The FEBS Journal

FEBS PRESS science publishing by scientists

REVIEW ARTICLE

Computational approaches for high-throughput single-cell data analysis

Helena Todorov[1,2,3] and Yvan Saeys[1,2]

1 Data Mining and Modelling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium
2 Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
3 Centre International de Recherche en Infectiologie, Inserm, U1111, Université Claude Bernard Lyon 1, CNRS, UMR5308, École Normale Supérieure de Lyon, Univ Lyon, France

Keywords
bioinformatics; computational tools; proteome; single cell; transcriptome

Correspondence
Y. Saeys, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Technologiepark 927, 9052 Gent, Belgium
Fax: +32 9 221 76 73
Tel: +32 9 331 37 40
E-mail: yvan.saeys@ugent.be

(Received 22 February 2018, revised 4 June 2018, accepted 25 July 2018)

doi:10.1111/febs.14613

During the past decade, the number of novel technologies to interrogate biological systems at the single-cell level has skyrocketed. Numerous approaches for measuring the proteome, genome, transcriptome and epigenome at the single-cell level have been pioneered, using a variety of technologies. All these methods have one thing in common: they generate large and high-dimensional datasets that require advanced computational modelling tools to highlight and interpret interesting patterns in these data, potentially leading to novel biological insights and hypotheses. In this work, we provide an overview of the computational approaches used to interpret various types of single-cell data in an automated and unbiased way.

Introduction

Single-cell technologies are currently revolutionising the way life scientists are studying biological systems from different perspectives. Three major classes of technologies can be distinguished: imaging-based techniques, techniques based on flow or mass cytometry and techniques based on next-generation sequencing. However, this is only a rough classification, as some recent innovations combine elements of different classes of techniques. While many of the early data preprocessing steps are specific to each class of techniques, several downstream computational analyses are generally applicable to any form of single-cell data, and one of the goals of this work is to provide a unifying overview of these generally applicable approaches.

Historically, microscopy-based techniques were the first methodology to study organisms at single-cell resolution [1]. While initially consisting largely of manual labour and thus being very low-throughput, automated image acquisition and segmentation have enabled high-throughput image-based screening, by analysing up to hundreds of thousands of cells in single-well plates [2]. Similarly, many other microscopy-based techniques allow the extraction of information at the single-cell level, although at a lower throughput. These include most types of light and electron microscopy, with a broad variety of applications. Common to all these image-based approaches is the fact that advanced image-analysis pipelines are needed to arrive at single-cell resolution [3]. A typical image processing pipeline first performs segmentation of the single cells from the image, followed by a feature extraction step, typically extracting several hundreds of features for each

Abbreviations
DE, differential expression; HVGs, highly variable genes; scRNA-Seq, single-cell RNA sequencing; TI, trajectory inference.

individual cell [4]. In comparison to other single-cell approaches where cells are dissociated in suspension, a major advantage of image-based single-cell profiling methodology is that it inherently provides the user with two- or three-dimensional spatial information, as knowing a cell's spatial context is often the key to discover novel biological findings.

Flow cytometry allows profiling and analysing cells in a high-throughput fashion and is based on passing cells through a laser beam in a rapidly flowing fluid stream. This core technology is in essence very similar to the original design from the late 1960s [5], illustrating the robustness of the technology [4,6]. The field of flow cytometry has emerged as a powerful methodology for single-cell analysis due to continuous innovations such as (a) multicolour assays enabling the measurement of a large number of proteins simultaneously [7], (b) spectral flow cytometry [8] in which classical mirrors, optics and detectors are replaced by dispersive optics and a linear array of detectors allowing highly complex fluorochrome combinations, (c) imaging flow cytometry [9] combining flow cytometry and microscopy for high-throughput imaging of single cells, and (d) acoustic-based focusing and sorting [10]. In addition, other technological advances such as mass cytometry have replaced the fluorescent labelling and readout using optics by labelling using heavy isotopes, and subsequent readout by mass spectrometry [11]. This eliminates the problem of spectral overlap in classical flow cytometry, allowing the theoretical measurement of up to 100 proteins simultaneously. Mass cytometry can also be performed on tissue slices, thereby scanning the tissue spot-by-spot and performing a single experiment per spot. This approach, named imaging mass cytometry, allows performing spatial proteomics in a high-throughput fashion [12]. The ability to measure increasing amounts of proteins simultaneously [7] complicates the analysis of this type of data, which can no longer be analysed manually as was done with datasets containing a few markers per cell, but needs new computational approaches to correctly identify cell populations [13].

Recent developments in microvolume sequencing have led to a new wave of single-cell '-omics' profiling technologies [14–18], permitting the quantification of whole genomes, epigenomes and transcriptomes at the single-cell level. Novel computational tools are being developed in order to deal with the continuously increasing dimensionality of these datasets, since a single experiment can quantify molecular characteristics of up to tens of thousands of cells, measuring tens of thousands of parameters (e.g. transcripts in the case of single-cell transcriptomics). A high level of resolution is provided by single-cell omics tools, as they aim to sequence all of the cell's content, instead of focusing on a set of user-defined targets as is done in cytometry. This allows performing novel types of analyses, such as studying the heterogeneity of cell populations in much greater detail, identifying rare cell types, and studying the dynamics of cellular systems. Furthermore, the field continues to evolve by combining single-cell RNA sequencing with other technologies such as spatial transcriptomics [19] and CRISPR-mediated knockout screens (Perturb-Seq [20]/CRISP-seq [21]). Recent approaches combine transcriptomics with other types of omics data at a single-cell resolution such as single-cell proteomics (CITE-seq [22]/REAP-seq [23]), single-cell genomics (G&T-seq [24]) and single-cell methylomics (scM&T-seq [25]). These emerging 'single-cell multi-omics' technologies [26] integrate several types of measurements on the same single cell and are likely to be part of the everyday methodology of molecular biologists in the future.

While all techniques described above provide the user with information at single-cell level, the throughput, resolution, cost and type of information acquired differ drastically between technologies. We will take a computational perspective here, and compare the main dataset characteristics for the three major classes of single-cell data introduced above. Classical imaging-based techniques typically offer a low throughput, measuring a few hundreds of cells, while more advanced high-content screening methods allow high-throughput measurements of hundreds of thousands to millions of cells. When applying segmentation and feature extraction, for example using popular pipelines such as CELLPROFILER [27], almost a thousand image-derived features can be extracted per cell. However, many of those capture redundant information and thus are very correlated. Flow and mass cytometry allow measuring cells at high throughput, up to millions of cells for classical flow cytometry. Only a few tens of parameters can be quantified simultaneously per single cell, but these parameters often represent very complementary information, as they are manually chosen by an expert. Single-cell omics technologies offer medium throughput, measuring thousands to tens of thousands of cells in a single run. However, these data are very rich in information, measuring thousands of transcripts in the case of single-cell transcriptomics.

