REVIEW ARTICLE

Computational approaches for interpreting scRNA-seq data

Raghd Rostom[1], Valentine Svensson[2], Sarah A. Teichmann[1] and Gozde Kar[2]

1 Wellcome Trust Sanger Institute, Cambridge, UK
2 The European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

Correspondence

G. Kar, The European Bioinformatics
Institute (EMBL-EBI), Cambridge, UK
Fax: +44 01223 494468
Tel: +44 01223 834244
E-mail: gkar@ebi.ac.uk

The recent developments in high-throughput single-cell RNA sequencing technology (scRNA-seq) have enabled the generation of vast amounts of transcriptomic data at cellular resolution. With these advances come new modes of data analysis, building on high-dimensional data mining techniques. Here, we consider biological questions for which scRNA-seq data is used, both at a cell and gene level, and describe tools available for these types of analyses. This is an exciting and rapidly evolving field, where clustering, pseudotime inference, branching inference and gene-level analyses are particularly informative areas of computational analysis.

Keywords: single-cell analysis methods and tools; single-cell genomics

While transcriptomic studies have, for many years, provided insight into mRNA expression and regulation, technological advances have allowed the quantification of transcripts at an unprecedented resolution. By sequencing the mRNA component of individual single cells, it has now become possible to study gene expression at an entirely new level, opening the door to novel biological questions which were not possible using population-level RNA sequencing. For example, the variability in splicing [1–5] and allelic expression [3,6–8] between cells has been shown, along with analysis of the stochastic gene expression and transcriptional kinetics [9,10]. Furthermore, single-cell RNA-sequencing (scRNA-seq) data have allowed fine-grained analysis of developmental trajectories [11–13] and identification of rare cell types [14,15].

In order to obtain scRNA-seq data, cells must first be isolated individually in an accurate and rapid manner. Initially, microscopic manipulation provided a reliable method to isolate single cells through physical separation using a capillary pipette, and may still play an important role in systems where few cells are available. However, the high labour and low-throughput nature of this technique has resulted in it being surpassed in much current research by higher throughput methods. Fluorescence-activated cell sorting (FACS) provides an efficient way to isolate a large number of cells in a rapid manner, and also allows the possibility of labelling cells with multiple fluorescent proteins. Size or marker selection is commonly used, and through 'index sorting', the data for each cell can be recorded as a reference in downstream analysis. Despite the prevalence of this method, the high number of starting cells required, along with the potential damage caused by the staining and physical stress of the process, means it may be a problematic approach. More recently, microfluidics have emerged as a key method for capturing single cells, allowing isolation in small volumes within a closed system, often followed directly by amplification and downstream reactions. The small volume in which these reactions occur increases the capture efficiency and lowers the reagent

cost. Finally, techniques involving the isolation of single cells in microdroplets, such as DropSeq [16] and InDrop [17], have rapidly expanded the high-throughput nature of scRNA-seq – allowing processing of tens of thousands of cells in a short space of time. The small volume of reactions, once again, decreases the cost per cell. Over time, these methods will continue to increase in speed, efficiency and reliability, further improving throughput of single-cell isolation.

Many protocols exist for the subsequent reverse transcription (RT), amplification, and library preparation prior to sequencing. Poly(T) priming is used to select polyadenylated mRNA for reverse transcription, however, only an estimated 10–20% of transcripts are sampled. This produces a lot of noise at the RT stage, and particularly affects lowly expressed genes [18]. Methods then differ in their approach to second-strand synthesis, either using poly(A) tailing, leading to a 3′ bias, or template-switching to produce full-transcript coverage. Amplification can be achieved through two methods: linear *in vitro* transcription (IVT) or exponential PCR, each with its own advantages and drawbacks. Ziegenhain *et al.* [19] and Svensson *et al.* [20] provide a comprehensive experimental and computational comparison of most of the protocols commonly used. Following cDNA amplification, library preparation is most commonly carried out using the commercially available Nextera kit and sequencing on the Illumina platform, although other methods are available.

As a relatively new field, it is key to understand the structure and complexities of scRNA-seq data, ensuring that appropriate analytical and statistical methods are applied [21]. Particularly challenging is the high level of noise [22], which derives primarily from the nature of single-cell experiments (called 'technical variation' and is mainly due to factors such as mRNA capture efficiency and cDNA amplification bias), along with the biological heterogeneity of cells ('biological variation'). Furthermore, unlike with conventional RNA-sequencing where experimental biases are well studied [23,24], there are biases which are still not fully understood in single-cell experiments, such as 'dropouts' due to the low amounts of starting material, leading to false negative expression.

Single-cell RNA-sequencing is a lossy technique, and it is not completely understood what causes the different failure modes for samples. Practically, this means the first step after acquiring reads from a scRNA-seq experiment is to perform quality control. Reads are processed in a similar manner to bulk RNA-seq, allowing expression quantification. It is important to check the quality of both the raw data (which can be performed using tools developed for bulk RNA-seq, such as FastQC [25]

or Kraken [26]), along with the aligned output. Imperative in scRNA-seq is the cell-by-cell quality control [27,28], ensuring that cells of poor quality are removed from subsequent analysis. Many metrics can be used to measure cell quality, such as the number of reads or genes detected, the proportion of reads mapping to mitochondrial genes (which may signify leaking of cytoplasmic RNA or cells undergoing apoptosis), or the proportion of reads mapping to externally spiked-in RNA molecules if used in the experiment [29].

Depending on the analysis task, appropriate normalization of the data is needed. Several normalization methods have been developed, many of which adjust for differences in sequencing depth and/or make use of spike-in molecules and/or unique molecular identifiers (UMIs) when available (reviewed in detail in [30]).

