular state of cells, we need to measure multiple molecular readouts and their interplay from the same single cell. Since the type and state of cells are usually defined by the cells' transcriptomes, and the protocols for profiling the single-cell transcriptome of polyadenylated mRNAs are the most developed among single-cell omics technologies, considerable effort has been applied to combining the single-cell transcriptome with other molecular readouts in the same single cell (Chappell et al., 2018). Several methods that simultaneously profile genomic DNA and mRNA from the same single cell, including DNA-RNA sequencing (DR-seq) (Dey et al., 2015) and genome and transcriptome sequencing (G&T-seq)(Macaulay et al., 2015),

have been developed for linking genomic variation with transcriptomic heterogeneity. DNA methylation (Angermueller et al., 2016; Hu et al., 2016) has also been integrated with the transcriptome to reveal the interplay between the epigenome and transcriptome at single-cell resolution. Recent single-cell multiomics methods have combined more than two genomic and epigenomic layers with the transcriptome. For example, single-cell triple-omics sequencing (scTrio-seq) profiles genomic copy number variation, DNA methylation, and the transcriptome of a single cell (Hou et al., 2016). Another method, scNMT-seq, combines the two epigenomic features of DNA methylation and chromatin accessibility with the transcriptome of a single cell (Clark et al., 2018). Single-cell multiomics technologies have not been applied to a large number of cells, because they require manually separating the transcriptome library from the genome or epigenome library. A recent method based on the split-pool barcoding approach integrated the transcriptome with chromatin accessibility in thousands of single cells, demonstrating the feasibility of high-throughput single-cell multiomics technologies (Cao et al., 2018).

The technologies for single-cell proteomics are still in their infancy because the methods for shotgun proteomics, such as liquid chromatography and tandem mass spectrometry (LC-MS/MS), require a large amount of input material and it is not possible to amplify proteins (Bantscheff et al., 2012; Budnik et al., 2018). Most protocols for single-cell protein quantification use high-affinity antibodies to measure the expression levels of a small number of targeted proteins. These antibodies are usually conjugated with fluorophores for flow cytometry (Perfetto et al., 2004), metal isotopes for mass cytometry (Spitzer and Nolan, 2016), or DNA barcode sequences for quantitative PCR or sequencing (Ullal et al., 2014). The idea of using DNA barcode-conjugated antibodies has been extended to develop methods for jointly profiling the transcriptome and expression levels of targeted cell surface proteins in single cells (Peterson et al., 2017; Stoeckius et al., 2017).

COMPUTATIONAL ANALYSIS OF SCRNA-SEQ DATA

As scRNA-seq has become a well-established method for dissecting cellular heterogeneity in complex tissues, the associated computational tools necessary for analyzing single-cell transcriptomic data continue to be designed and developed. As of November 2018, 325 tools have been deposited at the scRNA-tools database (www.scRNA-tools.org), and the number of tools being added is growing exponentially (Zappia et al., 2018). Compared with the analysis of bulk RNA-seq, scRNA-seq data analysis has several unique features. First, the gene-by-cell count matrix is very sparse owing to inefficient capture rates of mRNA molecules and low sequencing depth per cell, which results in higher technical variability in gene expression across cells. Second, tens of thousands of single cells are analyzed in a typical single-cell experiment, whereas the number of samples in bulk RNA-seq is usually three per condition, highlighting the importance of computational efficiency in tools for analyzing scRNA-seq data. Third, since the type and state of each cell are generally unknown, the expectation is that such information will be inferred from scRNA-seq data through unsupervised analysis, such as visualization and cell type identification. However,

