Molecular Aspects of Medicine 59 (2018) 114–122

Contents lists available at ScienceDirect

Molecular Aspects of Medicine

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

Identifying cell populations with scRNASeq

Tallulah S. Andrews, Martin Hemberg*

Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK

ARTICLE INFO

Article history:
Received 12 May 2017
Received in revised form
22 June 2017
Accepted 12 July 2017
Available online 25 July 2017

ABSTRACT

Single-cell RNASeq (scRNASeq) has emerged as a powerful method for quantifying the transcriptome of individual cells. However, the data from scRNASeq experiments is often both noisy and high dimensional, making the computational analysis non-trivial. Here we provide an overview of different experimental protocols and the most popular methods for facilitating the computational analysis. We focus on approaches for identifying biologically important genes, projecting data into lower dimensions and clustering data into putative cell-populations. Finally we discuss approaches to validation and biological interpretation of the identified cell-types or cell-states.

1. Introduction

There are an estimated 4×10^{13} cells in the human body (Bianconi et al., 2013), and they exhibit a stunning diversity in terms of both form and function. The traditional classification into ~200 cell-types is mainly based on morphology (Junqueria et al., 1992) rather than molecular features. Since the middle of the last century, immunofluorescence and flow cytometry have enabled more refined classification based on the presence or absence of various surface proteins (Coons et al., 1941; Fulwyler, 1965). However, these technique are limited to easily dissociable tissues, e.g. blood cell lineages (Roussel et al., 2010), and they only allow for a relatively small number of surface markers.

The development of single-cell RNA sequencing (scRNASeq) has enabled cell-type to be determined using the entire transcriptome of thousands of individual cells. scRNASeq has already been used to study several different tissues and organs, both during development and at a fixed point in time. These studies include various regions of the brain (Darmanis et al., 2015; Karlsson and Linnarsson, 2017; Liu et al., 2016; Tasic et al., 2016; Zeisel et al., 2015), retina ((Baron et al., 2016; Jaitin et al., 2014; Macosko et al., 2015; Zheng et al., 2017)), pancreas (Baron et al., 2016; Segerstolpe et al., 2016; Wang et al., 2016), immune cells (Jaitin et al., 2014; Villani et al., 2017), early embryonic development (Biase et al., 2014; Goolam et al., 2016; Xue et al., 2013) and in hematopoiesis (Velten et al., 2017; Wilson et al., 2015).

Here we will overview the main computational methods used by these and similar studies to process scRNASeq data for the identification and characterization of cell populations (Fig. 1). In addition, we discuss different protocols and experimental consideration that need to be taken into account when designing a scRNASeq experiment since they affect downstream analysis. We also discuss characteristics of scRNASeq which pose a challenge to the identification of biologically-relevant cell populations and statistical approaches to overcoming them. This is followed by an overview of methods available for performing unsupervised clustering on scRNASeq data that are used to group cells. Finally we discuss approaches to validating the identified groups of cells represent true distinct cell populations.

2. Experimental design considerations for scRNA-seq

Single-cell RNASeq (scRNASeq) is not a single method. It is a collection of protocols suitable for various applications that vary in terms of strengths and limitations. Consequently they are appropriate for different systems and different scientific questions. For example, one popular application is to identify rare cell populations (<1%) (Campbell et al. (2017); Grün et al., 2015; Jiang et al., 2016; Segerstolpe et al., 2016), which means that a large number of cells must be examined. For example, Campbell et al., 2017. sequenced 20,921 cells from mouse hypothalamus and they were able to identify neuronal subpopulations comprised of fewer than 50 cells (<0.2%). Another use of scRNA-seq is to characterize differences between similar cell-types, a task that requires methods with high detection rates for lowly expressed genes, and low levels of technical noise. For instance, dissecting differences among hematopoietic stem-cells requires detection of relatively lowly

* Corresponding author.
E-mail address: mh26@sanger.ac.uk (M. Hemberg).

http://dx.doi.org/10.1016/j.mam.2017.07.002

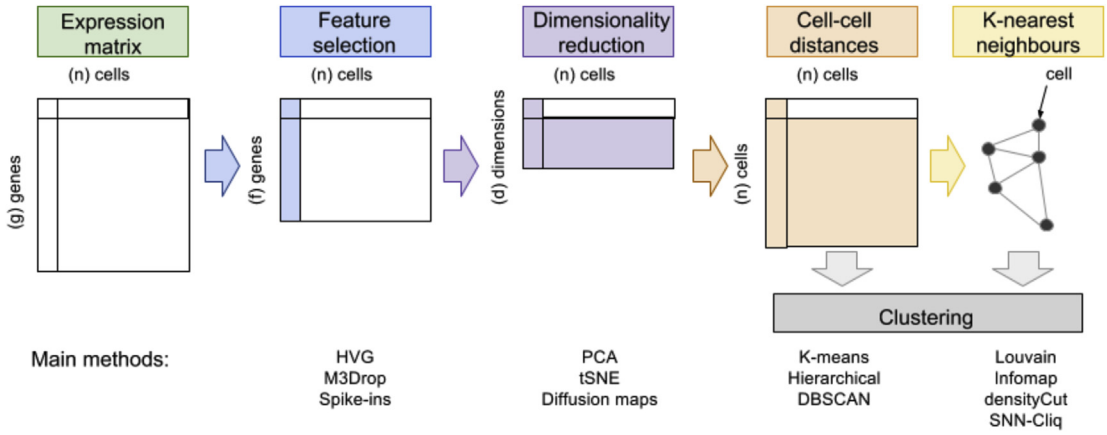

Fig. 1. Overview of methods covered in this review. Colour indicates which parts of the expression matrix are adjusted after each step, for instance feature selection only removes rows from the expression matrix, whereas dimensionality reduction calculates a new matrix composed of meta-features. Preprocessing steps not covered in detail in this review include quality control and normalization. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

expressed transcription factors, which in turn requires highly sensitive scRNASeq protocols (Tsang et al., 2015) or targeted approaches such as RT-qPCR (Wilson et al., 2015).

2.1. Experimental protocols

Briefly, each single-cell RNASeq protocol involves three main steps: i) isolation of single cells, ii) library preparation, and iii) sequencing. Isolation of cells requires dissociation of the sample followed by sorting into separate wells of a PCR plate, or capturing individual cells in separate droplets, microwells or microfluidic chambers. Library preparation involves reverse transcribing and amplifying either the full-length or only the 3'/5' "tagged" end of each mRNA. Sequencing is generally highly multiplexed and depth can vary from an average of 25,000 reads per cell (Macosko et al., 2015) to an average of 5 million reads per cell (Kolodziejczyk et al., 2015).

For studies requiring high throughput, droplet-based protocols such as InDrop (Klein et al., 2015), Drop-seq (Macosko et al., 2015) or 10X Chromium (Zheng et al., 2017), have gained wide popularity since they support cost-effective capture and library production for thousands to millions of cells. However, sequencing such a large number of cells can be prohibitively expensive. Fortunately, it has been shown that the minimum sequencing depth required to determine cell-type identity can be as few as 25,000–50,000 reads per cell (Jaitin et al., 2014; Pollen et al., 2014). Nevertheless, droplet-based methods tend to have lower detection rates and poorer mRNA capture efficiencies compared to other protocols (Svensson et al., 2017; Ziegenhain et al., 2017). Recent alternatives to droplets for high-throughput experiments include microwell-based approaches (Fan et al., 2015; Gierahn et al., 2017) and combinatorial indexing (Cao et al., 2017). However, these methods require cell-specific barcodes to be added prior to fragmentation thus only support 3'/5' sequencing.