While the profiling methodology and dataset characteristics in each of these technologies are very different, many of the applications and computational workflows are quite similar. In the remainder of the paper, we will discuss the differences and commonalities in computational workflows for the different applications.

Computational workflow for single-cell experiments

Regardless of the specific technology used to generate a single-cell dataset, a common pipeline can be devised, starting with the experimental design, data generation, technology-specific preprocessing, quality control and subsequent data analysis (Fig. 1). A detailed design of the experiment is a crucial step towards minimising technical variation and improving scientific reproducibility. This not only includes standardisation of experimental protocols and equipment, but also careful planning and consultation with statisticians and/or bioinformaticians regarding sample size, specific setup related to the biological questions that should be answered or specific types of computational analyses that should be carried out. Subsequently the experiment should be performed, ensuring that standardised procedures are followed for sample preparation, handling equipment and data acquisition while appropriate controls are added at multiple steps of the experiments.

The next step in the pipeline is the preprocessing and quality control. This step will likely take a considerable amount of time, as it is crucial to start from good quality data if good quality results are desired. Therefore, it is important to perform technology-specific preprocessing steps, a topic that will be covered in the section 'Data preprocessing and quality control'. After data preprocessing, an initial exploration of the data can be performed using visualisation techniques, in order to perform early detection of any possible batch effects or unexpected subpopulations. Applying visualisation techniques may also help to visualise the population structure within samples, and to compare this structure between different samples. In this step, interesting populations or trends may be observed that require further investigation.

Next, several types of in-depth analyses can be performed, in most cases starting with an automated clustering of the cells into cell types. This clustering allows quantifying and comparing different cell types in the samples and identifying new cell types or transition states. Novel computational approaches to model gradual transitions between cell states (trajectory inference) can also be applied at this stage. Other alternatives include specific predictive modelling approaches such as classification, regression and survival analysis modelling. All these approaches have the potential to extract novel biomarkers from single-cell data, with important diagnostic and therapeutic potential. Finally, more advanced computational approaches can be applied to single-cell omics data. The correlations in gene expression within cells can be studied to assess gene regulatory networks (network inference). In the case of multi-omics datasets, data integration

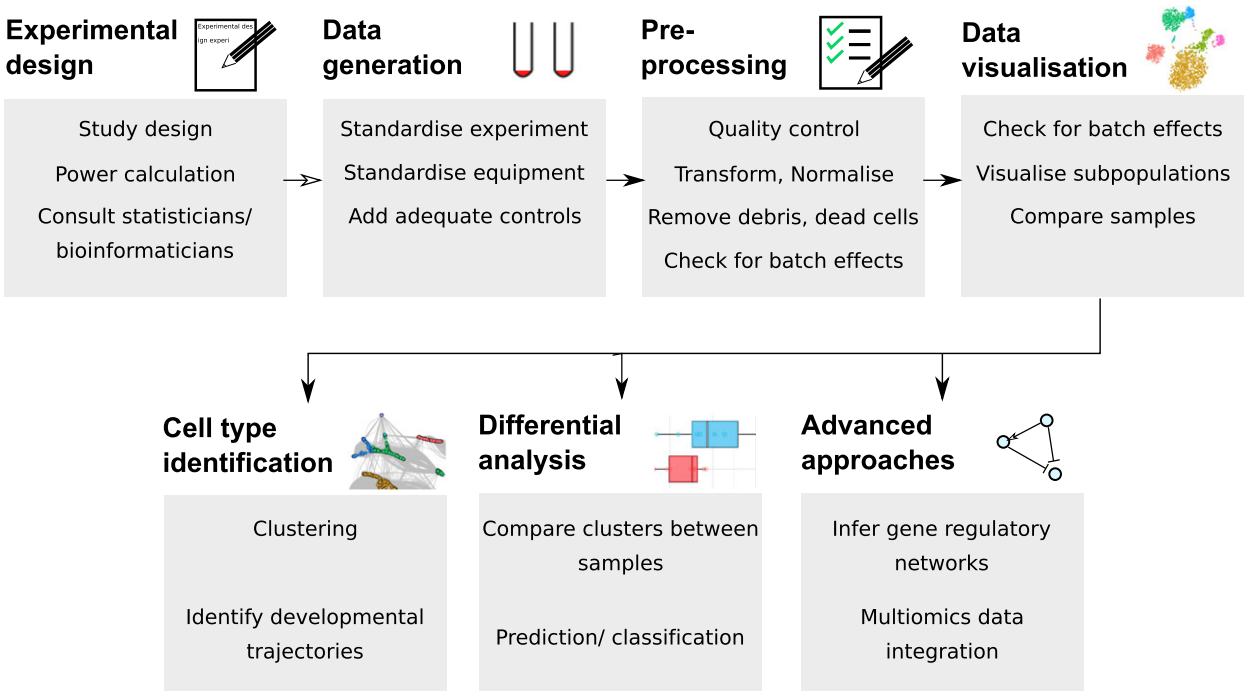

Fig. 1. The computational workflow for single-cell experiments detailed in steps.

approaches can be used to combine the information on single-cell mechanisms.

Data preprocessing and quality control

Single-cell imaging

The preprocessing of single-cell imaging data usually starts by accounting for batch effects through illumination correction, and image-wise processing such as noise removal, aligning or cropping [28,29]. This procedure is commonly followed by the segmentation of the individual cells within the images, and finally by a feature extraction process that yields a vector of numeric features for each individual cell, usually in a tabular format.

CELLPROFILER [27] is widely used to extract numerical features from two-dimensional microscopy images (such as in high-content screening assays). The main difficulty faced by CELLPROFILER is the segmentation of the cells or objects of interest present in the image. CELLPROFILER contains several fast algorithms that can extract well-separated objects; however, in many cases, these objects appear clumped, hindering their segmentation and making it prone to both false negatives (when the borders between objects cannot be found) and false positives (when the sensitivity of the detection is too high). In order to deal with this difficulty, CELLPROFILER also provides a more complex segmentation algorithm that follows a hierarchical process: first, it finds primary level objects that are typically well-separated (such as cell nuclei, visible on DNA-stain channels); then, the boundaries of secondary level objects (such as cell edges) are searched around the primary level objects.