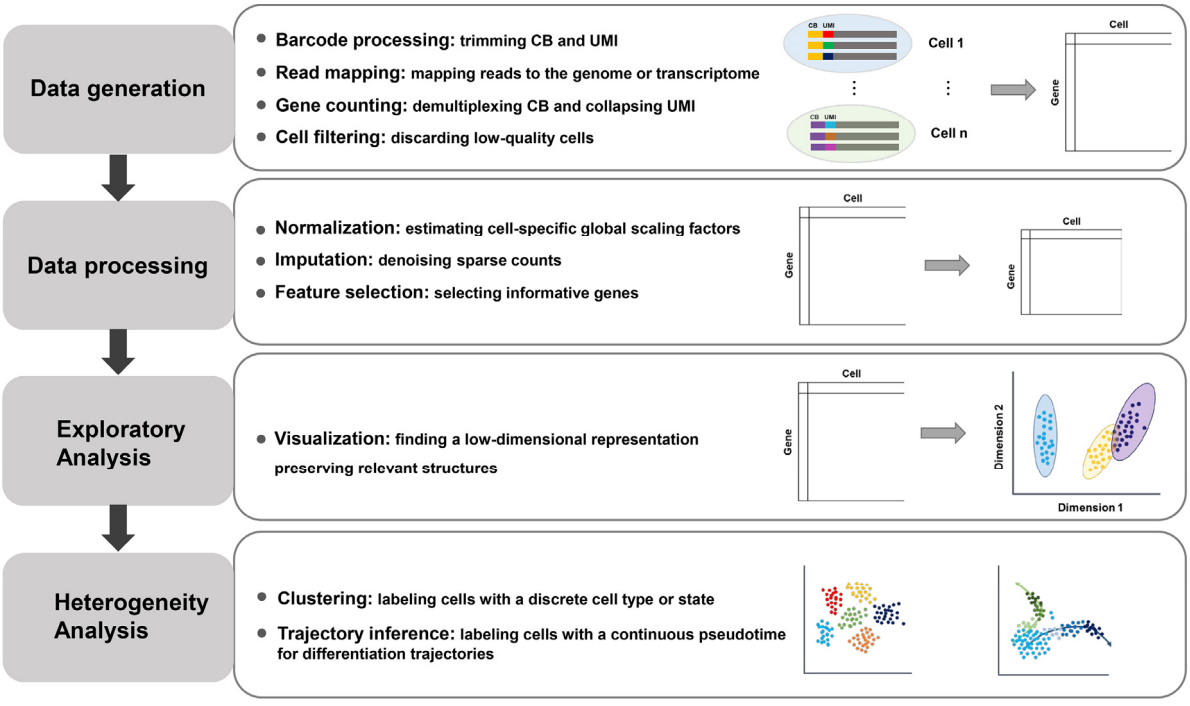

Fig. 1. Computational workflow for analyzing scRNA-seq data.

for bulk RNA-seq data, in which the class label of each sample is known a priori, genes that are differentially expressed between classes are usually identified through supervised analysis and hypothesis testing. Finally, there are single-cell-specific biological questions that cannot be addressed by bulk-level analysis. For example, it is possible to infer how individual tissue stem cells differentiate into multiple lineages during tissue homeostasis by estimating the ordering of cells along differentiation trajectories from a mixture of cells with heterogeneous differentiation states. The workflow of scRNA-seq data analysis includes four steps: data generation, data preprocessing, exploratory analysis, and heterogeneity analysis (Fig. 1).

Data generation: generating a count matrix

The basic pipeline for generating a gene-by-cell count matrix from high-throughput scRNA-seq data consists of four common steps: barcode processing, read mapping, gene counting, and cell filtering. Several tools have been developed for this purpose, including Cell Ranger (Zheng et al., 2017a), UMI-tools (Smith et al., 2017), umis (Svensson et al., 2017), ESAT (Derr et al., 2016), dropEst (Petukhov et al., 2018), scPipe (Tian et al., 2018) and zUMIs (Parekh et al., 2018). In the first step (barcode processing), we reformat each read pair in paired-end FASTQ files by trimming the CB and UMI from one read and adding this information to the sequence identifier line of the other read in the pair. Sequencing errors introduced into CBs and UMIs can optionally be corrected by filtering out read pairs with low-quality CBs and UMIs according to Phred quality scores. The reformatted reads are then mapped to the genome or transcriptome using any of the popular aligners developed for bulk RNA-seq data. Exon mapped reads from output BAM files are assigned to genes by a gene annotation GTF file and demultiplexed by CBs. For single-nuclei RNA-seq data, in which precursor mRNAs are abundant, both exon and intron mapped reads can be considered in gene counting to improve the number of detected genes (Parekh et al., 2018). PCR duplicates are removed by collapsing reads that are assigned to the same gene and share an identical UMI. Optionally, both sequencing and amplification errors in UMI sequences can be accounted for by collapsing UMIs if their edit distance is small and one UMI has a much higher read count than others. UMI-tools (Smith et al., 2017) uses a more elaborate method for UMI collapsing. It constructs UMI networks in which each node is labeled with a UMI sequence and read count, and two nodes are connected if their edit distance is 1. UMI collapsing is done by detecting modules in UMI networks based on adjacency and read counts.

After demultiplexing CBs and collapsing UMIs, a raw count matrix is obtained in which only a subset of CBs corresponds to intact cells. In plate-based protocols, CBs for intact cells can easily be identified and sequence errors in CBs can be corrected by comparing them with a list of known well-specific CBs. In droplet-based protocols, multiple heuristic methods have been proposed for filtering out CBs that correspond to empty droplets. The most popular method is to detect the threshold at the "knee point" in the barcode rank plot, where all cell barcodes are sorted by the total UMI counts in descending order. All CBs with a total UMI count less than the threshold are considered empty droplets and discarded (Macosko et al., 2015; Zheng et al., 2017b). Empty droplets contain cell-free transcripts in the cell suspension, which is the major source of non-zero total UMI counts for these CBs. A recent method has proposed a statistical framework for testing whether a CB is significantly different from cell-free transcript profiles, and combined this testing framework with the knee point method (Lun et al., 2018). This approach is implemented in DropletUtils (Lun et al., 2018) and Cell Ranger 3.0. If the expected number of cells is known, CBs can be discarded using a manually set threshold, and CBs corresponding to low-quality cells can be further filtered out based on multiple cell-level quality control (QC) metrics (Tian et al., 2018).