For smaller-scale experiments, the two main classes are PCR plate-based methods, including Smartseq2 (Picelli et al., 2013), SCRB-seq (Soumillon et al., 2014), CEL-seq (Hashimshony et al., 2012) and MARS-seq (Jaitin et al., 2014), for which single cells are typically isolated with cell-sorters or microfluidic chips (e.g. Fluidigm C1) which combine cell-capture and library preparation. These methods tend to be less cost-effective for capturing cells, but have higher detection rates (Svensson et al., 2017; Ziegenhain et al., 2017). In addition, these methods can support both 3'/5' tag-based and full-transcript sequencing. It has been shown that sequencing such samples to a depth of 1 million reads per cell maximizes gene detection rates (Svensson et al., 2017; Ziegenhain et al., 2017), but additional sequencing is necessary for accurate quantification of isoforms or more lowly abundant ncRNAs (Huang and Sanguinetti, 2017; Sims et al., 2014).

A key feature of scRNASeq protocols to consider is the doublet rate. Doublets occur when two (or more) cells are captured in a single droplet or reaction chamber, and only through careful analysis, e.g. (Segerstolpe et al., 2016; Wang et al., 2016) can they avoid being mistaken for novel intermediate cell types. For high-throughput methods there is a trade-off between cell-capture efficiency and doublet rates, and common practice is to aim for 1–5% doublet rates (Ziegenhain et al., 2017). A similar tradeoff exists for microfluidic chips though higher capture rates are typically achievable with doublet rates 1–10% (Fluidigm Corporation, 2017) though older versions had doublet rates as high as 30% (Macosko et al., 2015). For plate-based methods there is no such explicit trade-off. In addition to contamination through doublets, mixed libraries could result from sequencing library "leakage", which has been reported at rates of 5–10% of reads in Illumina HiSeq 4 000 sequencing (Sinha et al., 2017), though another study has failed to detect this leakage in Illumina HiSeq X data (Owens et al., 2017).

Doublets are just one experimental challenge that can confound cell-population identification. Another major challenge is batch effects (Hicks et al., 2015; Tung et al., 2017). Batch effects result from minor differences in experimental efficiencies or cell state between experimental replicates prepared at different times or by different experimenters. If biological conditions of interest (e.g. mutant vs wild-type) are processed in different batches (e.g. on different days or on different plates), then it is impossible to statistically resolve biological vs technical effects (Hicks et al., 2015). Batch effects can be removed through careful experimental design which involves spreading each biological condition over all experimental batches. Consequently, cell populations identified in a single experimental batch may be a result of technical confounders and should thus be treated with skepticism.

2.2. Managing technical noise

Single-cell RNASeq protocols are typically used in conjunction with unique molecular identifiers (UMIs) and/or exogenous RNA spike-ins to address the high technical noise. UMIs eliminate amplification noise by enabling reads to be assigned to individual reverse-transcription events, thereby estimating original molecule

counts (Islam et al., 2014; Kivioja et al., 2011). Whereas spiking-in exogenous RNAs at known concentrations to each cell lysate can be used to model technical noise (Buettner et al., 2015; Vallejos et al., 2015), employed in normalization (Ding et al., 2015; Risso et al., 2014), and used to estimate absolute transcript counts from observed read counts (Owens et al., 2016). Plate-based methods support both the use of spike-ins and UMI tagging, whereas droplet-based and microwell-based methods exclusively employ UMI tagging (Gierahn et al., 2017; Macosko et al., 2015). Microfluidic devices may or may not be compatible with UMIs or spike-ins depending on their design.

The standard set of RNA spike-ins were chosen by the ERCC consortium from bacterial sequences (Baker et al., 2005; Jiang et al., 2011). Thus, they differ from mammalian transcriptomes in terms of transcript length, nucleotide content, poly-A tail length, and absence of introns. In addition, it has been shown that ERCC spike-ins have lower capture efficiencies than endogenous mRNA (Svensson et al., 2017). Moreover, ERCC spike-ins exhibit high technical variability which may exceed that observed for endogenous genes in some circumstances (Robinson and Oshlack, 2010; SEQC/MAQC-III Consortium, 2014), and spike-in counts can be influenced by biological effects, thereby invalidating them as a true control (Risso et al., 2014; Tung et al., 2017). New spike-in RNAs derived from human sequences are more representative of mammalian transcripts, and may alleviate some of these issues (Paul et al., 2016).

UMIs are 4–10bp barcodes added to the 5′ or 3′ end of each cDNA during reverse transcription (Islam et al., 2014); and hence are used in combination with transcript-end sequencing. As such isoform information is lost and fewer genetic variants will be captured hence it is much more difficult to assess allelic expression. The main advantage of 5'/3′ sequencing is improved power at low sequencing depth due to the elimination of amplification noise, by incorporating UMIs, and elimination of gene length biases (Phipson et al., 2017). While full-length protocols capture all parts of the transcript, they suffer from either a 3′ and/or a 5' bias (Archer et al., 2016; Ziegenhain et al., 2017). Non-uniform coverage is also evident in bulk RNASeq (Lahens et al., 2014). However, correcting for non-uniform coverage is more complex in scRNASeq due to the diversity of library preparation methods used in scRNASeq, e.g. Smartseq2, CEL-seq and MARS-seq, each with different coverage biases (Archer et al., 2016). The advantages of full-length protocols is greater sensitivity (Svensson et al., 2017), isoform usage in highly expressed genes (Shalek et al., 2013) and coverage of genetic variants to assess allelic expression (Deng et al., 2014; Jiang et al., 2017).

Multiplexed-sequencing of scRNASeq results in unequal numbers of reads across cells. Normalization methods are available for correcting for different sequencing depths across cells and removing batch effects (see: Vallejos et al. (2017) review). Sequencing depth can be corrected using counts/transcripts per million, or downsampling. Methods developed specifically for scRNASeq include scran (Lun et al., 2016), which has advantages for datasets with many differentially expressed genes, and SCnorm (Bacher et al., 2017), which accounts for different effects of sequencing depth of genes with different expression levels. As discussed above, if spike-ins were included in the dataset they may be used for normalization strategies which are robust to differentially expressed genes and preserves differences due to total RNA content (Buettner et al., 2015; Grün et al., 2014; Owens et al., 2016; Risso et al., 2014; Vallejos et al., 2015). Batch effects can be statistically removed in experiments where each batch contains cells from multiple biological conditions and each biological condition is spread across multiple batches, i.e. a "balanced" design (Hicks et al., 2015). Methods for batch correction include RUVs (Risso et al., 2014), ComBat (Stein et al., 2015) and linear mixed-modelling (Tung et al., 2017).

3. Strategies for dealing with high dimensionality

A scRNASeq experiment provides information about all genes, which is very useful for uncovering new biology, but simultaneous analysis of thousands of genes introduces statistical challenges. The total number of genes measured in a dataset is referred to as the dimensionality, and for mammalian samples there are often ~10^4 dimensions. The difficulties arising when comparing any high dimensional data are well known and they are often referred to as the "curse of dimensionality". When comparing cells in a high dimensional gene expression space, distances between cells become more homogenous, making it difficult to distinguish differences between populations from variability within a population.

There are two main approaches to dealing with the curse of dimensionality. Firstly, data can be projected into a lower dimensional space (generally referred to as "dimensionality reduction"). The lower dimensional space is generally defined by an algorithm to optimally preserve some characteristic(s) of the original data. Since information is always lost during projection, the choice of projection method involves a prioritization of a specific set of properties. Secondly, uninformative genes can be removed, referred to as "feature selection" in machine learning, to reduce the number of dimensions used in the analysis. Reducing the number of genes not only facilitates visualization, but it may also reduce noise and speed up calculations. Below we discuss some of the most popular methods for unsupervised dimensionality reduction and feature selection for scRNASeq data.