However, it is also possible that the primary level objects appear clumped, which is why CELLPROFILER divides their detection into several steps following the guidelines of previously published algorithms [30–34]. Clumped objects are first detected, segmented and separated by dividing lines, thus avoiding false negatives. Finally, some of the objects are either removed or merged to reduce the false positive rate. Once the primary level objects are properly detected, it becomes simpler to find secondary level objects around them. CELLPROFILER provides an improved algorithm to properly detect the borders even when the objects are clumped against each other. Once the objects have been segmented, multiple features can be extracted from each of them in a per-channel basis (area, shape, intensity, texture, etc.) or at the whole-image level (number of cells, background intensity, etc.). CELLPROFILER has a modular structure that allows the user to select and configure the individual algorithms that will be applied, which in turn defines the specific preprocessing applied and the features that are obtained at the end of the pipeline. The resulting features can later be used for visualisation, clustering or differential downstream analyses for instance.

Flow/mass cytometry

In conventional flow cytometry, the first preprocessing step is typically compensation of the spectral overlap, to correct for spillover of the fluorescent signal into neighbouring channels. This is typically accounted for in the experimental procedure, by measuring the fluorescence of single stains in the different channels, allowing for the calculation of a compensation matrix. In mass cytometry, this issue is largely avoided by using rare isotopes instead of light measurements, although the measurement of certain isotopes can still be polluted due to metal impurity levels, oxidation and abundance sensitivity [35]. Mass cytometry panels should therefore be designed with caution by pairing strong intensity markers with less sensitive channels in order to avoid interference between channels [36]. The data is then transformed through a biexponential or hyperbolic arcsine transformation, which improves the separation between negative and positive cells for the different markers. Fluctuations in measurements can also be caused by an unsteady flow rate. Typically, up to 10 000 cells are measured per second at a steady rate in flow cytometry. Mass cytometry has a slightly lower throughput, measuring a few thousand cells per second. However, obstructions in the fluid stream and manual interventions can disturb the flow, which also impacts the amount of protein levels measured. To remove these technical artefacts, the data needs to be either manually gated against time or screened by tools such as FLOWCLEAN [37], FLOWQ [38] and FLOWAI [39], which can automatically identify and remove sections in which the flow was perturbed.

The acquisition level of cytometers can slightly change from one day to another, or even within hours. The use of control tubes to calibrate the machine before running an experiment can help to make different samples more comparable, but batch effects are often observed between two experiments. The resulting slight shift in protein expression can be accounted for manually, by shifting the gates of every sample that differs, or in an automated way using the FLOWSTATS [40] package. In mass cytometry, beads are commonly used in the experiments, allowing normalisation of the data based on the signal of these beads to have more

comparable samples. Some markers can also be used to barcode cells, and then pool several samples together, to avoid technical bias between different experimental conditions. When performing experiments on different days, it may be advisable to include additional control samples, such as an aliquot from the same sample that is taken along all different experiment days, in order to allow normalisation between experiment days later on. Once batch effects have been accounted for, debris, doublets and other low quality cells can be removed either by manual gating or using OPENCYTO [41], or FLOWDENSITY [42].

As flow cytometry allows the measurement of proteins at the single-cell level while preserving the integrity of the cells, it is sometimes used to sort specific cells into wells before sequencing their transcriptome. The cells can either be sorted by cell population, based on a set of common markers, or index-sorted, in which case single cells are sorted into wells and barcoded, so that their protein expression profile is kept. In this case, doublets and empty wells might occur, which should be carefully removed from the analysis before any further processing step.

Single-cell omics

Preprocessing single-cell omics data based on NGS technologies further builds on the wide availability of NGS preprocessing tools that are already available from experiments on bulk RNA or DNA. However, single-cell omics technologies lead to a number of additional challenges when going through the process from the individual reads to the mapped genomes or transcriptomes. We will focus here more specifically on methods for single-cell transcriptomics, as this is the most widely used type of single-cell omics data at present. Several scRNA-Seq protocols were developed, usually focusing either on sequencing a large number of cells, or a high amount of genes at an increased sequencing depth [43]. Due to the low amount of transcripts in the cells, scRNA-Seq data usually contain a lot of technical variance, requiring specific computational tools to perform quality control, normalisation and downstream analyses [44–47].

When performing a computational analysis on scRNA-Seq data coming from multiple experiments, batch effects can arise, leading to an increased interexperimental variability. Two recently published algorithms can be used in order to reduce batch effects. These algorithms either identify a gene correlation structure [48], or a subset of cells coming from the same population [49], that are shared between the datasets coming from different experiments. Proper data transformation is then applied to align similar cell populations, resulting in more consistent datasets that can be further analysed together.

Several quality control metrics, such as the library size and the percentage of mitochondrial genes, are used to filter out abnormal cells, in order to reduce the technical variance of the data [50]. Additionally, a great part of intercellular variability can be caused by the cell cycle, and it is up to the user to decide whether this variability should be removed from the data or not. Cyclone [51] is a method that can be used to predict the cell cycle stage, which can subsequently be used to either remove cycling cells, or tag them so that they can be easily identified later in the analysis. F-scLVM [52] is another algorithm that identifies the amount of variability across the expression of each gene that is due to cell cycle differences. It can be used to infer 'corrected' gene expression values, removing the effect of the cell cycle.

The next step in the process regards the normalisation of the count data, since a large part of the observed variability can be due to differences in size, viability, capturing efficiency and amplification biases between cells. Some methods aim to standardise the total number of reads per cell (RPKM [53], TPM [54], downsampling) or proportions of the total number of reads per cell (UQ, full quantile [55]). However, these methods can be seriously impacted by false negative counts [56]. Indeed, the number of transcripts in a cell being very low for certain genes, there is a high probability that these transcripts will be missed, resulting in a zero count in the final expression data. These missed transcripts are called dropouts, and lead to a high technical variance that can affect the final results. High-throughput scRNA-Seq protocols typically show higher dropout rates [43], but high amounts of sequenced cells can help to infer dropout probabilities. ZIFA [57] is a method which identifies zero counts that are most likely resulting from dropout events, and gives less weight to these counts. ZINB-WAVE [58] is another method which not only assesses the probability for a zero to be a dropout based on the sequencing depth, but also accounts for batch effects between samples, and computes global-scaling normalisation factors, which allow it to be used directly on non-normalised data.