Once cleaned data are obtained, there are many routes of analysis depending on the biological question under investigation (Fig. 1). In this review, we will consider these analysis from two viewpoints: cell-level approaches, such as the grouping of cells and trajectory ordering, along with gene-level investigations, such as gene variability and noise, coexpression and identification of differentially expressed genes.

Cell-level analysis

Visualizing and clustering cells

The cataloguing and classification of cells is a long-standing biological challenge. Traditionally, cell types were determined morphologically or based on molecular cell surface markers. However, with the availability of genome-wide expression data, the possibility of transcriptome-based analysis of cell similarity provides an alternative indicator of cell type.

The first step in understanding the distribution of cells is often to apply dimensionality reduction techniques: this represents the thousands of dimensions (genes) found in scRNA-sequencing data with a much smaller number, attempting to maintain a representation of some variation in interest. Furthermore, by considering only a two or three dimensional space, visualization provides a mean to qualitatively explore the data. There are hundreds of dimensionality reduction methods available (Table 1), which the researcher can elect to apply either to all observed genes or a selected subset of genes of interest. The most widespread is Principal Component Analysis (PCA) [31], where weighted sums of dimensions represent the data. The dimensions for each sample are known as principal components. These dimensions explain decreasing amounts of variation in the original data, with the first

scRNA-seq

Cell-level analysis

Clustering

Clustering & Visualization

Cellular trajectory inference & branching

Gene-level analysis

Identifying marker genes of cell type

Gene expression dynamics through pseudotime analysis

Gene-gene correlations
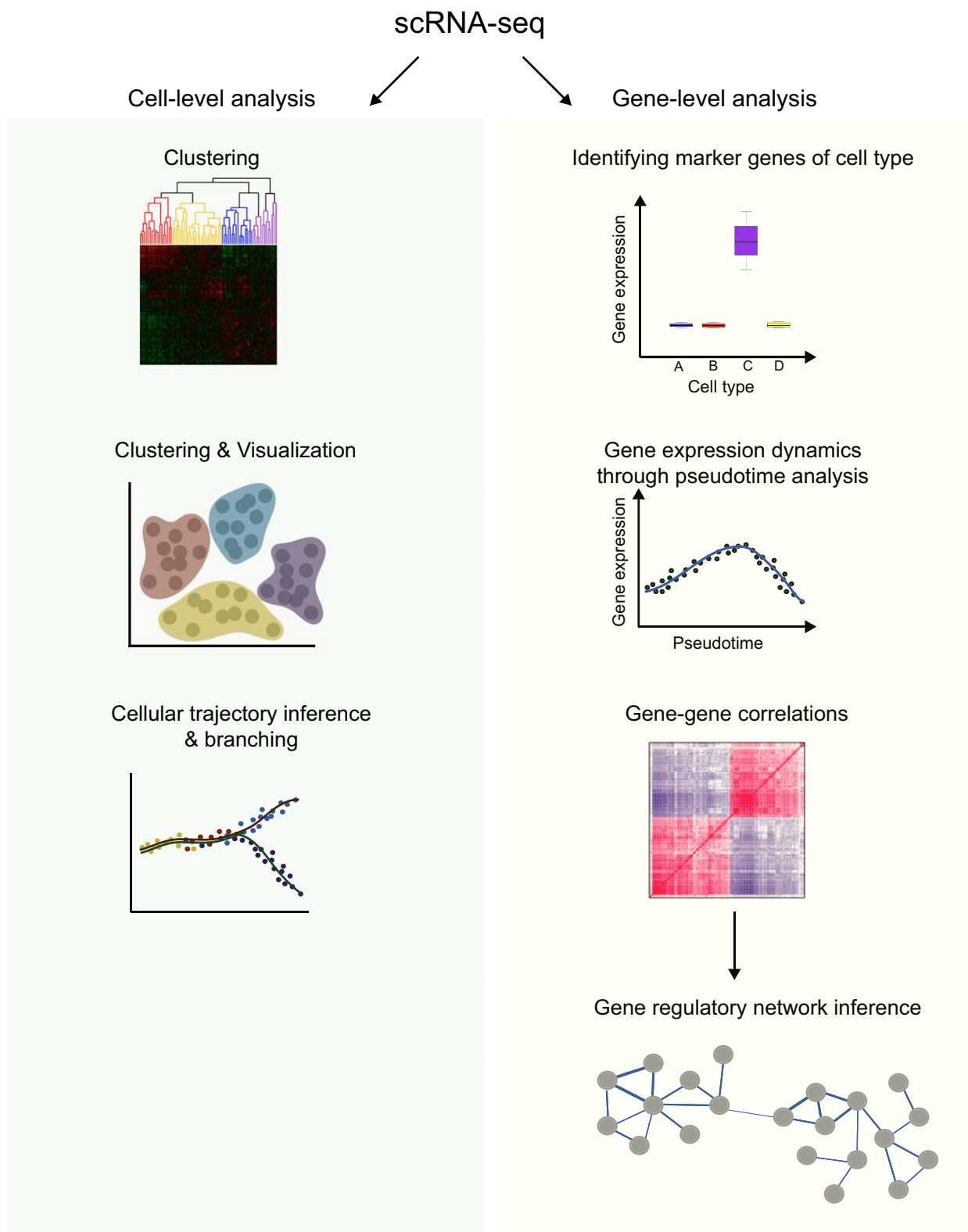

Gene regulatory network inference

Fig. 1. Overview of analysis methods for the interpretation of scRNA-seq data.