It is essential to discard low-quality cells, such as damaged or dying cells to avoid unwanted variation and misleading results in downstream analyses driven by these cells (Ilicic et al., 2016). Two types of cell-level QC features are widely used to distinguish low- from high-quality cells (Ilicic et al., 2016): (1) technical features that are proportional to total mRNA content, such as total UMI count, number of detected genes and proportion of reads mapped to spike-ins; and (2) biological features related with cell death or cell rupture, such as the proportion of reads that map to mitochondrial DNA. Although some methods use machine learning classifiers to automatically detect low-quality cells (Ilicic et al., 2016; Petukhov et al., 2018), the characteristics of low-quality cells are data-specific. Therefore, it is still recommended to visually inspect outliers corresponding to low-quality cells, with the aid of multiple diagnostic plots of cell-level QC metrics. Several tools, including scater (McCarthy et al., 2017) and scPipe (Tian et al., 2018), are available for computing QC metrics and visualizing them in diagnostic plots.

Data preprocessing: normalization, imputation, and feature selection

The next step is to estimate the true expression level of each gene in each cell by removing cell-specific biases in the gene-by-cell count matrix. The assumption in this analysis is that the expected count of a gene in a cell is proportional to the product of the relative expression level of the gene and the cell-specific global scaling factor. The global scaling factor represents cell-specific systematic biases affected by cell-to-cell differences in cell size, capture and RT efficiency, amplification factor, dilution factor, and sequencing depth (Vallejos et al., 2017). Cell-specific biases can be removed by normalizing the raw counts within each cell by a single scaling factor, applied to all genes in a cell. The cell-specific scaling factor can be estimated based on library size (e.g., reads per million (RPM) or transcripts per kilobase million (TPM)(Li et al., 2010)), upper quantile values of counts (Bullard et al., 2010), or normalization factors (e.g., size factor of DESeq (Anders and Huber, 2010) or trimmed mean of M-value of edgeR (Robinson and Oshlack, 2010)), developed for bulk RNA-seq normalization. However, normalization by library size is sensitive to a few highly expressed genes, and the

other normalization methods are problematic for sparse scRNA-seq data, since estimated scaling factors are unstable and inaccurate owing to zero inflation (Vallejos et al., 2017). Several normalization methods have been proposed for robustly estimating the cell-specific scaling factors in the presence of excessive zero counts (Lun et al., 2016a; Vallejos et al., 2015). For example, scran estimates pooled size factors from a pool of cells by summing expression values across these cells and then deconvolves the pooled size factors obtained from multiple pools to their cell-specific size factors (Lun et al., 2016a).

A high frequency of zero counts, which is driven by stochastic gene expression (Kim and Marioni, 2013), low mRNA capture efficiency and low sequencing depth, is a key characteristic of high-throughput scRNA-seq data. This zero inflation leads to high technical variability in gene expression, an effect that should be carefully accounted for in downstream analyses requiring accurate measurements of gene expression. Because global scaling normalization methods are unable to address this issue, computational approaches that recover the true expression levels of zero counts have been proposed (Chen and Zhou, 2018; Huang et al., 2018; Li and Li, 2018; van Dijk et al., 2018). These imputation methods take a normalized count matrix (usually log-transformed) as input and replace input data with de-noised values, estimated by borrowing information across similar cells (Chen and Zhou, 2018; Li and Li, 2018; van Dijk et al., 2018) or genes (Huang et al., 2018). These imputed expression values can be used to recover regulatory interactions between genes (Huang et al., 2018; van Dijk et al., 2018), increase the accuracy of estimates of cell-to-cell variability in gene expression (Huang et al., 2018), and improve cell clustering and differential gene expression analysis (Chen and Zhou, 2018; Huang et al., 2018; Li and Li, 2018). However, despite the potential of these imputation methods to recover true expression levels, it should be noted that all such methods introduce unexpected biases, including spurious gene-to-gene correlations, artificial cell subpopulation structure, and removal of rare cell types and transient cell states. Because these biases have not been rigorously examined, imputation should be applied with caution and is not included in the general workflow for scRNA-seq data analysis.

The normalized count matrix contains many genes whose expression levels are associated with a high level of technical noise. These genes mask the reliable detection of different cell types and states within a heterogeneous population of cells. It is necessary to filter out such genes to improve the extraction of biologically interesting patterns in the scRNA-seq data, a process known as feature selection. The most widely used approach is to evaluate the biological cell-to-cell variability in the expression of each gene, and then take genes showing significantly high biological variability as input in downstream unsupervised analyses such as visualization and clustering (Brennecke et al., 2013; Lun et al., 2016b; Vallejos et al., 2015). The key idea in evaluating biological variability is to decompose the observed variance of gene expression levels into its technical and biological components according to the law of total variance. To estimate the technical variability, we assume that the mean technical variance of each gene is a nonlinear function of its mean expression level. The nonlinear function can be estimated by fitting a curve to the mean-variance data of external RNA spike-ins (Brennecke et al., 2013; Kim et al., 2015; Vallejos et al., 2015) or all endogenous genes, under the assumption that the observed variance of most genes is dominated by technical noise (Kolodziejczyk et al., 2015b; Lun et al., 2016b). By subtracting the estimated technical variance from the observed variance, we can estimate the biological variance and choose highly variable genes that show significant non-zero biological variance.