Some methods rely on spike-ins to distinguish technical variability from biologically relevant changes in gene expression [59] (BASICS [60], GRM [61], SAMSTRT [62]). Spike-ins are control RNA transcripts which are added in the same quantity to all the samples to be sequenced. They can be used to normalise the data, as all cells should have exactly the same amount of

spike-ins after sequencing, and the differences in spike-in amounts should only be the consequence of technical artefacts. However, the most commonly used spike-in set (ERCC [63]) cannot always faithfully account for the intrinsic gene variability, as they have been shown to have a length and GC content that differ from mammalian transcripts [58]. Moreover, choosing the quantity of spike-ins that should be added to the cells can be challenging, as a significant amount of spike-ins has to be used in order to reflect faithfully the intercellular variability, but may eclipse the intracellular transcripts of interest. However, ERCC spike-ins are still commonly used to filter out low quality cells [50]. Overall, the views on the use of spike-ins for single-cell RNA Seq normalisation are still conflicting [64–66].

The methods cited above apply global scaling factors to all cells equally, assuming that the relation between the number of genes measured per cell and the sequencing depth is the same for all genes. However, this assumption of a constant gene-count/sequencing depth ratio has been shown to hold on bulk RNA data, but not in single-cell datasets [67]. Applying global scaling factors to scRNA-Seq data might therefore lead to biased correction of lowly and highly expressed genes. Two algorithms can be used to perform single-cell specific normalisation of scRNA-Seq datasets. The SCnorm method [67] relies on the fact that the normalisation should not be applied in the same way to all the genes, as they differ in various properties such as transcript length and GC content. SCnorm first groups genes with similar dependencies on sequencing depth and subsequently estimates different scale factors for each group of genes. Alternatively, SCRAN [50], first groups cells with similar expression profiles together, and applies intragroup normalisation before performing intergroup normalisation.

Visualising high-dimensional single-cell data

Once the data has been preprocessed, visualisation tools can help to get a first insight into the structure of the data. A quick principal component analysis (PCA) plot of the data can, for instance, allow identifying any remaining source of technical variability between samples, which should be removed by normalisation. Structures in the data or biological differences between the samples may then be investigated using different approaches: dimensionality reduction techniques, clustering techniques, or the novel class of techniques to model cell trajectories and state transitions. A list of visualisation tools and their principal characteristics is provided in Table 1.

Table 1. Dimensionality reduction based- and clustering based-tools for visualisation of single-cell high-dimensional data.

Class of method	Name	Description
Dimensionality reduction	PCA	Linear reduction in the dimensions holding the highest variance into orthogonal principal components
	MDS	Nonlinear reduction in the dimensions by preserving the intercellular distances of high dimensions in the lower dimensions
	tSNE	Nonlinear dimensionality reduction, preserves the local similarities between cells
	Diffusion maps	Nonlinear dimensionality reduction, computes transition probabilities between cells
	SPRING	k-Nearest Neighbour force directed graph, preserves the high-dimensional relationships between cells
Clustering	SPADE	Hierarchical clustering of the cells followed by the representation of these clusters in a minimal spanning tree
	FLOWSOM	SOM clustering followed by the representation of these clusters in a minimal spanning tree
	Scaffold Maps	Semisupervised method: new cells are grouped with the user-provided cell populations to which they are most similar
	FLOWMAP	Hierarchical clustering of the cells, followed by the representation of these clusters in a strong connected graph structure
	Phenograph	Groups cells which share the same neighbours together and identifies communities which maximise the Louvain modularity

Dimensionality reduction tools aim to capture the structure of the high-dimensional data by projecting it to a lower dimensional space that keeps the most important structural properties of the original, high-dimensional space. The lower dimensional projection allows the human expert to visualise and explore the data. Dimensionality reduction can be performed either in a linear way (the lower dimensional projections are a linear combination of the original dimensions), or in a nonlinear way. PCA is a linear dimensionality reduction technique, in which the features with the largest variability are preserved in principal components. The main sources of variability in the data can then be optimally laid out. A PCA can

therefore be applied to check for batch effects in the data, or to identify any main source of variability. The use of nonlinear dimensionality reduction methods (e.g. tSNE [t-stochastic neighbour embedding, 68], MDS [multidimensional scaling, 69], diffusion maps [70], SPRING [71]) allows optimal plotting of the data in two dimensions while preserving the local similarities between cells.

Clustering-based visualisation methods group similar cells together and may be combined with a subsequent visualisation step, for example by laying out the resulting clusters in two dimensions. This reduces computation time and can simplify the understanding of the resulting plot. Several methods have been proposed for the visualisation of clusters in single-cell data (SPADE [Spanning-tree Progression Analysis of Density-normalized Events] [72], FLOWSOM [73], FLOWMAP [74]). These methods represent the clusters under the form of a graph in which the most similar clusters are linked by an edge. FLOWSOM also allows performing meta-clustering, grouping clusters into larger populations, which has shown to return results very similar to manual labelling of cytometry data [75]. Single-Cell Analysis by Fixed Force- and Landmark-Directed (Scaffold) maps [76] were specifically designed to simplify the identification of user defined cell populations in cytometry data. Finally, Phenograph [77] identifies closely linked communities of cells in a graph structure. This algorithm therefore identifies populations without any previous knowledge on the number of expected populations, which can be very useful in discovery studies. While most of these methods were initially developed for flow cytometry data, FlowSOM and Phenograph are scalable to high dimensional datasets. These methods can therefore be applied to mass cytometry and scRNA-Seq datasets, or to features extracted from images, allowing the visualisation of structure in the data.

However, scRNA-Seq and image derived data typically contain much more dimensions than the usual 10–30 colour panels used in cytometry. When dealing with features extracted from images, a first step can consist in performing principal component analysis, which will help to reduce the redundancy of these highly correlated features. One can then choose to work with the principal components containing 95% of the data variability. These principal components can be analysed as new features, using visualisation or clustering techniques. scRNA-Seq datasets tend to contain noise which might bias clustering studies, especially due to the high amount of lowly expressed genes and dropouts. Therefore, the highly variable genes (HVGs) can first be filtered on this type of data [50,78], which considerably reduces the number of features and the noise they contain, while preserving the main biologically relevant sources of variability. Another algorithm was implemented in the SEURAT R package [79] to filter HVGs. Visualisation, clustering or any downstream analysis algorithms can then be applied either to the HVGs, or, if the dimensions of the data are still too high, on the principal components of a PCA run on these HVGs.