3.1. Dimensionality reduction

Principal component analysis (PCA) is a deterministic algorithm which projects data into a reduced number of independent dimensions. Dimensions are linear and they capture the highest variance possible. PCA is relatively fast, and when used with sparse-matrix representations it can scale to very large datasets. PCA generally preserves both long-range and short-range relationships amongst data points. A drawback is that PCA is restricted to linear dimensions and assumes approximately normally distributed data; two assumptions that may not be appropriate for scRNASeq datasets. A variation of PCA which explicitly deals with the large number of zero-values in scRNASeq data has been developed (Pierson and Yau, 2015) but the zero-inflation model employed may not fit all datasets (Andrews and Hemberg, 2016). Recently Risso et al. (2017) proposed a method similar to PCA based on a zero-inflated negative binomial model instead of a Gaussian model.

T-distributed stochastic neighbor embedding (tSNE) is a stochastic method designed for visualizing large high dimensional datasets (Maaten et al., 2008). A two or three dimensional embedding of the high dimensional data which preserves local structure amongst cells is calculated, but as a trade-off long-range information is lost (Fig. 2). Due to the probability distributions used to estimate the embedding, tSNE specifically projects data into isolated clusters, making it a popular choice for visualizing cell populations in scRNASeq data (Baron et al., 2016; Campbell et al., 2017; Macosko et al., 2015; Muraro et al., 2016; Segerstolpe et al., 2016).

A drawback of tSNE is the stochastic nature of the algorithm, and applying tSNE to the same dataset multiple times will produce different embeddings. Although the differences are often small and insignificant, best practice is to run the algorithm multiple times to ensure the stability of results. In addition, tSNE embeddings are sensitive to the choice of the "perplexity" parameter. Thus, it is necessary to run the algorithm multiple times to determine the

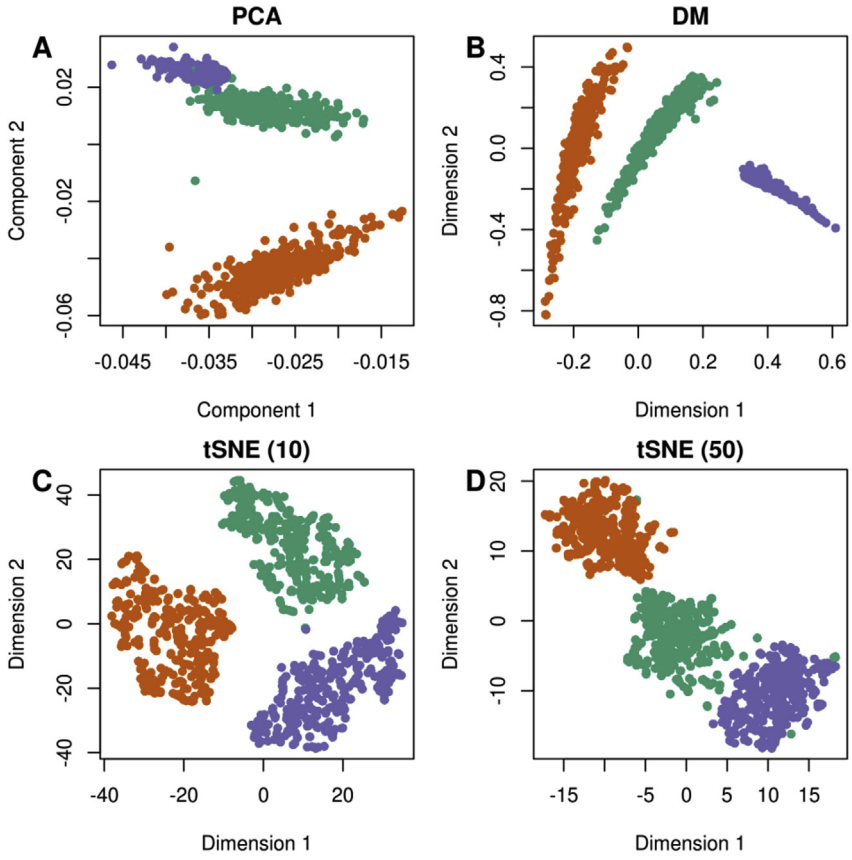

Fig. 2. A synthetic RNASeq dataset projected into two dimension using different methods. The data contains three groups, two of which are more similar to each other (green & purple) than to the third (orange), in addition the purple group follows a different trajectory than the other two groups. (A) Principal component analysis (PCA) preserves variance within the data. (B) Diffusion maps (DM) finds non-linear trajectories within the data. (C & D) t-distributed stochastic neighbor embedding (tSNE) highlights clustering structure within the data at the expense of long-range information. tSNE is also sensitive to the choice of perplexity, here we show perplexity = 10 (C) and perplexity = 50 (D) for the same data.

appropriate perplexity for a particular dataset (Fig. 2C,D). The authors of the method recommend only using tSNE for visualization purposes and not as a dimensionality reduction method (Maaten et al., 2008).

Diffusion maps (DM) is a nonlinear projection method which has predominantly been used for analyzing continuous progressions of cells (Moon et al., 2017; Angerer et al., 2016; Haghverdi et al., 2016). DM is based on models of a diffusion process to embed high dimensional data in low dimensional space. It is assumed that the low dimensional space is smooth and that it can be inferred from the distances between the cells. Unlike tSNE, DMs preserve both local and distant relationships between points. Since DM assumes a relatively smooth continuum of cells it performs well on large RT-qPCR and large scRNASeq experiments (i.e. > 1000 cells assayed) but performance drops for datasets with few cells or the presence of very distinct cell populations (Qiu et al., 2017).

3.2. Feature selection

Michaelis-Menten modelling of dropouts (M3Drop) uses the relatively tight relationship between dropout rate (i.e. the frequency of zeros) and mean expression to perform feature selection. Genes with high dropout rate relative to their expression are likely to be differentially expressed across subpopulations of cells within the dataset. Thus, identifying outliers from the fitted relationship is an effective method of feature selection for scRNASeq, and it can be shown that the method improves clustering and allows for batch effect corrections (Andrews and Hemberg, 2016).

Highly variable genes (HVG) is based on the assumption that genes with high variance relative to their mean expression are due to biological effects rather than just technical noise. The method seeks to identify genes that have a higher variability than expected by considering the relationship between variance and mean expression. This relationship is difficult to fit, and in practice genes are ranked by their distance from a moving median (Kolodziejczyk et al., 2015) or another statistic derived from variance is used, e.g. the squared coefficient of variation (Brennecke et al. (2013)).

Spike-in based methods use a similar idea to HVG and M3Drop to identify features of interest. Here, technical noise is explicitly modelled using data from spike-in RNAs to identify genes exhibiting dropout rates or variance significantly higher than those of spike-ins with similar expression levels. Examples of spike-in based methods include those by Brennecke et al., 2013, BASiCS (Vallejos et al., 2015), and scLVM (Buettner et al., 2015).

Correlated expression is a different approach to identifying biologically relevant genes specifically for identifying cell populations (Andrews and Hemberg, 2016). Genes differentially expressed between a pair of cell-types will be correlated with each other. Correlations will be positive if they are co-expressed in the same cell-type and negative if they are expressed in different cell-types. Feature selection proceeds using either the magnitude and/or significance of the correlations. An alternative which combines high variability and correlation information is to use gene-loading from PCA e.g. (Macosko et al., 2015; Pollen et al., 2014; Usoskin

et al., 2015). PAGODA (Fan et al., 2016) performs a variant on this method which combines HVG and PCA loadings to identify important sets of genes which either were highly correlated in the dataset or which share functional annotations.