Table 1. Tools for the visualization and clustering of cells.

Dimensionality reduction and clustering of cells

Method	Description	Input	Availability
PCA	Linear dimensionality reduction, producing a set of uncorrelated components, explaining decreasing amounts of variation in the data.	Expression table	[31]
t-SNE	Nonlinear dimensionality reduction: t-distributed Stochastic Neighbour Embedding.	Expression table	[32]
ZIFA	A linear dimensionality reduction technique, using the factor analysis framework, that explicitly models dropout characteristics.	Log-transformed count values	https://github.com/epierson9/ZIFA [33]
Destiny	A fast implementation of diffusion maps for R.	Expression matrix (with a suggested variance stabilized transformation, for example, square root).	http://bioconductor.org/packages/release/bioc/html/destiny.html [34]
SNN-cliq	Graph-theory-based algorithm; uses shared nearest neighbour (SNN) graph based upon a subset of genes. The number of clusters is automatically chosen.	Log-transformation of normalized expression (e.g. RPKM)	http://bioinfo.uncc.edu/SNNCliq/ [35]
RaceID	Iterative K-means clustering of a Pearson correlation matrix, with number of clusters chosen using the gap statistic.	Raw gene expression matrix	https://github.com/dgrun/RaceID [14]
SC3	Distance is calculated first, followed by k-means clustering. Instead of optimizing parameters (e.g. distance metric, matrix transformation), SC3 combines several clustering outcomes and outputs an averaged result.	Normalized expression values	https://bioconductor.org/packages/release/bioc/html/SC3.html [36]
SIMLR	Learns a similarity measure from scRNA-seq data to perform dimensionality reduction, clustering and visualization.	Raw gene expression estimates and number of cell population.	https://bioconductor.org/packages/release/bioc/html/SIMLR.html [37]

principal component capturing as much of the variance as possible. PCA is a simple special case of linear factor analysis. Another commonly applied method is t-SNE (t-Distributed Stochastic Neighbour Embedding) [32], a nonlinear visualization technique which considers local distances between data points (cells) by combining dimensionality reduction with random walks on the nearest neighbour network with the goal of separating far-apart clusters, while also ensuring all data points can be seen by eye to allow for comparisons of cluster size. This is a variation of Multidimensional Scaling (MDS), where PCA is applied on pairwise Euclidean distances to preserve pairwise distances in a low-dimensional space.

While powerful, and popular, these techniques can be heavily affected by the problematic abundance of zeroes in single-cell data; an issue which several methods account for. ZIFA (zero-inflated factor analysis) [33] extends the linear factor analysis framework, (based on correlations in the data rather than covariances), accounting for dropout characteristics in the data. The R-package Destiny provides an alternative, nonlinear method using diffusion maps [38]: distance between cells reflects the transition probability based

on several paths of random walks between the cells. This assumes a smooth nature of the data, and also includes imputation of dropouts.

Unsupervised clustering techniques provide a mechanism to group cells by similarity. While this unbiased approach has benefits, the small number of samples and absence of a way to validate if groupings are 'real' poses a problem, along with prior information on the number or type of groups. The features of single-cell data discussed above, such as dropouts, biases and noise, also add to the difficulty of accurate clustering. Despite these problems, several tools have been developed for use with scRNA-seq, along with traditional methods such as hierarchical clustering [39]. SNN-Cliq [35] achieves clustering by considering similarity calculated using a graph-based approach in which a shared nearest neighbour (SNN) network is constructed using rankings of similarities based on expression levels; dense clusters of nodes (cells) are then found. RaceID [14], while also using similarity in expression between cells (based on Pearson correlation), utilizes a different approach: k-means clustering. In k-means clustering each sample is associated with one of k prototypes, so that the total squared distance (inverse of similarity)

from samples to prototypes is minimal. After the initial step, RaceID uses an outlier detection algorithm and identifies cells which do not fit the model accounting for technical and biological noise. This has been used in the detection of rare cell populations. Another k-means-based tool, Single Cell Consensus Clustering (SC3) [36], uses consensus clustering [40], an ensemble strategy, to average over parameter choices in an attempt to make cluster assignments more robust. A recent method, SIMLR [37], uses multiple-kernel learning to infer similarity in a gene expression matrix with a given number of cell populations. As multiple kernels are used, it is possible to learn a distance measurement between cells that is specific to the statistical properties of the scRNA-seq set under investigation.

Cellular trajectory inference and branching analysis

Trajectory analysis is a strictly simpler version of dimensionality reduction, where the assumption is that a 1-dimensional 'time' can describe the high-dimensional expression values. The theory is that during a biological process, changes will happen gradually, so biological observations can be ordered compared to each other in terms of pairwise similarity. While clustering techniques have been used to define discrete population and states for a long time, trajectory inference is younger in the field of scRNA-seq.

One of the initial methods for so called Pseudotime analysis of single cells was Monocle [41], which used a minimum spanning tree (MST) strategy to order cells by the distance to a start cell, based on a technique for putting microarray samples on a trajectory [42]. In the updated versions of Monocle, the MST strategy has been replaced by a more sophisticated tree-embedding strategy [43,44]. Monocle is a comprehensive R-package for single-cell analysis with functionality for normalization, clustering and differential expression analysis, but the main feature is the pseudotime inference.