In order to highlight the differences between the different methods cited above, we applied two dimensionality reduction tools (PCA and tSNE) and two clustering-based tool (FLOWSOM, Phenograph) on a publicly available scRNA-Seq dataset [16] of 3000 peripheral blood mononuclear cells (PBMCs) from the 10X Genomics platform (Fig. 2). We first preprocessed the dataset as described in the data preprocessing section by filtering out low quality cells and genes. We then selected the most highly variable genes, to which we applied the different visualisation methods. This filtering on highly variable genes has two advantages. It significantly reduces the size of the dataset, therefore reducing the analysis time, and it helps to focus on the genes that are driving heterogeneity across cells [50]. The PBMC dataset had previously been expert-labelled in the Seurat R pipeline [79], which allowed us to use the cell identities to simplify the comparison of the outputs from the different methods. The different methods provided complementary information on the structure of the data. For instance, all methods except PCA identified the rare megakaryocyte cell population, and all methods except FlowSOM represented these megakaryocyes close to the monocyte cell population. As a general guideline, it is often advisable to apply several techniques in parallel to acquire a deeper understanding of the data structure.

Cell type identification

While the clustering approach to single-cell analysis assumes that cells are forming well separated groups, other types of techniques focus on better detecting cells that are in transition between cell states. In the first case, the expression of certain markers is expected to differ drastically, providing hard separations between cell populations. In the second case, the markers are seen as continuous variables which smoothly change from one cell to another, leading to structural patterns in the data which can be seen as developmental trajectories (Fig. 3). The choice between the two sets of methods depends on the biological question, but a good practice can be to first apply a clustering algorithm to identify the main populations in the data,

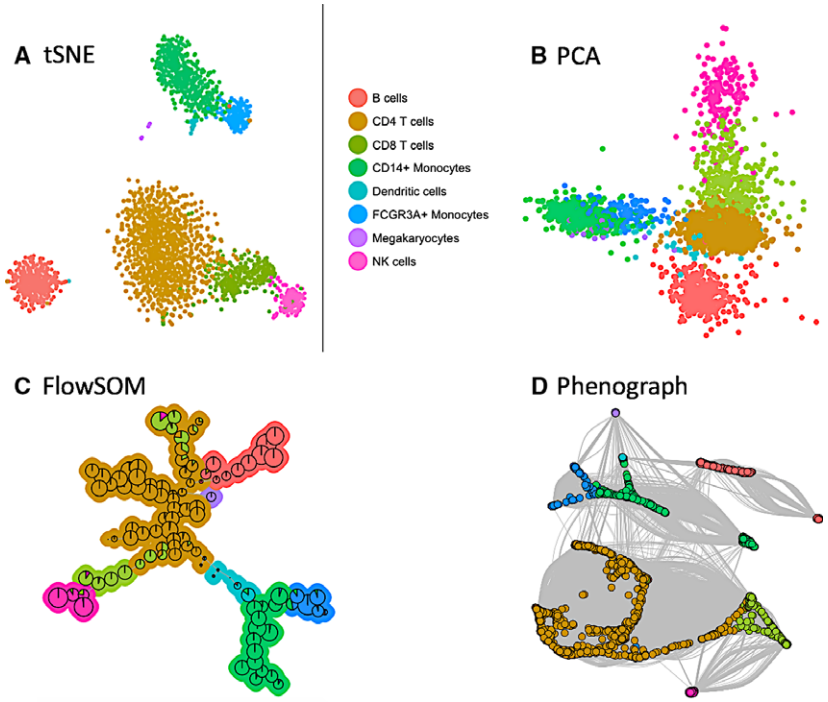

Fig. 2. Comparison of (A) tSNE, (B) PCA, (C) FLOWSOM and (D) Phenograph on the PBMC dataset. (A) The cell colours correspond to the labels provided by experts in the Seurat R pipeline. (B) The main differences between cell types can be seen on the horizontal (1st principal component) and vertical (2nd principal component) axis. (C) The colours inside the pies correspond to the cell colours on the tSNE plot. The background colours correspond to the meta-clusters identified by FlowSOM. Discrepancies between the pie colour and the background colour highlight the cells for which FlowSOM's results diverged from the manual annotation. (D) The similarities between the different cell types are nicely laid out on a Phenograph plot.

and then perform trajectory inference on a specific group of similar cells. Indeed, trajectory inference tools will tend to identify trajectories in any dataset, so they should be applied to specifically delineated sets of cells. The identification of trajectories in highly variable datasets is a current challenge, which is only described recently in the literature [80].

Clustering-based approaches

Several tools have been implemented in order to identify similar groups of cells in cytometry data, comparing either the similarities between cells (SPADE [81], FLOWSOM [73]), the distances between cells in a lower dimensional space (Accense [82]) or the shared neighbours in a graph (Phenograph [77]). A benchmark study of clustering tools, the FLOWCAP I [83] challenge, provided several mammalian datasets to assess the ability of different clustering methods to identify cell populations accurately. Most tools provided a good delineation of cell populations compared to manual gating, and ensemble methods which merged the outputs of several clustering methods showed the best results. However, due to the increasing number of markers used in cytometry data, there is a need to perform benchmark studies regularly, as tools which were very efficient with low-dimensional datasets might not necessarily perform equally well in higher dimensions [84]. Another study [75] compared 18 clustering methods for conventional flow and mass cytometry data, taking into account the clustering accuracy as well as the computational time, which becomes more important when dealing with large datasets. The FLOWSOM [73] algorithm showed the best clustering accuracy and was one of the fastest methods when applied to large datasets, with a linear complexity with respect to the number of cells. CytoCompare [85] is a tool which was created to perform the comparison of the clustering results of three methods: SPADE, ViSNE/Accense [82] and Citrus [86].

The clustering algorithms described above can also be applied to image derived features, although, as was the case for visualisation techniques, the high correlation between features might bias clustering results. The redundancy of the features can be reduced by first applying a PCA to this type of data, and performing