The methods for dealing with high dimensional data presented here are not mutually exclusive, and it is common practice to apply multiple approaches. Dimensionality reduction methods, i.e. PCA, tSNE, and DM, are susceptible to batch effects and technical noise which may obscure structure within the data (Finak et al., 2015; Hicks et al., 2015; Tung et al., 2017). Performing feature selection to remove genes with little biological signal prior to dimensionality reduction can greatly reduce these effects (Andrews and Hemberg, 2016). Examples of such approaches include iteratively performing spike-in based feature selection followed by PCA (Liu et al., 2016; Tasic et al., 2016), HVG feature selection followed by tSNE (Segerstolpe et al., 2016), and HVG feature selection followed by dimensionality reduction with both PCA and tSNE (Campbell et al., 2017).

4. Unsupervised clustering methods for identification of cell populations

One of the most popular uses of scRNASeq is to identify and characterize cell-populations. From a biological point of view, cell-populations are often different cell-types, e.g. neurons and glia in a brain sample, but they can also correspond to different states of identical cell-types, e.g. stimulated and unstimulated T-cells. From a mathematical point of view, *de novo* identification of cell-populations in scRNASeq data is an unsupervised clustering problem. As such, the problem has been widely studied in the machine learning literature, and there are several well-established strategies that have been adapted for scRNASeq data. We will discuss some of the central issues as well as the main classes of clustering algorithms which have been applied to single-cell RNASeq data below.

The number of different possibilities for grouping a large number of cells into k clusters is typically astronomical, making it infeasible to consider all possible partitionings. Instead an optimal solution is found using various heuristic methods which balance partitioning quality and scalability. The quality of a clustering is based on a metric comparing intra- and inter-cluster similarity employ, different metrics make different assumptions about the underlying distribution of the data, e.g. “modularity” assumes a sparse graph structure of data, whereas distance to cluster centroid, used by k-means, assumes roughly equally sized round clusters in the data. Applying a method to data which violates its assumption will result in incorrect cluster identification, and consequently no clustering algorithm works well in all situations (Wiwie et al., 2015).

K-means is a commonly used clustering algorithm for single-cell analysis (Burns et al., 2015; Grün et al., 2015; Kiselev et al., 2017; Muraro et al., 2016; Tsang et al., 2015), and it is generally used after feature selection and dimensionality reduction, e.g. (Burns et al., 2015; Grün et al., 2015). K-means is a very fast method which iteratively assigns cells to the nearest cluster centre (or “centroid”), and then recomputes the cluster centroids. However, k-means requires the number of clusters to be predetermined and uses stochastic starting locations for each cluster, thus requiring it to be run multiple times to check robustness to these parameters. These many results can subsequently be combined by calculating a consensus, e.g. as done by SC3 (Kiselev et al., 2017).

A shortcoming of k-means is that the method assumes a predetermined number of round equally-sized clusters. If these assumptions are violated, k-means may identify many adjacent clusters along a differentiation trajectory and merge rare cells with a more prevalent cell-type (Fig. 3A). Rare cell-populations can be identified by combining k-means with outlier detection methods, e.g. RaceID (Grün et al., 2015). However, RaceID performs poorly when the data does not contain rare cell-populations (Li et al., 2017; Lin et al., 2017).

Hierarchical clustering is another popular general-purpose clustering method commonly used to identify cell-populations (Baron et al., 2016; Guo et al., 2015; Patel et al., 2014; Wilson et al., 2015). Different variants of hierarchical clustering make different assumption, but the most common ones, i.e. Ward's (Ward, 1963) and “complete”, assume round equally-sized clusters like k-means (Fig. 3C,D). However, hierarchical clustering is slower than k-means, but has the advantage of being able to determine relationships between clusters of different granularities since the result can be visualized as a dendrogram. This dendrogram is then “cut” at different heights to generate different numbers of clusters. Methods which tailor hierarchical clustering for single-cell RNASeq data include pcaReduce (Žurauskienė and Yau, 2016), SINCERA (Guo et al., 2015), and CIDR (Lin et al., 2017). Extensions to hierarchical clustering which perform feature selection after each split have been used in the analysis of the neuronal cell-types in brain (Zeisel et al., 2015) and islet cell-types in pancreas (Baron et al., 2016). However these methods tend to identify many clusters which likely represent the same cell type (Baron et al., 2016; Li et al., 2017).

Density-based clustering identifies clusters as contiguous regions with a high density of cells. Unlike hierarchical clustering or k-means, density-based clustering does not assume clusters of a particular shape or size (Fig. 3 B). However, density-based methods often assume that all clusters are equally dense, i.e. cell populations are equally homogenous. In addition, this density must be provided to the algorithm through one or more parameters. Setting the density parameters is analogous to choosing the number of clusters for k-means or selecting where to cut the tree for hierarchical clustering. Since density-based clustering requires a large number of samples to accurately estimate densities, it works well on droplet-based datasets, large RT-qPCR experiments and cytometry experiments containing data for thousands to millions of cells (Campbell et al., 2017; Jiang et al., 2016; Macosko et al., 2015). The dominant method is DBSCAN (Ester et al., 1996), which has been combined with dimensionality reduction in the original Seurat (Macosko et al., 2015) or a rare cell-type sensitive feature selection method in GiniClust (Jiang et al., 2016).

Graph clustering, also known as “community detection”, is an extension of density-based clustering specifically for data represented as a graph, i.e. a set of cells connected to each other by “edges”. Since graphs can easily represent complex nonlinear structure with minimal assumptions, cell populations of different sizes, densities, and shapes can be identified (Lancichinetti and Fortunato, 2009). An additional advantage of graph-based clustering methods is that they can scale to millions of cells (Blondel et al., 2008; Danon et al., 2005; Rosvall and Bergstrom, 2008; Schaeffer, 2007.

Density in a graph can be measured as the number of edges connecting a set of cells and can be readily compared to a null hypothesis, e.g. a fully random graph or a degree-controlled random graph, using a metric called modularity. The most popular methods using modularity are based on the Louvain algorithm (Blondel et al., 2008; Lancichinetti and Fortunato, 2009), which is used in PhenoGraph (Levine et al., 2015) and version 1.4 of Seurat. Alternatively, density can be estimated by modelling random walks on the graph and using the proportion of time spent at each cell, a strategy used by densityCut (Ding et al., 2016). Another approach to estimating density, that is employed by SNN-Cliq (Xu and Su, 2015), is to use the overlaps between the k-nearest neighbours of each cell.