Exploratory analysis: dimensionality reduction

By selecting informative genes, such as highly variable genes, the dimension of scRNA-seq data is reduced to the number of chosen genes, but the results still suffer from high dimensionality, which makes it difficult to comprehend and visualize the patterns of cellular heterogeneity. Dimensionality reduction is performed to find a low-dimensional representation that preserves the relevant structure of the original high-dimensional data. In the context of scRNA-seq data analyses, two different relevant structures are considered: a local structure that preserves cell-to-cell distance within a local neighborhood of cells, and a global structure that preserves cell-to-cell distance on the low-dimensional manifold associated with the underlying biological process. Capturing local structure in a low-dimensional representation is important for clustering cells of the same type or state close together. In contrast, capturing global structure is useful for preserving distance between clusters and revealing underlying biological processes for cell-to-cell variability in gene expression. Principal component analysis (PCA), a linear method used for dimensionality reduction, projects high-dimensional data onto a low-dimensional linear space by maximizing the variance of the projected data. PCA is also a popular method for data pre-processing since it removes redundancies among genes owing to its orthogonal linear projection. Many dimensionality reduction methods use PCA as a preprocessing step to reduce distortions incurred because of irrelevant dimensions in the calculation of pairwise distances between cells.

Although PCA has been successfully applied to capture the global structure of cellular heterogeneity in low-throughput scRNA-seq data (Brennecke et al., 2013; Hashimshony et al., 2012; Picelli et al., 2013; Shalek et al., 2013), it is limited by its frequent failure to visualize the local structure essential for cell clustering and cell type identification. This issue was addressed by introducing t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) to the field of single-cell genomics (Amir et al., 2013). t-SNE is a nonlinear dimensionality reduction method for capturing the local structure in which dissimilar cells in the original high-dimensional space are modeled by large distances, and similar cells are modeled by small distances. Thus, t-SNE generates a low-dimensional representation in a two- or three-dimensional space displaying multiple isolated clusters. However, global structures, such as the distance between clusters, are not well captured in the t-SNE map. The current state-of-the-art method for dimensionality reduction that

captures both local and global structure in scRNA-seq data is uniform manifold approximation and projection (UMAP) (Becht et al., 2018; McInnes et al., 2018). It has been shown that UMAP is able to arrange clusters along differentiation trajectories and preserve a differentiation continuum of transient cells (Becht et al., 2018). Understanding the captured local and global structure in the low-dimensional representation can be facilitated by overlaying the expression of a marker gene or the activity of a set of genes associated with a biological process of interest on the two- or three-dimensional map, a step that is useful for exploratory data analysis.

Heterogeneity analysis: clustering and trajectory inference

Two computational approaches for dissecting cellular heterogeneity in scRNA-seq data have been developed based on the assumption that a latent variable generates the observed cell-to-cell variability: 1) a discrete latent variable approach that labels each cell with a discrete cluster indicator for cell type or state, and 2) a continuous latent variable approach that labels each cell with a continuous pseudotime for differentiation trajectories The correct reference is (Wagner et al., 2016).

The discrete latent variable approach can be formulated as an unsupervised clustering problem which has been extensively studied in the field of statistics and machine learning. Diverse clustering algorithms, such as k-means, hierarchical, density-based, and graph-based clustering, have been applied to identify cell clusters in scRNA-seq data (Andrews and Hemberg, 2018; Kiselev et al., 2017; Satija et al., 2015). A number of considerations should be taken into account to ensure that each cluster is associated with a distinct cell type or state. First, selecting genes showing differential expression across multiple cell types is essential for improving the quality of clustering results. Such relevant genes can be identified by selecting genes that are highly variable across cells. Both feature selection and dimensionality reduction (e.g., PCA and t-SNE) can be sequentially applied to extract informative features that are taken as input to clustering algorithms (Andrews and Hemberg, 2018; Duo et al., 2018). Second, because the optimal number of clusters is dependent on the definition of cell types or states and subjective clustering resolution, it cannot be generally estimated from data. It is generally recommended that the number of clusters should be chosen by a user with domain-specific knowledge. Third, identifying rare cell types, such as stem cells and short-lived progenitors, in a heterogeneous population requires careful examination of outliers within a large cluster (Grun et al., 2015) or selection of genes that are specifically expressed in a minor population of cells as features (Jiang et al., 2016). Fourth, if samples are processed in multiple batches and technical batch effects largely account for the observed variability, batch effects should be adjusted while preserving global structure. If the biological condition is not confounded by batch information, regression-based batch correction methods originally designed for bulk RNA-seq can be applied (Buttner et al., 2017; Kolodziejczyk et al., 2015b). However, in a confounded design, which is common in the droplet-based protocols, the batch correction methods regress out both biological and technical variability. One solution is to project the expression profile of each cell to a feature space by calculating the correlation coefficient between the expression vector of single cells and the expression vector of the reference bulk panel of diverse cell types (Li et al., 2017). Although this approach improves clustering accuracy in the presence of batch effects, obtaining a reference panel that contains all cell types of single cells is not straightforward. A more general strategy is to merge multiple scRNA-seq data with shared subpopulations using canonical correlation analysis (Butler et al., 2018) or by identifying mutual nearest neighbors (Haghverdi et al., 2018).