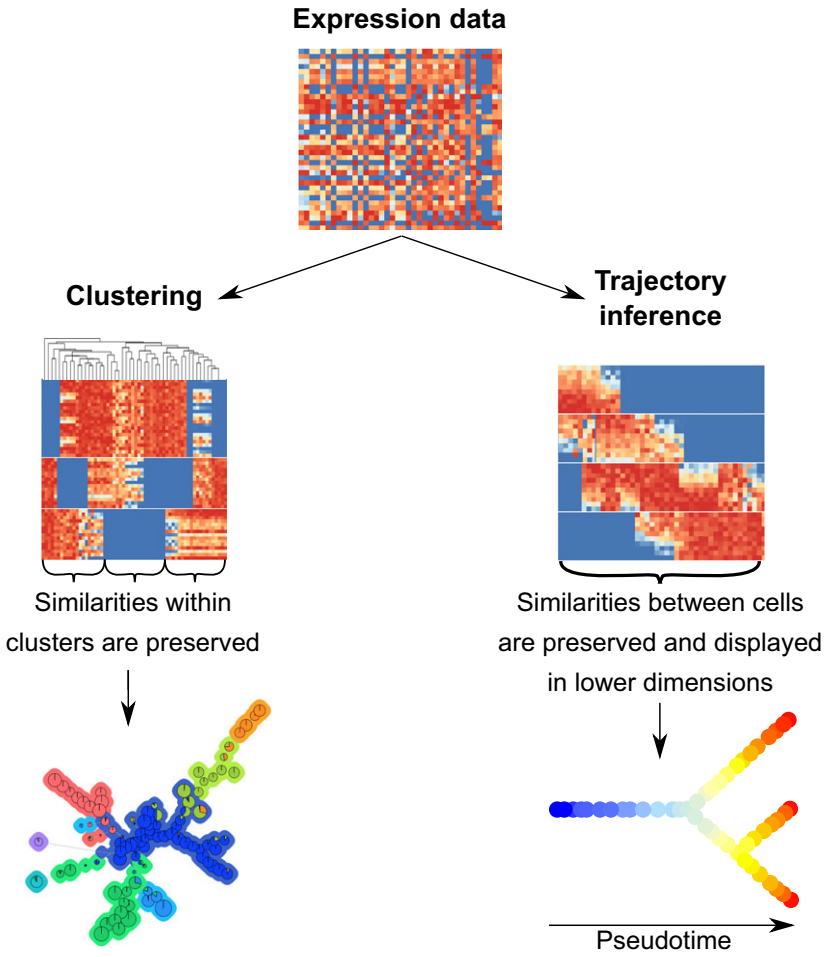

Fig. 3. In order to identify structures in an expression data matrix, two types of methods can be used. Clustering-based methods will tend to maximise the similarities between cells within clusters while maximising the differences between clusters. These methods thus help to identify homogeneous groups of cells in the data. On the other hand, trajectory inference methods will tend to preserve the local similarities between cells, ordering them along trajectories which represent gradual changes between similar cells.

clustering on the principal components of the PCA. In scRNA-Seq data, clustering is more tricky because the gene expression contains noise and the data is very sparse. Cells may mistakenly be grouped together based on technical noise attributed to sequencing depth or library size, rather than actual biological effects. This raises the need for new tools, which are able to overcome this issue. Several tools do not compare the expression patterns of cells directly anymore, but apply tricks to perform more accurate clustering: SC3 [87] computes a consensus clustering over several kmeans runs at the cost of a high computational cost, BackSPIN [88] uses a biclustering method and DIMM-SC [89] was designed specifically for droplet-based single-cell RNA seq data.

Another characteristic of scRNA-Seq data is the high amount of dropout events. Some clustering methods were specifically designed to deal with this artefact, either by imputing the expected value of dropout candidates (CIDR [90]), or by computing the similarities between cells with techniques that are robust to dropouts (SIMLR [91], SNN-Cliq [92], SCENIC [93]). The PAGODA [94] algorithm also accounts for technical biases such as the expression magnitude and the cell cycle.

Approaches for modelling gradual transitions

Another set of approaches, called trajectory inference (TI) methods, aim to reconstruct the developmental process that cells are undergoing. The resulting trajectory consists of states and transitions, with each cell mapped to a pseudotemporal location in the trajectory (Fig. 4A). Various visualisation techniques can aid in interpreting