Recently, diffusion pseudotime (dpt) has been developed [12]. In this technique, geodesic pairwise distances between samples on the data manifold are approximated using a diffusion map representation. Trajectory is then defined as the distance from a start cell along these distances. A different strategy for trajectory inference is to consider a generative model for the data, treating 'time-points' as hidden (or latent). This leads to the probabilistic interpretation of PCA, which in turn leads to factor analysis and ZIFA. Here, the expression of each gene can be described as a linear function of an unknown 'time'.

Nonlinearity in the data, as described in [41] precludes PCA from being an effective technique for this task. The Gaussian Process Latent Variable Model (GPLVM) allows gene expression to follow any smooth (nonlinear) function over time [45]. While more computationally demanding than linear versions, this allows cells to be put in the most likely ordering [45,46]. This means that the most number of genes exhibit smooth expression curves with as little noise as possible. Being a probabilistic model, the benefits are that uninteresting structure in the data can be accounted for directly, such as batch effects or technical factors. It is also possible to incorporate more information about your experimental design through priors [47]. There are many implementations of this method. For Python, there is GPy and GPFlow, and for R there are the DELOREAN (https://github.com/JohnReid/DeLorean) and PSEUDOGP (https://github.com/kieranrcampbell/pseudogp) packages.

The Ouija method [48] takes a different approach to pseudotime in a couple of ways. Firstly, it defines a generative model for gene expression in scRNA-seq data based on ZIFA, to deal with the most common types of measurement noise. Secondly, it is based on the assumption that a small number of switch-like markers for a biological process of interest are known. The cells are then ordered according to the most likely ordering to confer with the switching genes. Ouija is available as an R-package on BIOCONDUCTOR and is compatible with the popular SCATER package [27].

A unique problem in single-cell developmental data is that a set of progenitor cells can develop into multiple distinct cell types. This means the cells will not follow a single trajectory in the high-dimensional space. A couple of heuristics have been published: in Wishbone [49], cells are clustered by the pairwise detour distance relative to a reference cell, using geodesic distance. This method is reported to be correctly recovering the known stages and bifurcation point of T-cell development in mouse. Another method, that has been introduced by Haghverdi *et al.* [12], measures transition between cells using a random-walk-based distance.

More principled model based approaches have been presented with SCUBA, which considers transition of cells clusters over time [50]. As well as with GPfates/OMGP [47], where multiple smooth trajectories are explicitly modelled. After inference, each cell gets assigned a posterior probability of having been sampled from a particular trajectory. This method has been shown to be efficient in reconstructing the developmental trajectories of Th1 and Tfh cell populations during Plasmodium infection in mice (Table 2).

Gene-level analyses

Unwanted factor removal

Uninteresting, largely technical variation can be observed in both bulk RNA-seq and scRNA-seq experiments. This variation is usually correlated with some common experimental factor, such as room temperature or stock of reagents. This form of variation is known as batch effects. It is possible to handle batch effects by having a carefully balanced experimental design, such as uniformly distributing replicate conditions across batches. For statistical analysis and inference, if the samples are spread over multiple batches, this information can directly be accounted for [52,53]. Additionally, several statistical methods have been developed to adjust for batch effects [54,55]. One example is ComBat, which removes known batch effects using a linear model of expression from batches where variance is based on an empirical Bayesian framework [54].

Technical variation in scRNA-seq experiments could be mainly due to mRNA capture efficiency, cDNA amplification bias and the rate cDNAs in a library are sequenced. To estimate technical variation, several methods use spike-in molecules, which are added with each cell in the same quantity. Risso *et al.* have developed a sleuth of strategies called RUVSeq that either performs factor analysis on a set of control genes such as ERCC spike-ins or samples within replicate libraries to identify technical factors which can be adjusted for [56]. Similar strategies have also been made by others [57–59].

Grün *et al.* [60] have estimated technical noise in data by fitting a model that incorporates sampling noise and global sample-to-sample variability in sequencing efficiency. Subtracting technical noise from total noise has led to inferring the biological noise component, which has been shown to be consistent with single-molecule FISH, a highly sensitive imaging-based method for transcript counting [61]. An accurate noise model is needed in statistical analysis tasks to avoid overfitting.

A substantial amount of variation also results from differences in cell size or cell cycle stage of each cell. To adjust for cell cycle effects, Buettner *et al.* [62] have developed single-cell latent variable model (scLVM), which is a two-step approach that reconstructs cell cycle state before using this information to obtain adjusted gene expression levels by linear regression. They have also shown that removing cell cycle effects in T cells reveals subpopulations associated with T-cell differentiation [62]. This implies the importance of dissecting biological variation into interesting and uninteresting parts in correctly characterizing subpopulations.

Identification of highly variable genes

Several methods have been developed to identify genes that show high biological variability (Table 3). Brennecke *et al.* [22] have first estimated technical noise using spike-in molecules, and modelled the mean–variance relationship to identify highly variable genes. Kim *et al.* [7] have presented a statistical framework to decompose the total variance into the technical and biological variance based on a generative model, which would help in identifying variable genes. Another method, BASiCS, uses a Bayesian model which jointly models spike-ins and endogenous genes and provides posterior distributions for the extent of biological variability [63].

Identification of differentially expressed genes and marker genes

Identification of differentially expressed genes and marker genes of subpopulations is a simple yet important analysis in scRNA-seq studies. Although originally developed for bulk RNA-seq experiments, methods such as DESeq2 [64] and EdgeR [65] are also widely used in scRNA-seq experiments. DESeq2 identifies differentially expressed genes by fitting a GLM for each gene, uses shrinkage estimation to stabilize variance and fold changes, and applies a Wald or likelihood ratio (LR) test for significance testing [64]. EdgeR fits a GLM with negative binomial (NB) noise for each gene, estimates dispersions by conditional maximum likelihood, and identifies differential expression using an exact test adapted for overdispersed data [65]. Monocle also fits a GLM, but dispersion is estimated directly from the data for each gene, since most single-cell studies have enough samples to allow this [41]. For relative abundance data, dropouts are handled by using a tobit noise model, while using a NB noise model with imputed dropouts for count data.