Finally, the identified clusters are annotated as cell types or states using the expression of known marker genes. To automate this annotation, researchers have developed correlation-based scoring methods (Aran et al., 2019; Kiselev et al., 2018) or machine learning classifiers (Alavi et al., 2018; Alquicira-Hernandez et al., 2018) with the aid of reference bulk transcriptomes (Aran et al., 2019) or reference single-cell transcriptomes (Alavi et al., 2018; Alquicira-Hernandez et al., 2018; Kiselev et al., 2018). The identity of cell clusters can also be inferred by examining differentially expressed genes across cell clusters and their enriched functional categories of genes. Although statistical methods designed for differential expression analysis in scRNA-seq have been developed (Finak et al., 2015; Kharchenko et al., 2014), their performance is comparable or sometimes inferior to methods designed for bulk RNA-seq or general purpose two-sample tests, such as the t-test and Wilcoxon rank sum test (Soneson and Robinson, 2018).

The continuous latent variable approach, pioneered by Monocle (Trapnell et al., 2014), is referred to as trajectory inference or pseudotemporal ordering. The main assumption underlying this approach is that there exists a dynamic cellular process that shapes the transcriptional landscape and each individual cell can be placed along the process. Many dynamic cellular processes, including differentiation (Velten et al., 2017), reprogramming (Treutlein et al., 2016), and cell cycling (Kowalczyk et al., 2015), continuously progress along single or multiple trajectories, passing through transient cell states. The temporal progression of each cell along these trajectories, termed pseudotime, is the continuous latent variable that is inferred from data. If a large number of cells covering transient states are sampled from a mixed population of cells whose cell-to-cell variability is largely driven by a given cellular process, trajectories can be accurately reconstructed. Over the last 4 years, more than 60 computational tools have been developed for pseudotemporal ordering (Zappia et al., 2018). Most of these tools operate based on the assumption that cells showing similar expression profiles should be placed close together on the same trajectories (Kester and van Oudenaarden, 2018). They use a recurring framework that consists of two steps: 1) constructing a low-dimensional representation of cells, and 2) modeling trajectories with graphs or curves in the low-dimensional representation (Cannoodt et al., 2016).

In the first step, two different classes of representation are used: (1) a two- or three-dimensional feature space generated using dimensionality reduction algorithms, and (2) a k-

nearest neighbor graph (k-NNG) in which each cell is represented as a node and each node is linked with its k nearest neighbors. The low-dimensional feature space can be constructed by applying diverse dimensionality reduction algorithms, including PCA (Shin et al., 2015), independent component analysis (Trapnell et al., 2014), t-SNE (Marco et al., 2014), diffusion map (Haghverdi et al., 2016), or UMAP (Becht et al., 2018), after selecting genes relevant to the cellular process of interest. In principle, algorithms that preserve the global structure in the low-dimensional feature space, such as diffusion map and UMAP, should be used. The k-NNG is usually constructed after projecting cells to the low-dimensional feature space using dimensionality reduction methods (Bendall et al., 2014; Setty et al., 2016). For better visualization, k-NNGs can be arranged in a two-dimensional space using the force-directed layout embedding (Briggs et al., 2017; Schiebinger et al., 2017). For feature selection, there is no consensus on the best practice for selecting genes that are informative with respect to constructing the low-dimensional representation. Widely used criteria for this process include highly expressed genes, highly variable genes across cells, differentially expressed genes across cell clusters (Qiu et al., 2017; Trapnell et al., 2014), genes that show gradual changes within a local neighborhood (Welch et al., 2016), and a set of known genes related to the cellular process.

In the second step of modeling trajectories, a backbone of trajectories is constructed with graphs or curves in the low-dimensional representation, and then the pseudotime of cells is evaluated by projecting cells onto the backbone. Constructing the backbone, which usually requires prior information, such as the structure of trajectories and a root cell with a pseudotime of 0, is the key step for determining the accuracy of inferred trajectories. Early methods fixed the structure of trajectories as linear(Bendall et al., 2014; Shin et al., 2015) or bifurcating (Haghverdi et al., 2016; Setty et al., 2016). A more complex structure of trajectories is difficult to correctly reconstruct from data, since it becomes more sensitive to outlier cells, requires more prior information, and needs sampling of a sufficient number of cells. The most widely used strategy for addressing this issue is to group cells into clusters that represent distinct cell types or states. The backbone is constructed by linking clusters, and the trajectories are inferred by specifying the start clusters (Street et al., 2018), both start and end clusters (Lummertz da Rocha et al., 2018), or all clusters on a given trajectory (Wolf et al., 2018). Several methods for identifying the least differentiated cells (or stem cells) have been proposed for facilitating construction of the backbone (Grun et al., 2016; Teschendorff and Enver, 2017). In addition, the direction and the speed of differentiation can be inferred from RNA velocity, but this is sensitive to the set of input genes (La Manno et al., 2018). After reconstructing trajectories, the dynamics of gene regulation along the inferred trajectories can be analyzed (Aibar et al., 2017).