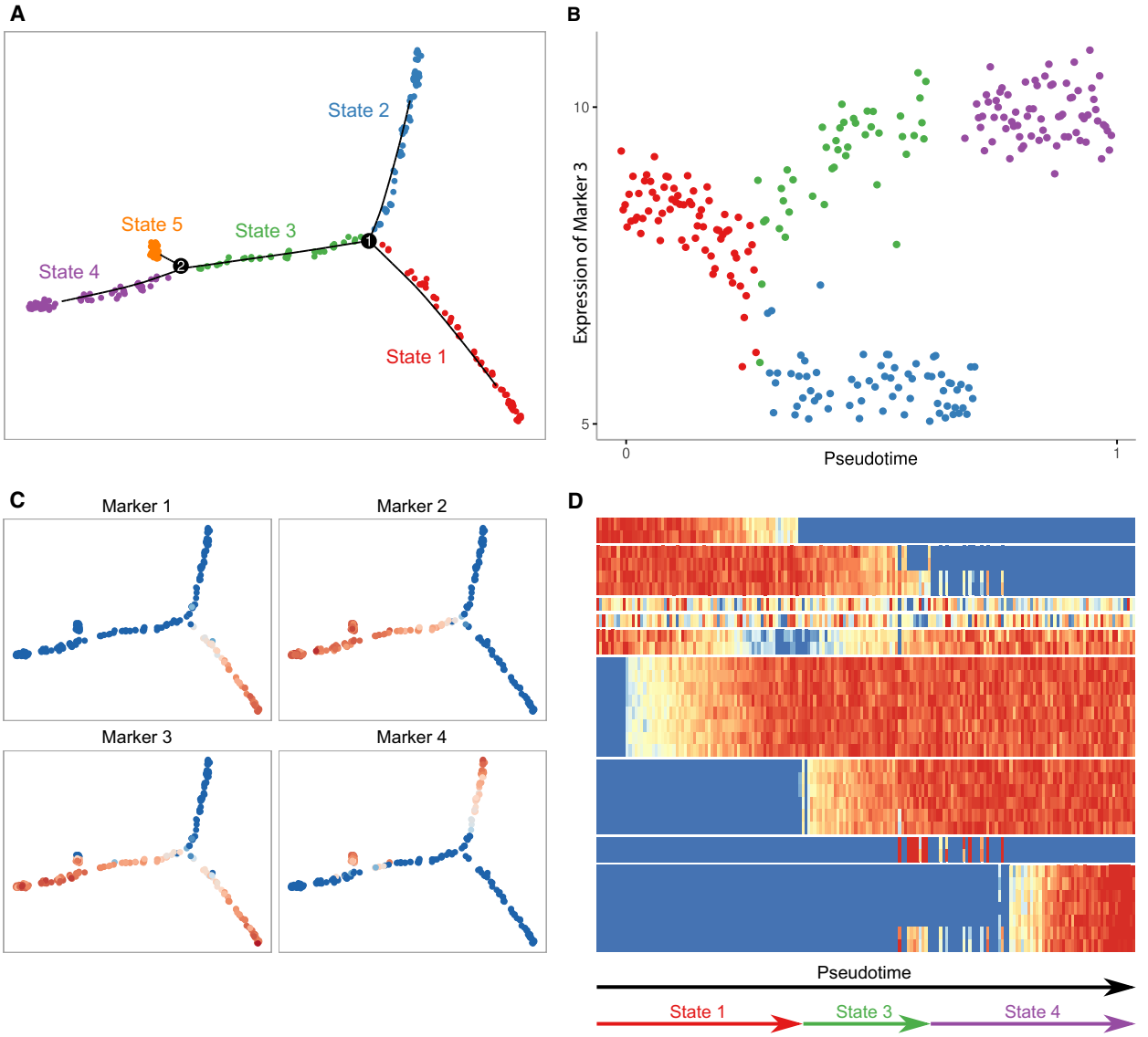

Fig. 4. There are several approaches to visualising trajectory models inferred by TI methods. (A) The most common visualisation is a dimensionality reduction where similar cells are placed close together. The cells are typically coloured based on prior knowledge (e.g. cell type) or computationally inferred clustering, and are overlaid by the trajectory inferred by the TI method. (B) A scatter plot can be used to demonstrate a response in gene expression over pseudotime. (C) Colouring of the cells in the dimensionality reduction plot can also be used to compare the gene expression profiles. (D) In order to obtain an overview of the dynamics of a large number of genes, these genes can be grouped together into modules, and one path along the trajectory can be visualised in the form of a heatmap.

the cell state- and branching point delineation, by visualising the expression value of a marker over time (Fig. 4B), comparing the gene expression values in cells within the reduced dimensions (Fig. 4C), or grouping genes together in pseudotemporally coregulated modules (Fig. 4D). Cannoodt *et al.* [95] provide an overview of several commonly used TI methods, organising them by the different components they are based on.

Trajectory inference was first explored on mass cytometry in order to reconstruct the differentiation of hematopoietic stem cells into naive B cells [96]. Since then, TI methods have been used increasingly to reconstruct cell developmental trajectories. There are several strategies TI methods use to tackle this complexity, and the choice of which method is most appropriate will thereby depend on the characteristics of the given dataset [97]. Pioneering TI methods were often specialised in producing a fixed trajectory type (e.g. linear [96,98], bifurcating [70,99] or cyclical [100]). Some methods require specific input [101], while others

are capable of inferring the trajectory structure in an unbiased way [72,102]. A recent comparative review [97] assessed the performance of more than thirty TI methods on both synthetic and real scRNA-Seq datasets, providing useful practical guidelines to choose the most appropriate methods. Notably, no method consistently outperformed the others on all datasets. Rather, various sets of methods were better suited to specific trajectories in the datasets, with some methods better identifying linear trajectories, and others efficiently identifying cycles. A good practice would therefore be to identify a set of TI methods to apply to the data based on the expected structure, and comparing the results of at least 2–3 methods to confirm the biological findings.

Differential analysis

Cytometry-based approaches

In order to identify cell populations which differ between different experimental conditions (e.g between samples of patients with different clinical outcomes), cytometry data can first be clustered, and these clusters can be compared between the conditions. In FLOWSOM [73], the user can provide a fold-change threshold, to colour clusters which differ between the conditions. The Citrus [86] and COMPASS [103] algorithms both perform model selection to identify the clusters which are best associated with a certain condition. A similar method was implemented, which groups cells into hyperspheres instead of clusters (Cydar [59]). Convolutional neural networks have also been used to identify subpopulations of cells which differ the most between two conditions (CellCNN [104]). However, none of these methods directly cope with complex experiments and may therefore be sensitive to batch effects, which might be misinterpreted as the main difference between the conditions. One solution is to first remove possible batch effects in a preprocessing step before performing differential analysis. A CYTOF workflow [105] has been proposed, which first applies clustering and then uses Gaussian linear mixture models to perform differential analysis while accounting for possible batch effect, paired experiments and other sources of technical variance in the data.

Sequencing-based approaches

The technical biases which have to be dealt with are even larger in single-cell and bulk RNA-Seq data, as many genes are lowly expressed and noisy. Several methods were proposed to specifically tackle differential expression (DE) of genes in scRNA-Seq data (SCDE [106], MAST [107], scDD [108]). These methods use mixture models or Bayesian modelling frameworks to identify both the technical effects between samples (mainly caused by the gene detection rate) and the variance which is related to the condition being tested. Another method, CENSUS [72], normalises the single-cell gene expression into relative transcript counts (accounting for technical variability between cells) in time series studies specifically, allowing for the identification of genes whose expression varies along time. These single-cell specific DE methods aim to free themselves from the idea that gene expression is unimodal across cells. Indeed, as many cells often show unmeasured genes, either due to biological or technical effects, these methods model gene expression through more elaborate distributions.

However, a recent study [109], which compared 36 differential gene expression approaches, concluded that methods that were largely used for the DE analysis of bulk RNA datasets (such as DESEQ2 [110], edger [111], VOOM [112]), were in fact not performing worse than single-cell specific DE methods on scRNA-Seq datasets. Single-cell specific DE approaches also required more computational time, although they scaled well with increasing cell numbers. This comparative study highlighted the fact that an important trend that generally improved a DE analysis results was accurate gene filtering, which reduces noise in lowly expressed genes, leading to less false positive genes being identified as differentially expressed.

Advanced computational approaches

Network inference

Single-cell transcriptomics provide a rich source of data, by quantifying the expression profiles of thousands of cells. The intercellular heterogeneity which naturally results from biological stochasticity [113] allows inferring mechanisms of gene regulation involving transcription factors and their target genes. More complex, nonlinear interactions between genes can be studied at the single-cell level, as was shown with the PIDC [114] algorithm, which was able to infer regulatory networks involved in developmental processes from sc-qPCR datasets. However, inferring one global regulatory network from thousands of cells might not always prove accurate. Different subpopulations of cells in the data might be undergoing different regulatory processes, which is why some methods were implemented specifically to compute differential regulatory networks. These methods derive one regulatory

network for each cell subtype (CSRF [115], Pólya tree models [116]).

In order to improve the inference of gene regulatory networks, external sources of information can be provided. As was discussed in the section 'Approaches modelling gradual transitions', cells can be ordered along developmental trajectories. Some network inference methods can include the information from these inferred trajectories to reconstruct dynamic regulatory networks (AR1MA1 [117], SCODE [118]). Another source of external information could come from perturbational studies, in which genes are knocked out and the consequences on the transcriptome can be observed [21]. New tools will be needed to optimally use this type of data in order to infer regulatory networks.