One of the recent methods developed for scRNA-seq experiments, called MAST, uses a two-part generalized linear model that is adjusted for cellular detection rate (dropouts) [66]. Another method, M3Drop, applies Michaelis–Menten modelling of dropouts in scRNA-seq, that is used to identify genes differentially dropped out [67]. SCDE is a bayesian method to compare two groups of single cells, taking into account variability in scRNAseq data due to dropout and amplification biases and uses a two-component mixture for testing for differences in expression between

Table 2. Tools for the ordering of cells & bifurcation/branch identification.

Method	Description	Input	Availability
Pseudo-temporal ordering of cells			
PQ-trees	Samples are ordered by a minimum spanning tree of data, using a PQ-tree construction.	Expression table	[42]
Monocle2	A principal graph is embedded in the transcriptome space, distance along the graph from a start cell defines pseudotime.	Expression table, Batch effect formula, gene list (can be found through DE), dimensionality reduction options (method, number of dimensions)	Bioconductor package 'monocle' [3]
Wishbone	Diffusion maps on reduced k-NN graph (using waypoints).	Expression table, Start cell, number of waypoints, number of nearest neighbours k.	Python: https://github.com/ManuSetty/wishbone (MATLAB version only supports cytometry data) [49]
Wanderlust	Heuristic k-NN graph geodesic distance	Expression table	In CYT: https://www.c2b2.columbia.edu/danapeerlab/html/cyt-download.html [11]
DPT	Diffusion components are averaged for each sample based on spectral embedding, and used as a distance between samples.	Expression table, variance of Gaussian kernel, Start cell	For R and Matlab: http://www.helmholtz-muenchen.de/icb/research/groups/machine-learning/projects/dpt/index.html For Python: https://github.com/Teichlab/scrnatb [12]
GPLVM	Assume genes follow any smooth functions and infer time as latent parameter	Expression table or dimensionality reduction, covariance function, optional priors, Optional covariance function hyper parameters.	GPy GPFlow DeLorean pseudogp [45,46])
Ouija	Provided a small number of genes sigmoidal over trajectory, treat time as latent variable.	Expression table, list of assumed switch-like genes, optional priors of switching time and direction.	Bioconductor package 'ouija'. [48]
Branching analysis			
Wishbone	Two branches are detected by clustering detours between cells relative to a starting cells in terms of pseudotime.	Expression table	https://github.com/ManuSetty/wishbone[49]
Anticorrelation clustering	Branch points are identified when anticorrelated distances (relative to a start cell) become correlated. After this, cells can be segmented to belong to either of the two branches, or the trunk.	Expression table	[12]
OMGP/GPfates	Model data as a mixture of continuous processes. Each cell obtains a posterior probability of being generated by each of the branches.	Expression table	https://github.com/SheffieldML/GPclust[47]
Monocle	The principal graph fitted to the expression data explicitly has the concept of branches, which cells are assigned to.	Expression table, gene list	[5]
Mpath	Finding Minimum Spanning Tree in neighbourhood graph of landmarks.	Expression table	[51]

conditions [68]. Another method, SINCERA, identifies differentially expressed genes based on simple statistical tests such as Wilcoxon rank sum and *t*-tests [69]. In comparison to these methods, a more recent method, scDD, identifies genes where the overall distribution of values has changed between conditions. This answers a different question which might be of interest in scRNA-seq experiments [70]. Using a Bayesian

Table 3. Tools for gene-level analysis

Identification of differentially expressed genes

Method	Description	Input	Availability
Designed specifically for single cell RNA-seq data			
SCDE	Bayesian method to compare two groups of single cells, taking into account variability in scRNAseq data due to dropout and amplification biases.	Raw gene expression counts	http://hms-dbmi.github.io/scde/ [68]
MAST	Uses two-part generalized linear model that is adjusted for cellular detection rate.	Normalized gene expression values	https://github.com/RGLab/MAST [66]
M3Drop	Applies Michaelis-Menten modelling of dropouts to identify differential expression.	Raw gene expression counts	https://github.com/tallulandrews/ M3Drop [67]
scDD	A Bayesian modelling framework to identify genes that are differentially expressed and/or show a differential number of modes or differential proportion of cells within modes.	Normalized and log-scaled gene expression values	https://github.com/kdkorthauer/scDD [70]
SINCERA	Identifies DE genes based on simple statistical tests such as Wilcoxon rank sum and t-tests.	Raw gene expression values	https://research.cchmc.org/pbge/ sincera.html [69]
Designed originally for bulk RNA-seq data			
DESeq2	Fits a GLM for each gene, uses shrinkage estimation for dispersions and fold changes, applies a Wald or LR test for significance testing.	Raw gene expression counts	https://bioconductor.org/packages/ release/bioc/html/DESeq2.html [64]
EdgeR	Fits a negative binomial distribution for each gene, estimates dispersions by conditional maximum likelihood, identifies differential expression using an exact test adapted for overdispersed data. Supports arbitrary linear models.	Raw gene expression counts	http://bioconductor.org/packages/ release/bioc/html/edgeR.html [65]
Identification of highly variable genes			
Brennecke et al.	Biological variability of genes is inferred after quantifying the technical noise based on the square of coefficient of variation (CV^2) of the spike-in molecules.	Raw expression counts for both spike-ins and endogenous genes	[22]
Kim et al.	Presents a statistical framework to decompose the total variance into the technical and biological variance based on a generative model.	Raw expression counts for both spike-ins and endogenous genes	[7]
BASiCS	Uses a Bayesian approach that jointly models spike-ins and endogenous genes. Posterior probabilities associated to highly (or lowly) variable genes are provided.	Raw expression counts for both spike-ins and endogenous genes	https://github.com/catavallejos/ BASiCS [63]
Unwanted factor removal			
scLVM	Uses a Gaussian Process Latent variable model to dissect observed heterogeneity into different sources allowing removal of confounding factor of variation such as cell cycle-induced variations.	Raw gene expression counts and a set of genes associated with the latent factor	https://github.com/PMBio/scLVM [62]
Combat	Removes known batch effects based on an empirical Bayesian framework.	Normalized and log-scaled gene expression counts and batch information	https://github.com/brentp/combat. py/blob/master/R-combat.R [57]
OEFinder	Identifies potential artefacts (ordering effects) generated by the Fluidigm C1 platform using orthogonal polynomial regression.	A set of genes (and P-values) that are affected by the artefact	https://github.com/lengning/OEFinder [77]
RUVSeq	Adjusts for nuisance technical effects by performing factor analysis on a set of control genes such as spike-ins or samples such as replicate libraries.	Raw gene expression counts and a set of control genes, spike-ins or replicate libraries	https://github.com/drisso/RUVSeq [56]