The main drawback of graph-based methods is that scRNASeq

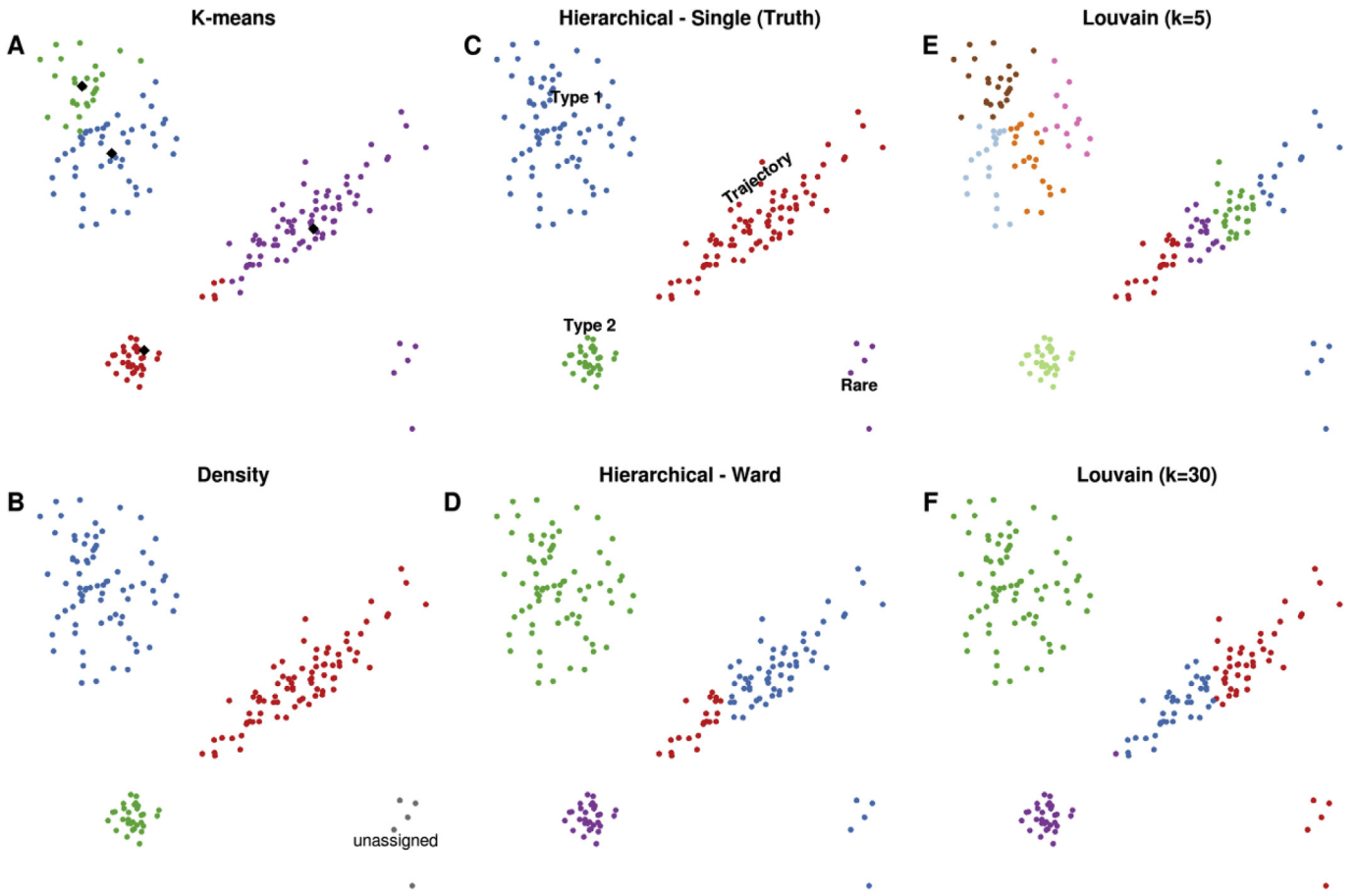

Fig. 3. Synthetic data to represent different structures within single-cell RNASeq data is clustered using different methods. Two dimensional data was simulated to represent a noisy population (Type 1), a tight population (Type 2), a developmental trajectory (Trajectory), and a rare cell population (Rare). Colours indicate clusters identified by different methods. These data were clustered with (A) k-means, $k = 4$, black diamonds are cluster centroids; (B) DBSCAN with density parameter: $eps = 0.5$; (C) single-linkage hierarchical, cut at $k = 4$; (D) Ward's hierarchical, cut at $k = 4$; (E) Louvain clustering after converting to a 5-NN graph; (F) Louvain clustering after converting to a 30-NN graph.

data does not have an inherent graph structure. Thus, the performance of these methods is reliant on how effectively the scRNASeq data is converted into graph-representation. scRNASeq data is typically converted to a graph-structure by representing cells as nodes connected by edges to their k nearest neighbours (kNN), e.g. (Ding et al., 2016; Satija et al., 2015). This representation assumes equally sized cell populations (Fig. 3 E,F). However, due to the curse of dimensionality, the identification of k-nearest-neighbours may not be a robust strategy (Beyer et al., 1999). Thus, it is often necessary to perform some form of dimensionality reduction and/or feature selection prior to defining kNN graphs to prevent biasing clustering algorithms (Radovanović et al., 2010).

A key decision for clustering methods is how many groups to identify. Coarse clusterings identify a small number of very distinct clusters which are more likely to correspond to cell-types; whereas fine clustering identifies a large number of less distinct clusters which may correspond to different cell-states. Most clustering algorithms require either the number of clusters (k) or parameters relating to the coarseness of clustering (e.g. density parameters) to be defined *a priori* by the user. Choosing an appropriate k is difficult since there is no generally accepted method to do so. In fact, for many samples there exists a hierarchy of cell-types and cell-states, and they may all be of interest. For example, Zeisel et al. (2015) clustered a brain sample and found a coarse level of 9 main cell-types which separated neurons from various non-neuronal cell-types such as glia, and a second finer level which subdivided neurons into seven layer-specific groups.

Although there are several software packages available for unsupervised clustering (Ding et al., 2016; Grün et al., 2015; Guo et al., 2015; Kiselev et al., 2017; Levine et al., 2015; Lin et al., 2017; Macosko et al., 2015; Xu and Su, 2015; Žurauskienė and Yau, 2016), the different methods have not yet been thoroughly and independently benchmarked. Benchmarking scRNASeq clustering is a challenging problem due to the difficulty of obtaining a ground truth through orthogonal methods. In general, we do not know the identity of any of the cells assayed *a priori* and reliable marker genes are only available for a few well-characterized cell-types. A possible solution is to use samples where the cells are derived from different cell-lines or early embryonic states (Kiselev et al., 2017), but such benchmarks may fail to recapitulate the complexity of a tissue sample.

Many of the computational tools discussed in the preceding sections are available in ASAP, a recently released user-friendly interactive web-tool (Gardeux et al., 2016). Other easy to use software packages which implement different combinations of feature selection, dimensionality reduction and clustering algorithm include Seurat (Satija et al., 2015), PAGODA (Fan et al., 2016), and SC3 (Kiselev et al., 2017).

5. Biological characterization of clusters

Interpreting groups identified by a clustering algorithm is not trivial. Firstly, due to the heuristic nature of clustering algorithms, they will always find some partitioning, even if presented with data generated from a uniform distribution. In addition, even when clusters are a result of biological effects, rather than noise, those effects may not represent differences in cell-type. Today, there are no accepted standards for the criteria required to label a cell population as a novel cell type. Defining cell-types based on transcriptional differences is difficult since transient differences in cell-state, e.g. cell-cycle stage, can have a larger effect on the global transcriptome than cell-type (Buettner et al., 2015). Furthermore, studies of differentiation processes in the intestine (Barker, 2014) and bone marrow (Velten et al., 2017) have revealed greater functional plasticity than previously thought, challenging the notion of rigid cell-type definitions in these systems. However, single cell studies of adult differentiated tissue are generally consistent with each other and recapitulate morphology-based cell-type classifications (Crow et al., 2017; Kiselev and Hemberg, 2017).

5.1. Computational approaches

Clustering algorithms always identify groups, even if these groups represent only slight differences in density due to noise in homogeneous datasets. Assessing the statistical significance of clusters, however, is very difficult due to the heuristic nature of most clustering algorithms. In general, to assess significance the algorithm must be re-run of multiple instances of a null-model and the results compared to those of the observed data. The null-model datasets may be drawn from probability distributions fit to the observed data, e.g. (Severson et al., 2017), or generated by randomly reordering the observed expression values for each gene independently.

A useful strategy for ensuring that a good clustering has been obtained is to apply several algorithms to the same data and make sure that the results are consistent. Since clustering methods make different assumptions, using more than one method for the same data can ensure the result is not dependent on those assumptions. In addition, for stochastic clustering methods, such as k-means (Hartigan and Wong, 1979) or Louvain maximum modularity (Blondel et al., 2008), running the algorithm multiple times and taking the consensus will give a more robust solution than the result from any single run (Goder and Filkov, 2008; Kiselev et al., 2017). Based on our experience, distinct clusters are consistent across clustering algorithms, whereas if there is little separation between clusters the results will vary between methods.