FUTURE DEVELOPMENTS

Over the past decade, technologies for single-cell transcriptomics have emerged as essential tools for dissecting cellular heterogeneity in individual tissues. Rapid technological advances are expected to expand the breadth and depth of the application of scRNA-seq. Comprehensive transcriptomic reference maps of all cell types in the body of diverse organisms, including humans (Luo et al., 2017) and mice (Han et al., 2018; Tabula Muris et al., 2018), are being constructed to provide a systematic framework for understanding the molecular characteristics of cell types or states, cellular trajectories and molecular mechanisms of development and differentiation, and regulatory interactions between cells. A more in-depth single-cell transcriptomic analysis that profiles non-mRNA species, such as microRNAs (Faridani et al., 2016) or full-length mRNA isoforms (Gupta et al., 2018), within a single cell is also being actively developed. Integrating the transcriptome with multiple omics (Chappell et al., 2018), genotypes (Dixit et al., 2016; Jaitin et al., 2016), cellular phenotypes (Cadwell et al., 2016; Fuzik et al., 2016), lineage tracing (Kester and van Oudenaarden, 2018), and spatial information (Lein et al., 2017) within the same cell is another active area of ongoing research. In parallel with technological advances, computational methods that integrate diverse molecular and cellular information from the same cell and infer hidden biological structures from large-scale single-cell data should be developed.

ACKNOWLEDGMENTS

This work was supported by grants from the National Research Foundation of Korea funded by the Ministry of Science, ICT and Future Planning (2017R1C1B2007843, 2017M3C7A1048448, 2017M3A9B6073099, 2017M3A9D5A01052447) and from Business for Cooperative R&D between Industry, Academy, and Research Institute funded by the Ministry of SMEs and Startups (C0452791).

REFERENCES

Aibar, S., Gonzalez-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083-1086.

Alavi, A., Ruffalo, M., Parvangada, A., Huang, Z., and Bar-Joseph, Z. (2018). A web server for comparative analysis of single-cell RNA-seq data. Nat. Commun. *9*, 4768.

Alquicira-Hernandez, J., Nguyen, Q., and Powell, J.E. (2018). scPred: scPred: Cell type prediction at single-cell resolution. bioRxiv, 369538.

Amir, E.D., Davis, K.L., Tadmor, M.D., Simonds, E.F., Levine, J.H., Bendall, S.C., Shenfeld, D.K., Krishnaswamy, S., Nolan, G.P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat. Biotechnol. *31*, 545-552.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Andrews, T.S., and Hemberg, M. (2018). Identifying cell populations with scRNASeq. Mol. Aspects Med. *59*, 114-122.

Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S., Ponting, C.P., Voet, T., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat. Methods *13*, 229-232.

Aran, D., Looney, A.P., Liu, L., Fong, V., Hsu, A., Wolters, P.J., Abate, A., Butte, A.J., and Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163-172.

Bagnoli, J.W., Ziegenhain, C., Janjic, A., Wange, L.E., Vieth, B., Parekh, S., Geuder, J., Hellmann, I., and Enard, W. (2018). Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. Nat. Commun. *9*, 2937.

Bantscheff, M., Lemeer, S., Savitski, M.M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal. Bioanal. Chem. *404*, 939-965.

Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38-44.

Bendall, S.C., Davis, K.L., Amir el, A.D., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell *157*, 714-725.

Brennecke, P., Anders, S., Kim, J.K., Kolodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods *10*, 1093-1095.

Briggs, J.A., Li, V.C., Lee, S., Woolf, C.J., Klein, A., and Kirschner, M.W. (2017). Mouse embryonic stem cells can differentiate via multiple paths to the same state. Elife *6*, e26945.

Budnik, B., Levy, E., Harmange, G., and Slavov, N. (2018). SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. Genome Biol. *19*, 161.

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics *11*, 94.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411-420.

Buttner, M., Miao, Z., Wolf, A., Teichmann, S.A., and Theis, F.J. (2017). Assessment of batch-correction methods for scRNA-seq data with a new test metric. bioRxiv, 200345.

Cadwell, C.R., Palasantza, A., Jiang, X., Berens, P., Deng, Q., Yilmaz, M., Reimer, J., Shen, S., Bethge, M., Tolias, K.F., et al. (2016). Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. Nat. Biotechnol. *34*, 199-203.

Cannoodt, R., Saelens, W., and Saeys, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. Eur. J. Immunol. *46*, 2496-2506.

Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., et al. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science *361*, 1380-1385.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. Science *357*, 661-667.

Chappell, L., Russell, A.J.C., and Voet, T. (2018). Single-cell (multi)omics technologies. Annu. Rev. Genomics Hum. Genet. *19*, 15-41.

Chen, M., and Zhou, X. (2018). VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. Genome Biol. *19*, 196.

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.

Clark, S.J., Argelaguet, R., Kapourani, C.A., Stubbs, T.M., Lee, H.J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J.C., et al. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat. Commun. *9*, 781.

Derr, A., Yang, C., Zilionis, R., Sergushichev, A., Blodgett, D.M., Redick, S., Bortell, R., Luban, J., Harlan, D.M., Kadener, S., et al. (2016). End sequence analysis toolkit (ESAT) expands the extractable information from single-cell RNA-seq data. Genome Res. *26*, 1397-1410.

Dey, S.S., Kester, L., Spanjaard, B., Bienko, M., and van Oudenaarden, A. (2015). Integrated genome and transcriptome sequencing of the same cell. Nat. Biotechnol. *33*, 285-289.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell *167*, 1853-1866 e1817.

Duo, A., Robinson, M.D., and Soneson, C. (2018). A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Res *7*, 1141.

Eldar, A., and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. Nature *467*, 167-173.