Table 3. (Continued).

Method	Description	Input	Availability
Pseudotime Analysis			
Monocle	Spline regression using VGAM	Expression table, gene list	[5]
SwitchDE	Find genes which are explained as sigmoid curves over pseudotime.	Expression table	Bioconductor package 'switchde' [72]
ImpulseDE	Find genes which follow an impulse model.	Expression table	Bioconductor package 'impulsede' [73]
GP Regression	Find genes which follow any non-linear smooth function.	Expression table	GPy GPFlow Many others

modelling framework, scDD classifies each gene into one of the four types of changes across two biological conditions: shifts in unimodal distribution, differences in the number of modes, differences in the proportion of cells within modes, or both differences in the number of modes and shifts in unimodal distribution [70].

Gene-centric expression dynamics through pseudotime analysis

Using an inferred trajectory as described above, samples can be analysed using a continuous time covariate instead of a few discrete time-points. This enables the use of more sophisticated time series-based analysis techniques for modelling gene expression dynamics, and allows us to ask more complex questions from the data.

The popular scRNA-seq package Monocle provides a wrapper for the VGAM linear modelling package to investigate how expression changes over the trajectory. Splines are used to model expression dependence on pseudotime to allow nonlinear trends. The VGAM package allows for more than just expression levels to be modelled by the splines: with appropriate link functions, allelic expression balance or isoform usage can be modelled [3]. Splines require several parameters to be chosen however, and the choices greatly affect the results. A nonparametric nonlinear alternative to spline regression is Gaussian Process regression, which can be used in a likelihood ratio-based fashion to identify genes which are dependent on pseudotime [45,71].

Often, we want to ask particular questions from the data, in which case parametric models are useful. In the SwitchDE method, genes which sequentially switched on or off can be identified, along with a parameter letting you learn when the switch happens [72]. Similarly, an assumption can be that genes be described as a transient pulse over the pseudotime. The package ImpulseDE identifies such genes, while providing parameters for when in pseudotime the pulse occurs [73].

Correlation analysis and network inference

One important application of scRNA-seq studies is the identification of coregulated modules of genes and gene-regulatory networks constructed using gene-to-gene expression correlations. Here, genes with highly correlated expression levels across cells are assumed to be coregulated. Using single-cell transcriptomic data of Th2 cells, Mahata *et al.* [74] demonstrated how gene–gene correlations can be used to reveal novel mechanistic insights; they have applied correlation analysis between steroidogenic enzyme *Cyp11a1* and cell surface genes and identified *Ly6c1/2* as a marker of the steroid-producing cell population in mouse.

One method to elucidate regulatory interactions in bulk RNA-seq studies is called the weighted gene coexpression network analysis (WGCNA) [75]. In such a network, nodes represent genes and edges represent coexpression as defined by correlation and relative interconnectedness. The method has also been applied in a scRNA-seq study where the authors have identified a number of functional modules of coexpressed genes that can describe each embryonic developmental stage in mouse [76].

Although these methods are useful, the inferred networks are undirected; that is, they do not provide direct regulatory relationships among genes. To reveal which gene is upstream/downstream in the regulatory cascade, perturbation experiments (such as knockdown of a gene of interest) are typically required (Table 3).

Conclusions and perspectives

While many tools have been developed to take into account key features of single-cell RNA-sequencing data, there is still a way to go. The community will work towards refining existing methods to deal with the complexities of the data, such as the large amount of noise and high level of dropouts. In addition, we are facing issues of scalability due to the increase in

experimental throughput, which will also need to be addressed, along with adaptation for changes in experimental protocols. An example of this is the ability to measure gene expression in single cells spatially.

Another major advance will be the combination of other -omics techniques, such as the study of methylation [78] and chromatin accessibility [79] in single cells, leading to the same increase in resolution and potential to tackle novel questions as scRNA-seq. The ability to capture two levels of information within the same cell will hold great power in understanding regulation and functionality at the single-cell level. This has already been shown with the combination of whole-genome sequencing and transcriptomics [80,81] and bisulphite-sequencing and transcriptomics [82,83].