Another way to assess the reliability of cell clusters is to subsample cells and/or genes and re-analyze the result (Joost et al., 2016; Tasic et al., 2016; Zeisel et al., 2015). Clusters resulting from outliers or low frequency doublets will be absent in any subsampling which don't contain them. However, a rare cell population may comprise only a handful of cells in the dataset, and thus will generally fail this approach, even if they represent true cell-types. Identifying large dominant clusters as well as rare outliers is a very challenging problem, and to the best of our knowledge there is no method available that performs well for both tasks.

The approaches discussed above improve the confidence that the obtained clusters are robust and reliable, but they do not demonstrate that the cell populations are biologically relevant. Cell-types and cell-states are generally associated with specific functional characteristics, for instance neurons transmit electrical signals whereas T-cells recognize foreign antigens and initiate the immune response (Trapnell, 2015). Linking cell-clusters from a scRNASeq experiment with biological function(s) is usually a challenging problem and there are no automated procedures or softwares to carry out this task in full.

5.2. Experimental approaches

Often the first step is to identify differentially expressed genes, i.e. genes that can reliably distinguish between two or more clusters, or marker genes, i.e. genes that are high in only one cluster. However, groups derived from unsupervised clustering methods will always have differentially expressed genes between them due to the nature of the algorithms. Instead, functional enrichment analysis using external annotations, e.g. Gene Ontology, needs to be applied to marker and differentially expressed genes to identify patterns that are biologically meaningful.

Crucially, marker genes can be used for experimental validation. For example, co-expression of marker genes could be replicated using RT-qPCR, high-throughput sequencing or cytometry (Burns et al., 2015; Jaitin et al., 2014; Muraro et al., 2016; Tasic et al., 2016). Marker genes may be used to isolate cell populations for culturing and functional assays. Examples of this are found in hematopoiesis; where Velten et al. (2017) demonstrated that different hematopoietic progenitor subgroups found in scRNASeq data were biased towards either myeloid or erythropoietic cell-fates while retaining the potential to differentiate into both classes, and immunology; where cell populations identified by scRNASeq were isolated and shown to release novel suites of cytokines (Jaitin et al., 2014; Wilson et al., 2015). In some cases functional assays can be directly combined with scRNASeq, such as measuring the excitability of neurons with patch-clamping experiments which was used to identify subsets of interneurons with quicker or slower excitation profiles (Fuzik et al., 2016).

Marker genes may also be used for *in situ* imaging of the putative cell population. Burns et al. (2015) used immunofluorescence to demonstrate spatial localization of different cell-types in the inner ear. Immunofluorescence may also be used to confirm co-expression or mutually exclusive expression of cell-type markers (Tirosh et al., 2016). Alternatively, cell-type specific markers can be targeted using FISH to validate their co-expression in addition to determining spatial distribution of the cell-type in the tissue. Both immunohistochemistry and single-molecule RNA-FISH were used to identify spatial location of different putative cell-types within hair follicles and dissect spatial vs differentiation related expression patterns (Joost et al., 2016).

A different approach to validating clusters is comparing across across different species, e.g. human and mouse, to determine whether a cluster is broadly conserved and thus likely to be a true cell-type. A good example of this approach is the comparison of radial glial progenitor populations in human, mouse and ferret by Johnson et al. (2015). They found two novel subpopulations which were present in human and ferret but absent in mouse which through comparative genomics of the respective marker genes were linked to gyrencephaly in mammals. Another example is the validation of neuron populations with high Dlx1 and Dlx2 in both human and mouse brain (Aibar et al., 2017). A conceptually similar approach to the cross-species comparison is to compare clusters from different states, e.g. healthy and diseased. Such studies have been carried out for pancreas cell-populations in both healthy individuals and those with type-II diabetes (Baron et al., 2016; Segerstolpe et al., 2016; Wang et al., 2016). Cell-types were broadly consistent in donors with type-II diabetes, but significant differences in expression of specific genes and in relative frequencies of cell-types have been noted. Instead of a disease state, gene knockouts can be used to interrogate identified cell populations. Manipulating the levels of key transcription factors associated with particular a cell population can validate the population

by showing it increases or decreases respectively. Olsson et al. (2016) created knockouts of Gfi1 and Irf8, which are associated with different hematopoietic progenitors. These resulted the respective cell-types, i.e. granulocytic progenitors for Gfi1 and monocytic progenitors for Irf8, being absent in the knockout populations.

Biological validation of cell populations is necessary as clustering methods can over-partition cells into multiple groups which represent the same single functional cell-type (Fuzik et al., 2016; Jiang et al., 2016; Severson et al., 2017). Validation experiments can also provide useful information on the specific function(s) of novel cell population or relevance to disease states.

6. Closing remarks

Identifying novel or known cell populations is likely to remain a key goal of scRNASeq experiments in the future. These endeavors will continue to drive the development of experimental protocols and analysis methods. However, due to the trade-offs between cell number and sensitivity, it is likely there will never be a single optimal platform for scRNASeq experiments. Likewise, no computational methods for dimensionality reduction, feature selection and unsupervised clustering will be optimal in all situations. In addition, many challenges remain for the analysis of cell populations identified by scRNASeq. Although novel cell populations can be readily identified using existing methods, these findings must be validated using external data or experiments to ensure they are not technical artefacts.

References

Aibar, S., Bravo González-Blas, C., Moerman, T., Wouters, J., Huynh-Thu, V.A., Imrichová, H., Kalender Atak, Z., Hulselmans, G., Dewaele, M., Rambow, F., Geurts, P., Aerts, J., Marine, J.-C., van den Oord, J., Aerts, S., 2017. BioRxiv. http://dx.doi.org/10.1101/144501, 144501.

Andrews, T.S., Hemberg, M., 2016. BioRxiv. http://dx.doi.org/10.1101/065094, 065094.

Angerer, P., Haghverdi, L., Büttner, M., Theis, F.J., Marr, C., Buettner, F., 2016. Bioinformatics 32, 1241–1243.

Archer, N., Walsh, M.D., Shahrezaei, V., Hebenstreit, D., 2016. Cell Syst. 3, 467–479 e12.

Bacher, R., Chu, L.-F., Leng, N., Gasch, A.P., Thomson, J.A., Stewart, R.M., Newton, M., Kendziorski, C., 2017. Nat. Methods 14, 584–586.

Baker, S.C., Bauer, S.R., Beyer, R.P., Brenton, J.D., Bromley, B., Burrill, J., Causton, H., Conley, M.P., Elespuru, R., Fero, M., Foy, C., Fuscoe, J., Gao, X., Gerhold, D.L., Gilles, P., Goodsaid, F., Guo, X., Hackett, J., Hockett, R.D., Ikonomi, P., External RNA Controls Consortium, 2005. Nat. Methods 2, 731–734.

Barker, N., 2014. Nat. Rev. Mol. Cell Biol. 15, 19–33.

Baron, M., Veres, A., Wolock, S.L., Faust, A.L., Gaujoux, R., Vetere, A., Ryu, J.H., Wagner, B.K., Shen-Orr, S.S., Klein, A.M., Melton, D.A., Yanai, I., 2016. Cell Syst. 3, 346–360 e4.

Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U., 1999. When is “nearest neighbor” meaningful? In: Beeri, C., Buneman, P. (Eds.), Database Theory — ICDT'99, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 217–235.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M.C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., Canaider, S., 2013. Ann. Hum. Biol. 40, 463–471.

Biase, F.H., Cao, X., Zhong, S., 2014. Genome Res. 24, 1787–1796.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. J. Stat. Mech. 2008, P10008.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., Heisler, M.G., 2013. Nat. Methods 10, 1093–1095.

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., Stegle, O., 2015. Nat. Biotechnol. 33, 155–160.

Burns, J.C., Kelly, M.C., Hoa, M., Morell, R.J., Kelley, M.W., 2015. Nat. Commun. 6, 8557.