Faridani, O.R., Abdullayev, I., Hagemann-Jensen, M., Schell, J.P., Lanner, F., and Sandberg, R. (2016). Single-cell sequencing of the small-RNA transcriptome. Nat. Biotechnol. *34*, 1264-1266.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., 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.

Fuzik, J., Zeisel, A., Mate, Z., Calvigioni, D., Yanagawa, Y., Szabo, G., Linnarsson, S., and Harkany, T. (2016). Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. Nat. Biotechnol. *34*, 175-183.

Gierahn, T.M., Wadsworth, M.H., 2nd, Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Love, J.C., and Shalek, A.K. (2017). Seq-well: portable, low-cost RNA sequencing of single cells at high throughput. Nat. Methods *14*, 395-398.

Grun, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature *525*, 251-255.

Grun, D., Muraro, M.J., Boisset, J.C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., van Es, J., Jansen, E., Clevers, H., et al. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. Cell Stem Cell *19*, 266-277.

Gupta, I., Collier, P.G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A.B., Sloan, S.A., et al. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. Nat. Biotechnol. *36*, 1197-1202.

Haghverdi, L., Buttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods *13*, 845-848.

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.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A.,

Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the mouse cell atlas by microwell-seq. Cell *172*, 1091-1107.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K.J., Rozenblatt-Rosen, O., et al. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-seq. Genome Biol. *17*, 77.

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.

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y., et al. (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. *26*, 304-319.

Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., Zhu, X., Wang, C.Y., Xue, Z., and Fan, G. (2016). Simultaneous profiling of transcriptome and DNA methylome from a single cell. Genome Biol. *17*, 88.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J.I., Raj, A., Li, M., and Zhang, N.R. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. Nat. Methods *15*, 539-542.

Ilicic, T., Kim, J.K., Kolodziejczyk, A.A., Bagger, F.O., McCarthy, D.J., Marioni, J.C., and Teichmann, S.A. (2016). Classification of low quality cells from single-cell RNA-seq data. Genome Biol. *17*, 29.

Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J.B., Lonnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. *21*, 1160-1167.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lonnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163-166.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776-779.

Jaitin, D.A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T.M., Tanay, A., van Oudenaarden, A., and Amit, I. (2016). Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. Cell *167*, 1883-1896.

Jiang, L., Chen, H., Pinello, L., and Yuan, G.C. (2016). GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. *17*, 144.

Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., et al. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol. *36*, 89-94.

Kester, L., and van Oudenaarden, A. (2018). Single-cell transcriptomics meets lineage tracing. Cell Stem Cell *23*, 166-179.

Kharchenko, P.V., Silberstein, L., and Scadden, D.T. (2014). Bayesian approach to single-cell differential expression analysis. Nat. Methods *11*, 740-742.

Kim, J.K., Kolodziejczyk, A.A., Ilicic, T., Teichmann, S.A., and Marioni, J.C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat. Commun. *6*, 8687.

Kim, J.K., and Marioni, J.C. (2013). Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol. *14*, R7.

Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. Nat. Methods *14*, 483-486.

Kiselev, V.Y., Yiu, A., and Hemberg, M. (2018). Scmap: projection of single-cell RNA-seq data across data sets. Nat. Methods *15*, 359-362.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187-1201.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015a). The technology and biology of single-cell RNA sequencing. Mol. Cell *58*, 610-620.

Kolodziejczyk, A.A., Kim, J.K., Tsang, J.C., Ilicic, T., Henriksson, J., Natarajan, K.N., Tuck, A.C., Gao, X., Buhler, M., Liu, P., et al. (2015b). Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell *17*, 471-485.

Kowalczyk, M.S., Tirosh, I., Heckl, D., Rao, T.N., Dixit, A., Haas, B.J., Schneider, R.K., Wagers, A.J., Ebert, B.L., and Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res. *25*, 1860-1872.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lonnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494-498.

Lein, E., Borm, L.E., and Linnarsson, S. (2017). The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. Science *358*, 64-69.

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2010). RNA-seq gene expression estimation with read mapping uncertainty. Bioinformatics *26*, 493-500.

Li, H., Courtois, E.T., Sengupta, D., Tan, Y., Chen, K.H., Goh, J.J.L., Kong, S.L., Chua, C., Hon, L.K., Tan, W.S., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat. Genet. *49*, 708-718.

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.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell *133*, 523-536.

Lummertz da Rocha, E., Rowe, R.G., Lundin, V., Malleshaiah, M., Jha, D.K., Rambo, C.R., Li, H., North, T.E., Collins, J.J., and Daley, G.Q. (2018). Reconstruction of complex single-cell trajectories using CellRouter. Nat. Commun. *9*, 892.

Lun, A., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes, T., and Marioni, J.C. (2018). Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. BioRxiv. 234872.

Lun, A.T., Bach, K., and Marioni, J.C. (2016a). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. *17*, 75.

Lun, A.T., McCarthy, D.J., and Marioni, J.C. (2016b). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res *5*, 2122.

Luo, C., Keown, C.L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J.R., Sandoval, J.P., et al. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science *357*, 600-604.

Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., Goolam, M., Saurat, N., Coupland, P., Shirley, L.M., et al. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat. Methods *12*, 519-522.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman,

M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202-1214.