Although technical and experimental advances will continue to expand the horizon of research within the single-cell field, the application to an increasing range of biological areas holds exciting prospects. There has already been significant research into fields in which heterogeneity is well known, such as development [84,85], immunology [86,87] and cancer [88,89]. However, the increase in throughput will allow larger investigations. The ability to profile thousands of cells opens the scRNA-seq field to possibilities such as examining the role of human genetics: how do differences in single-cell heterogeneity depend on the genetic background of the individual? Furthermore, we are now on the verge of defining all cell types and subpopulations organism wide – creating a 'Human Cell Atlas' (www.humancellatlas.org). A thorough description of human cell populations has huge potential to help in understanding disease, and may in future play an important role in clinical diagnosis and treatment.

As the single-cell field, and the data generation that accompanies it, continues to expand at an incredible rate, it is imperative to develop tools and statistical methods to analyse the data in the best possible way, extracting significant and insightful biological meaning.

Acknowledgements

Author contributions

All authors read and approved the final manuscript.

References

1 Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, 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.

2 Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM and Wold BJ (2014) From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res* **24**, 496–510.

3 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.

4 Welch JD, Hu Y and Prins JF (2016) Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Res* **44**, e73.

5 Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL and Pachter L (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**, 46–53.

6 Deng Q, Ramsköld D, Reinius B and Sandberg R (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196.

7 Kim JK, Kolodziejczyk AA, Ilicic T, Illicic T, Teichmann SA and Marioni JC (2015) Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat Commun* **6**, 8687.

8 Reinius B, Mold JE, Ramsköld D, Deng Q, Johnsson P, Michaëlsson J, Frisén J and Sandberg R (2016) Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat Genet* **48**, 1430–1435.

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

10 Kar G, Kim JK, Kolodziejczyk AA and Natarajan KN (2017) Flipping between Polycomb repressed and active transcriptional states introduces noise in gene expression. *bioRxiv*, https://doi.org/10.1101/117267

11 Bendall SC, Davis KL, Amir E-AD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP and Pe'er D (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725.

12 Haghverdi L, Büttner M, Wolf FA, Buettner F and Theis FJ (2016) Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13**, 845–848.

13 Lönnberg T, Svensson V, James KR, Fernandez-Ruiz D, Sebina I, Montandon R, Soon MSF, Fogg LG, Stubbington MJT, Otzen Bagger F *et al.* (2017)

Temporal mixture modelling of single-cell RNA-seq data resolves a CD4+ T cell fate bifurcation. *Sci Immunol* **2**, eaal2192.

14 Grün 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.

15 Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C *et al.* (2015) Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142.

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

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

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

19 Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I and Enard W (2017) Comparative analysis of single-Cell RNA sequencing methods. *Mol Cell* **65** (631–643), e4.

20 Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, Macaulay IC, Cvejic A and Teichmann SA (2017) Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* **14**, 381–387.

21 Stegle O, Teichmann SA and Marioni JC (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133–145.

22 Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC *et al.* (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* **10**, 1093–1095.

23 Love MI, Hogenesch JB and Irizarry RA (2016) Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat Biotechnol* **34**, 1287–1291.

24 Jones DC, Kuppusamy KT, Palpant NJ, Peng X, Murry CE, Ruohola-Baker H and Ruzzo WL (2016) Isolator: accurate and stable analysis of isoform-level expression in RNA-Seq experiments. *bioRxiv*, https://doi.org/10.1101/088765

25 Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

26 Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N and Enright AJ (2013) Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods* **63**, 41–49.

27 McCarthy DJ, Campbell KR, Lun ATL and Wills QF (2016) scater: Pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R. *bioRxiv*, https://doi.org/10.1101/069633

28 Illicic T and McCarthy D (2016) cellity: Quality Control for Single-Cell RNA-seq Data. R package version 1.4.0.

29 Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC and Teichmann SA (2016) Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* **17**, 29.

30 Grün D and van Oudenaarden A (2015) Design and analysis of single-cell sequencing experiments. *Cell* **163**, 799–810.

31 Pearson K (1901) LIII. On lines and planes of closest fit to systems of points in space. *Philos Mag Series 6* **2**, 559–572.

32 Van der Maaten L and Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* **9**, 85.

33 Pierson E and Yau C (2015) ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol* **16**, 241.

34 Angerer P, Haghverdi L, Büttner M, Theis FJ, Marr C and Buettner F (2016) destiny: Diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241–1243.

35 Xu C and Su Z (2015) Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31**, 1974–1980.

36 Kiselev VY, Kirschner K, Schaub MT, Andrews T, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR and Hemberg M (2016) SC3 – consensus clustering of single-cell RNA-Seq data. *bioRxiv*, https://doi.org/10.1101/036558

37 Wang B, Zhu J, Pierson E, Ramazzotti D and Batzoglou S (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods* **14**, 414–416.

38 Lafon S and Lee AB (2006) Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Trans Pattern Anal Mach Intell* **28**, 1393–1403.

39 Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* **58**, 236–244.

40 Strehl A and Ghosh J (2002) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J Mach Learn Res* **3**, 583–617.

41 Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS and Rinn JL (2014) The dynamics and regulators of cell fate

decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386.

42 Magwene PM, Lizardi P and Kim J (2003) Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics* **19**, 842–850.

43 Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner H and Trapnell C (2017) Reversed graph embedding resolves complex single-cell developmental trajectories. *bioRxiv*, https://doi.org/10.1101/110668

44 Mao Q, Wang L, Goodison S and Sun Y (2015) Dimensionality Reduction Via Graph Structure Learning. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 765–774. ACM, New York, NY, USA.