Campbell, J.N., Macosko, E.Z., Fenselau, H., Pers, T.H., Lyubetskaya, A., Tenen, D., Goldman, M., Verstegen, A.M.J., Resch, J.M., McCarroll, S.A., Rosen, E.D., Lowell, B.B., Tsai, L.T., 2017. Nat. Neurosci. 20, 484–496.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., Adey, A., Waterston, R.H., Trapnell, C., Shendure, J., 2017. BioRxiv. http://dx.doi.org/10.1101/104844, 104844.

Coons, A.H., Creech, H.J., Jones, R.N., 1941. Exp. Biol. Med. 47, 200–202.

Fluidigm Corporation, 2017. Redesign of C1 medium-cell 96 and HT IFCs improves single-cell capture efficiency [WWW document]. http://info.fluidigm.com/FY16Q2-C1WhitePaperUpdate_LP.html (accessed 4.24.17).

Crow, M., Paul, A., Ballouz, S., Huang, Z.J., Gillis, J., 2017. BioRxiv. http://dx.doi.org/10.1101/150524, 150524.

Danon, L., Díaz-Guilera, A., Duch, J., Arenas, A., 2005. J. Stat. Mech. 2005, P09008.

Darmanis, S., Sloan, S.A., Zhang, Y., Enge, M., Caneda, C., Shuer, L.M., Hayden Gephart, M.G., Barres, B.A., Quake, S.R., 2015. Proc. Natl. Acad. Sci. U. S. A. 112, 7285–7290.

Deng, Q., Ramsköld, D., Reinius, B., Sandberg, R., 2014. Science 343, 193–196.

Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., Wildberg, A., Wang, W., 2015. Bioinformatics 31, 2225–2227.

Ding, J., Shah, S., Condon, A., 2016. Bioinformatics 32, 2567–2576.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996, KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 226–231.

Fan, H.C., Fu, G.K., Fodor, S.P.A., 2015. Science 347, 1258367.

Fan, J., Salathia, N., Liu, R., Kaeser, G.E., Yung, Y.C., Herman, J.L., Kaper, F., Fan, J.-B., Zhang, K., Chun, J., Kharchenko, P.V., 2016. Nat. Methods 13, 241–244.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., Linsley, P.S., Gottardo, R., 2015. Genome Biol. 16, 278.

Fulwyler, M.J., 1965. Science 150, 910–911.

Fuzik, J., Zeisel, A., Máté, Z., Calvigioni, D., Yanagawa, Y., Szabó, G., Linnarsson, S., Harkany, T., 2016. Nat. Biotechnol. 34, 175–183.

Gardeux, V., David, F., Shajkofci, A., Schwalie, P.C., Deplancke, B., 2016. BioRxiv. http://dx.doi.org/10.1101/096222, 096222.

Gierahn, T.M., Wadsworth, M.H., Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Love, J.C., Shalek, A.K., 2017. Nat. Methods 14, 395–398.

Goder, A., Filkov, V., 2008. 2008 Proceedings of the Tenth Workshop on Algorithm.

Goolam, M., Scialdone, A., Graham, S.J.L., Macaulay, I.C., Jedrusik, A., Hupalowska, A., Voet, T., Marioni, J.C., Zernicka-Goetz, M., 2016. Cell 165, 61–74.

Grün, D., Kester, L., van Oudenaarden, A., 2014. Nat. Methods 11, 637–640.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., van Oudenaarden, A., 2015. Nature 525, 251–255.

Guo, M., Wang, H., Potter, S.S., Whitsett, J.A., Xu, Y., 2015. PLoS Comput. Biol. 11, e1004575.

Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F., Theis, F.J., 2016. Nat. Methods 13, 845–848.

Hartigan, J.A., Wong, M.A., 1979. J. Roy. Stat. Soc. C Appl. Stat. 28 (1), 100–108. http://dx.doi.org/10.2307/2346830.

Hashimshony, T., Wagner, F., Sher, N., Yanai, I., 2012. Cell Rep. 2, 666–673.

Hicks, S.C., Teng, M., Irizarry, R.A., 2015. BioRxiv. http://dx.doi.org/10.1101/025528, 025528.

Huang, Y., Sanguinetti, G., 2017. BioRxiv. http://dx.doi.org/10.1101/098517, 098517.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., Linnarsson, S., 2014. 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., Amit, I., 2014. Science 343, 776–779.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., Oliver, B., 2011. Genome Res. 21, 1543–1551.

Jiang, L., Chen, H., Pinello, L., Yuan, G.-C., 2016. Genome Biol. 17, 144.

Jiang, Y., Zhang, N.R., Li, M., 2017. BioRxiv. http://dx.doi.org/10.1101/109629, 109629.

Johnson, M.B., Wang, P.P., Atabay, K.D., Murphy, E.A., Doan, R.N., Hecht, J.L., Walsh, C.A., 2015. Nat. Neurosci. 18, 637–646.

Joost, S., Zeisel, A., Jacob, T., Sun, X., La Manno, G., Lönnerberg, P., Linnarsson, S., Kasper, M., 2016. Cell Syst. 3, 221–237 e9.

Junqueria, L.C., Carneiro, J., Kelly, R.O., 1992. Basic Histology, seventh ed. (Appleton and Lange).

Karlsson, K., Linnarsson, S., 2017. BMC Genomics 18, 126.

Kiselev, V.Y., Hemberg, M., 2017. BioRxiv. http://dx.doi.org/10.1101/150292, 150292.

Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., Hemberg, M., 2017. Nat. Methods 14, 483–486.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., Taipale, J., 2011. Nat. Methods 9, 72–74.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., Kirschner, M.W., 2015. Cell 161, 1187–1201.

Kolodziejczyk, A.A., Kim, J.K., Tsang, J.C.H., Ilicic, T., Henriksson, J., Natarajan, K.N., Tuck, A.C., Gao, X., Bühler, M., Liu, P., Marioni, J.C., Teichmann, S.A., 2015. Cell Stem Cell 17, 471–485.

Lahens, N.F., Kavakli, I.H., Zhang, R., Hayer, K., Black, M.B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R.S., Grant, G.R., Hogenesch, J.B., 2014. Genome Biol. 15, R86.

Lancichinetti, A., Fortunato, S., 2009. Phys. Rev. E Stat. Nonlin Soft Matter Phys. 80, 056117.

Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., Finck, R., Gedman, A.L., Radtke, I., Downing, J.R., Pe'er, D., Nolan, G.P., 2015. Cell 162, 184–197.

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., Wong, M., Choi, P.J., Wee, L.J.K., Hillmer, A.M., Tan, I.B., Robson, P., Prabhakar, S., 2017. Nat. Genet. 49, 708–718.

Lin, P., Troup, M., Ho, J.W.K., 2017. Genome Biol. 18, 59.

Liu, S.J., Nowakowski, T.J., Pollen, A.A., Lui, J.H., Horlbeck, M.A., Attenello, F.J., He, D., Weissman, J.S., Kriegstein, A.R., Diaz, A.A., Lim, D.A., 2016. Genome Biol. 17, 67.

Lun, A.T.L., Bach, K., Marioni, J.C., 2016. Genome Biol. 17, 75.