Marco, E., Karp, R.L., Guo, G., Robson, P., Hart, A.H., Trippa, L., and Yuan, G.C. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proc. Natl. Acad. Sci. USA *111*, E5643-5650.

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.

McGinnis, C.S., Patterson, D.M., Winkler, J., Hein, M.Y., Srivastava, V., Conrad, D.N., Murrow, L.M., Weissman, J.S., Werb, Z., Chow, E.D., et al. (2018). MULTI-seq: scalable sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. bioRxiv, 387241.

Mclnnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. arXiv. 1802.03426.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621-628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344-1349.

Papalexi, E., and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. Nat. Rev. Immunol. *18*, 35-45.

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., and Hellmann, I. (2018). zUMIs: a fast and flexible pipeline to process RNA sequencing data with UMIs. Gigascience *7*.

Perfetto, S.P., Chattopadhyay, P.K., and Roederer, M. (2004). Seventeen-colour flow cytometry: unravelling the immune system. Nat. Rev. Immunol. *4*, 648-655.

Peterson, V.M., Zhang, K.X., Kumar, N., Wong, J., Li, L., Wilson, D.C., Moore, R., McClanahan, T.K., Sadekova, S., and Klappenbach, J.A. (2017). Multiplexed quantification of proteins and transcripts in single cells. Nat. Biotechnol. *35*, 936-939.

Petukhov, V., Guo, J., Baryawno, N., Severe, N., Scadden, D.T., Samsonova, M.G., and Kharchenko, P.V. (2018). dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. Genome Biol. *19*, 78.

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.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods *14*, 979-982.

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216-226.

Ramskold, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol *30*, 777-782.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science *360*, 176-182.

Sasagawa, Y., Danno, H., Takada, H., Ebisawa, M., Tanaka, K., Hayashi, T., Kurisaki, A., and Nikaido, I. (2018). Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. Genome Biol. *19*, 29.

Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K.D., Imai, T., and Ueda, H.R. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol. *14*, 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.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Liu, S., Lin, S., Berube, P., Lee, L., et al. (2017). Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. bioRxiv, 191056.

Setty, M., Tadmor, M.D., Reich-Zeliger, S., Angel, O., Salame, T.M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe'er, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat. Biotechnol. *34*, 637-645.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236-240.

Shin, J., Berg, D.A., Zhu, Y., Shin, J.Y., Song, J., Bonaguidi, M.A., Enikolopov, G., Nauen, D.W., Christian, K.M., Ming, G.L., et al. (2015). Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis. Cell Stem Cell *17*, 360-372.

Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. Genome Res. *27*, 491-499.

Soneson, C., and Robinson, M.D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. Nat. Methods *15*, 255-261.

Spitzer, M.H., and Nolan, G.P. (2016). Mass cytometry: single cells, many features. Cell *165*, 780-791.

Stegle, O., Teichmann, S.A., and Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. *16*, 133-145.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods *14*, 865-868.

Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics *19*, 477.

Svensson, V., Natarajan, K.N., Ly, L.H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A., and Teichmann, S.A. (2017). Power analysis of single-cell RNA-sequencing experiments. Nat. Methods *14*, 381-387.

Tabula Muris, C., Overall, c., Logistical, c., Organ, c. p.,, Library, p. s.,, Computational data, a., Cell type a., Writing, g., and Principle i. (2018). Single-cell ranscriptomics of 20 mouse organs creates a Tabula Muris. Nature *562*, 367-372.

Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. Nature *541*, 331-338.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377-382.

Teschendorff, A.E., and Enver, T. (2017). Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nat. Commun. *8*, 15599.

Tian, L., Su, S., Dong, X., Amann-Zalcenstein, D., Biben, C., Seidi, A., Hilton, D.J., Naik, S.H., and Ritchie, M.E. (2018). scPipe: a flexible R/bioconductor preprocessing pipeline for single-cell RNA-sequencing data. PLoS Comput. Biol. *14*, e1006361.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381-386.

Treutlein, B., Lee, Q.Y., Camp, J.G., Mall, M., Koh, W., Shariati, S.A., Sim, S., Neff, N.F., Skotheim, J.M., Wernig, M., et al. (2016). Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. Nature *534*, 391-395.

Ullal, A.V., Peterson, V., Agasti, S.S., Tuang, S., Juric, D., Castro, C.M., and Weissleder, R. (2014). Cancer cell profiling by barcoding allows multiplexed protein analysis in fine-needle aspirates. Sci. Transl. Med. *6*, 219ra219.

Vallejos, C.A., Marioni, J.C., and Richardson, S. (2015). BASiCS: bayesian analysis of single-cell sequencing data. PLoS Comput. Biol. *11*, e1004333.

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.

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., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. Cell *174*, 716-729 e727.

Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. *19*, 271-281.

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. Nat. Biotechnol. *34*, 1145-1160.

Welch, J.D., Hartemink, A.J., and Prins, J.F. (2016). SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. Genome Biol. *17*, 106.

Wolf, F.A., Hamey, F., Plass, M., Solana, J., Dahlin, J.S., Gottgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2018). Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. bioRxiv, 208819.

Zappia, L., Phipson, B., and Oshlack, A. (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput. Biol. *14*, e1006245.

Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017a). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.

Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017b). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.