45 Macaulay IC, Svensson V, Labalette C, Ferreira L, Hamey F, Voet T, Teichmann SA and Cvejic A (2016) Single-cell RNA-sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Rep* **14**, 966–977.

46 Campbell KR and Yau C (2016) Order under uncertainty: robust differential expression analysis using probabilistic models for pseudotime inference. *PLoS Comput Biol* **12**, e1005212.

47 Lönnberg T, Svensson V, James KR, Fernandez-Ruiz D, Sebina I, Montandon R, Soon MSF, Fogg LG, Nair AS, Liligeto UN *et al.* (2017) Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria. *Sci Immunol* **2**, eaal2192.

48 Campbell K and Yau C (2016) Ouija: Incorporating prior knowledge in single-cell trajectory learning using Bayesian nonlinear factor analysis. *bioRxiv*.

49 Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, 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.

50 Marco E, Karp RL, Guo G, Robson P, Hart AH, 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–E5650.

51 Chen J, Schlitzer A, Chakarov S, Ginhoux F and Poidinger M (2016) Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nat Commun* **7**, 11988.

52 Simpson EH (1951) The interpretation of interaction in contingency tables. *J R Stat Soc Series B Stat Methodol* **13**, 238–241.

53 Yule GU (1903) Notes on the theory of association of attributes in statistics. *Biometrika* **2**, 121–134.

54 Johnson WE, Li C and Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127.

55 Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM and Marron JS (2004) Adjustment of systematic microarray data biases. *Bioinformatics* **20**, 105–114.

56 Risso D, Ngai J, Speed TP and Dudoit S (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* **32**, 896–902.

57 Leek JT, Johnson WE, Parker HS, Jaffe AE and Storey JD (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883.

58 Leek JT (2014) svaseq: Removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res*, https://doi: 10.1093/nar/gku864

59 Stegle O, Parts L, Piipari M, Winn J and Durbin R (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500–507.

60 Grün D, Kester L and van Oudenaarden A (2014) Validation of noise models for single-cell transcriptomics. *Nat Methods* **11**, 637–640.

61 Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A and Tyagi S (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* **5**, 877–879.

62 Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC and Stegle O (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* **33**, 155–160.

63 Vallejos CA, Marioni JC and Richardson S (2015) BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol* **11**, e1004333.

64 Love M, Anders S and Huber W (2014) Differential analysis of count data–the DESeq2 package. *Genome Biol* **15**, 550.

65 Robinson MD, McCarthy DJ and Smyth GK (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140.

66 Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, 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.

67 Andrews TS and Hemberg M (2016) Modelling dropouts allows for unbiased identification of marker genes in scRNASeq experiments. *bioRxiv*.

68 Kharchenko PV, Silberstein L and Scadden DT (2014) Bayesian approach to single-cell differential expression analysis. *Nat Methods* **11**, 740–742.

69 Guo M, Wang H, Potter SS, Whitsett JA and Xu Y (2015) SINCERA: A pipeline for single-cell RNA-seq profiling analysis. *PLoS Comput Biol* **11**, e1004575.

70 Korthauer KD, Chu L-F, Newton MA, Li Y, Thomson J, Stewart R and Kendziorski C (2016) A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol* **17**, 222.

71 Kalaitzis AA and Lawrence ND (2011) A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformat* **12**, 180.

72 Campbell KR and Yau C (2017) switchde: Inference of switch-like differential expression along single-cell trajectories. *Bioinformatics* **33**, 1241–1242.

73 Sander J, Schultze JL and Yosef N (2017) ImpulseDE: Detection of differentially expressed genes in time series data using impulse models. *Bioinformatics* **33**, 757–759.

74 Mahata B, Zhang X, Kolodziejczyk AA, Proserpio V, Haim-Vilmovsky L, Taylor AE, Hebenstreit D, Dingler FA, Moignard V, Göttgens B *et al.* (2014) Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Rep* **7**, 1130–1142.

75 Langfelder P and Horvath S (2008) WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformat* **9**, 559.

76 Xue Z, Huang K, Cai C, Cai L, Jiang C-Y, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE *et al.* (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593–597.

77 Leng N, Choi J, Chu L-F, Thomson JA, Kendziorski C and Stewart R (2016) OEFinder: A user interface to identify and visualize ordering effects in single-cell RNA-seq data. *Bioinformatics* **32**, 1408–1410.

78 Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W and Kelsey G (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* **11**, 817–820.

79 Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY and Greenleaf WJ (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490.

80 Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM *et al.* (2015) G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* **12**, 519–522.

81 Dey SS, 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.

82 Moroz LL and Kohn AB (2013) Single-neuron transcriptome and methylome sequencing for epigenomic analysis of aging. *Methods Mol Biol* **1048**, 323–352.

83 Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood SA, Ponting CP, Voet T *et al.* (2016) Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* **13**, 229–232.

84 Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, Lao K and Surani MA (2010) Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* **6**, 468–478.

85 Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E *et al.* (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol* **33**, 269–276.

86 Proserpio V and Mahata B (2016) Single-cell technologies to study the immune system. *Immunology* **147**, 133–140.

87 Vieira Braga FA, Teichmann SA and Chen X (2016) Genetics and immunity in the era of single-cell genomics. *Hum Mol Genet* **25**, R141–R148.

88 Navin NE (2015) The first five years of single-cell cancer genomics and beyond. *Genome Res* **25**, 1499–1507.

89 Cloney R (2017) Cancer genomics: single-cell RNA-seq to decipher tumour architecture. *Nat Rev Genet* **18**, 2–3.