Maaten, Laurens van der, Hinton, G., 2008. J. Mach. Learn. Res. 9, 2579–2605.
Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., Trombetta, J.J., Weitz, D.A., Sanes, J.R., Shalek, A.K., Regev, A., McCarroll, S.A., 2015. Cell 161, 1202–1214.
Moon, K.R., Dijk, D. v, Wang, Z., Chen, W., Hirn, M.J., Coifman, R.R., Ivanova, N.B., Wolf, G., Krishnaswamy, S., 2017. BioRxiv. http://dx.doi.org/10.1101/120378, 120378.
Muraro, M.J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M.A., Carlotti, F., de Koning, E.J.P., van Oudenaarden, A., 2016. Cell Syst. 3, 385–394 e3.
Olsson, A., Venkatasubramanian, M., Chaudhri, V.K., Aronow, B.J., Salomonis, N., Singh, H., Grimes, H.L., 2016. Nature 537, 698–702.
Owens, N.D.L., Blitz, I.L., Lane, M.A., Patrushev, I., Overton, J.D., Gilchrist, M.J., Cho, K.W.Y., Khokha, M.K., 2016. Cell Rep. 14, 632–647.
Owens, G.L., Todesco, M., Drummond, E.B.M., Yeaman, S., Rieseberg, L.H., 2017. BioRxiv. http://dx.doi.org/10.1101/142356, 142356.
Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., Louis, D.N., Rozenblatt-Rosen, O., Suvà, M.L., Regev, A., Bernstein, B.E., 2014. Science 344, 1396–1401.
Paul, L., Kubala, P., Horner, G., Ante, M., Hollaender, I., Alexander, S., Reda, T., 2016. BioRxiv. http://dx.doi.org/10.1101/080747, 080747.
Phipson, B., Zappia, L., Oshlack, A., 2017. F1000Res. 6 http://dx.doi.org/10.12688/f1000research.11290.1, 595.
Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G., Sandberg, R., 2013. Nat. Methods 10, 1096–1098.
Pierson, E., Yau, C., 2015. Genome Biol. 16, 241.
Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N., Szpankowski, L., Fowler, B., Chen, P., Ramalingam, N., Sun, G., Thu, M., Norris, M., Lebofsky, R., Toppani, D., Kemp, D.W., Wong, M., Clerkson, B., Jones, B.N., West, J.A.A., 2014. Nat. Biotechnol. 32, 1053–1058.
Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H., Trapnell, C., 2017. BioRxiv. http://dx.doi.org/10.1101/110668, 110668.
Radovanović, M., Nanopoulos, A., Ivanović, M., 2010. J. Mach. Learn. Res. 11, 1487–2531.
Risso, D., Ngai, J., Speed, T.P., Dudoit, S., 2014. Nat. Biotechnol. 32, 896–902.
Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., Vert, J.-P., 2017. BioRxiv. http://dx.doi.org/10.1101/125112, 125112.
Robinson, M.D., Oshlack, A., 2010. Genome Biol. 11, R25.
Rosvall, M., Bergstrom, C.T., 2008. Proc. Natl. Acad. Sci. U. S. A. 105, 1118–1123.
Roussel, M., Benard, C., Ly-Sunnaram, B., Fest, T., 2010. Cytom. A 77, 552–563.
Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A., 2015. Nat. Biotechnol. 33, 495–502.
Schaeffer, S.E., 2007. Comput. Sci. Rev. 1, 27–64.
Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E.-M., Andréasson, A.-C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M.K., Smith, D.M., Kasper, M., Ämmälä, C., Sandberg, R., 2016. Cell Metab. 24, 593–607.
SEQC/MAQC-III Consortium, 2014. Nat. Biotechnol. 32, 903–914.
Severson, D.T., Owen, R.P., White, M.J., Lu, X., Schuster-Boeckler, B., 2017. BioRxiv. http://dx.doi.org/10.1101/118919, 118919.
Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., Trombetta, J.J., Gennert, D., Gnirke, A., Goren, A., Hacohen, N., Levin, J.Z., Park, H., Regev, A., 2013. Nature 498, 236–240.
Sims, D., Sudbery, I., Ilott, N.E., Heger, A., Ponting, C.P., 2014. Nat. Rev. Genet. 15, 121–132.
Sinha, R., Stanley, G., Gulati, G.S., Ezran, C., Travaglini, K.J., Wei, E., Chan, C.K.F., Nabhan, A.N., Su, T., Morganti, R.M., Conley, S.D., Chaib, H., Red-Horse, K., Longaker, M.T., Snyder, M.P., Krasnow, M.A., Weissman, I.L., 2017. BioRxiv. http://dx.doi.org/10.1101/125724, 125724.
Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A., Mikkelsen, T.S., 2014. BioRxiv. http://dx.doi.org/10.1101/003236, 003236.
Stein, C.K., Qu, P., Epstein, J., Buros, A., Rosenthal, A., Crowley, J., Morgan, G., Barlogie, B., 2015. BMC Bioinforma. 16, 63.
Svensson, V., Natarajan, K.N., Ly, L.-H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A., Teichmann, S.A., 2017. Nat. Methods 14, 381–387.
Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., Bertagnolli, D., Goldy, J., Shapovalova, N., Parry, S., Lee, C., Smith, K., Bernard, A., Madisen, L., Sunkin, S.M., Hawrylycz, M., Zeng, H., 2016. Nat. Neurosci. 19, 335–346.
Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J.-R., Cohen, O., Shah, P., Lu, D., Genshaft, A.S., Hughes, T.K., Ziegler, C.G.K., Kazer, S.W., Garraway, L.A., 2016. Science 352, 189–196.
Trapnell, C., 2015. Genome Res. 25, 1491–1498.
Tsang, J.C.H., Yu, Y., Burke, S., Buettner, F., Wang, C., Kolodziejczyk, A.A., Teichmann, S.A., Lu, L., Liu, P., 2015. Genome Biol. 16, 178.
Tung, P.-Y., Blischak, J.D., Hsiao, C.J., Knowles, D.A., Burnett, J.E., Pritchard, J.K., Gilad, Y., 2017. Sci. Rep. 7, 39921.
Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnerberg, P., Lou, D., Hjerling-Leffler, J., Haeggström, J., Kharchenko, O., Kharchenko, P.V., Linnarsson, S., Ernfors, P., 2015. Nat. Neurosci. 18, 145–153.
Vallejos, C.A., Marioni, J.C., Richardson, S., 2015. PLoS Comput. Biol. 11, e1004333.
Vallejos, C.A., Risso, D., Scialdone, A., Dudoit, S., Marioni, J.C., 2017. Nat. Methods 14, 565–571.
Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., Boch, T., Hofmann, W.-K., Ho, A.D., Huber, W., Trumpp, A., Essers, M.A.G., Steinmetz, L.M., 2017. Nat. Cell Biol. 19, 271–281.
Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., Jardine, L., Dixon, D., Stephenson, E., Nilsson, E., Grundberg, I., McDonald, D., Filby, A., Li, W., De Jager, P.L., Rozenblatt-Rosen, O., Hacohen, N., 2017. Science 356.
Wang, Y.J., Schug, J., Won, K.-J., Liu, C., Naji, A., Avrahami, D., Golson, M.L., Kaestner, K.H., 2016. Diabetes 65, 3028–3038.
Ward, J.H., 1963. J. Am. Stat. Assoc. 58, 236–244.
Wilson, N.K., Kent, D.G., Buettner, F., Shehata, M., Macaulay, I.C., Calero-Nieto, F.J., Sánchez Castillo, M., Oedekoven, C.A., Diamanti, E., Schulte, R., Ponting, C.P., Voet, T., Caldas, C., Stingl, J., Green, A.R., Theis, F.J., Göttgens, B., 2015. Cell Stem Cell 16, 712–724.
Wiwie, C., Baumbach, J., Röttger, R., 2015. Nat. Methods 12, 1033–1038.
Xu, C., Su, Z., 2015. Bioinformatics 31, 1974–1980.
Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., Liu, J., Horvath, S., Fan, G., 2013. Nature 500, 593–597.
Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J., Linnarsson, S., 2015. Science 347, 1138–1142.
Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Bielas, J.H., 2017. Nat. Commun. 8, 14049.
Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., Enard, W., 2017. Mol. Cell 65, 631–643 e4.
Žurauskienė, J., Yau, C., 2016. BMC Bioinforma. 17, 140.