Single-cell transcriptomics data represent a rich source of information to infer interactions which occur between genes and transcription factors. However, new studies are highlighting the need to not only focus on a single-cell's transcripts, but also the methylation state of the DNA, the chromatin state and other epigenomic data that might enrich our knowledge of the gene regulation dynamics [119,120].

Single-cell multi-omics data integration

Single-cell transcriptomics, proteomics, genomics and epigenomics have provided a level of understanding of the cellular heterogeneity that could not be reached with bulk studies. However, the models which are inferred from single technologies are by definition incomplete. Indeed, the relationships between the genome, the amount of transcripts and proteins in a single cell are not always straightforward. Transcriptional regulatory mechanisms such as methylation may for instance alter the correlation between the gene copy number and the associated number of transcripts. Moreover, post-transcriptional mechanisms regulating protein translation and stability may also influence the relation between the number of transcripts and proteins in a cell. In order to fully understand and to start modelling the mechanisms involved in single cells, it will therefore be essential to integrate complementary types of data from the same single cells [26].

New experimental approaches have already been able to achieve a simultaneous and multiparameter measurement by combining methods. The study of the genome together with the transcriptome [24,121] for instance has confirmed the existence of a strong correlation between genes with high copy numbers and the number of mRNA transcripts. The joint analysis of the methylome together with the transcriptome [25] also corroborated the negative relation between the methylation of a gene and its transcription. More surprisingly, the measurement of both transcripts and proteins [122,123] in single cells has highlighted the fact that the amount of these two entities was poorly correlated. This could be due to the fact that transcription occurs in bursts, resulting in high discrepancies between the numbers of transcripts, whereas protein levels have been shown to be more stable for particular genes [124].

The experimental procedures cited above led to low-throughput datasets, typically containing 100 cells at most, and could therefore be analysed by regular correlation studies to assess the links between different omics entities. The recently published CITE-seq [22] and REAP-seq [23] methods have allowed the simultaneous measurement of the transcriptome as well as 100 proteins in thousands of cells, and have the potential to measure thousands of proteins in single cells, as these proteins are tagged with synthetic oligonucleotides. Some studies have also achieved a broader characterisation of single cells by combining proteomics- and imaging-based approaches [125,126]. As new experimental procedures keep providing larger and larger datasets, and new tools allow getting more insight into the mechanisms of regulations at the single-cell level [127,128], there is a great need for multi-omics integrative computational tools. These tools should have the ability to combine the information coming from complementary sources to infer complex global models.

Conclusions and future perspectives

Various high-throughput approaches currently allow studying cell populations into unprecedented depth. The rapid development of novel technologies or hybridisations between them is generating large and complex datasets that require designing novel computational approaches for preprocessing, visualising and extracting novel patterns from them. As novel technologies arise, the development of computational tools and the adequate benchmarking between them is lagging behind. Indeed, many computational approaches to study single-cell data are continuously being published, but the number of benchmark studies that objectively compare these methods is under-represented. Nevertheless, such benchmarks are essential to extract useful guidelines for biologists who want to use these tools, pinpoint limitations of current approaches and highlight novel directions for future tool development.

While current methods mainly focus on cells in suspension, novel advances that include the spatial context will stimulate novel classes of computational tools that will enable modelling cellular interactions and cell dynamics into much greater depth. Such techniques

will allow going from cells in isolation to tissues and organs, offering new perspectives for multiscale modelling. On the other hand, single-cell multi-omics approaches are providing complementary information that can relate epigenetic, transcriptional and translational information, paving the way for single-cell multi-omics and multi-source data integration.

All of these advances strengthen the idea that the life sciences are becoming even more data-driven sciences. To be able to analyse and correctly interpret the results of computational pipelines, young researchers thus should be trained adequately in properly using and understanding the principles of these novel computational approaches.

Acknowledgements

We thank Sofie Van Gassen, Robrecht Cannoodt, Niels Vandamme and Daniel Peralta for critical comments and valuable input. HT is funded by a BOF-IOP grant from Ghent University; YS is an ISAC Marylou Ingram scholar.

Conflict of interest

The authors declare no competing interests.

References

1 Liu Z, Lavis LD & Betzig E (2015) Imaging live-cell dynamics and structure at the single-molecule level. *Mol Cell* **58**, 644.
2 Abraham V, Taylor D & Haskins J (2004) High content screening applied to large-scale cell biology. *Trends Biotechnol* **22**, 15–22.
3 Goodman A & Carpenter AE (2016) High-throughput, automated image processing for large-scale fluorescence microscopy experiments. *Microsc Microanal* **22**, 538–539.
4 Kamentsky L, Jones TR, Fraser A, Bray MA, Logan DJ, Madden KL, Ljosa V, Rueden C, Eliceiri KW & Carpenter AE (2011) Improved structure, function and compatibility for Cell Profiler: modular high-throughput image analysis software. *Bioinformatics* **27**, 1179–1180.
5 Fulwyler MJ (1965) Electronic separation of biological cells by volume. *Science (New York, NY)* **150**, 910–911.
6 Robinson JP & Roederer M (2015) Flow cytometry strikes gold. *Science* **350**, 739–740.
7 Perfetto SP, Chattopadhyay PK & Roederer M (2004) Innovation: Seventeen-colour flow cytometry: unravelling the immune system. *Nat Rev Immunol* **4**, 648–655.
8 Nolan JP, Condello D, Nolan JP & Condello D (2013). Spectral flow cytometry. In Current Protocols in Cytometry, p. 1.27.1–1.27.13. John Wiley & Sons, Inc., Hoboken, NJ.
9 McGrath KE, Bushnell TP & Palis J (2008) Multispectral imaging of hematopoietic cells: where flow meets morphology. *J Immunol Methods* **336**, 91–97.
10 Goddard G, Martin JC, Graves SW & Kaduchak G (2006) Ultrasonic particle-concentration for sheathless focusing of particles for analysis in a flow cytometer. *Cytometry Part A* **69A**, 66–74.
11 Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, Pavlov S, Vorobiev S, Dick JE & Tanner SD (2009) Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal Chem* **81**, 6813–6822.
12 Giesen C, Wang HA, Schapiro D, Zivanovic N, Jacobs A, Hattendorf B, Schüffler PJ, Grolimund D, Buhmann JM, Brandt S *et al.* (2014) Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods* **11**, 417–422.
13 Saeys Y, Van Gassen S & Lambrecht BN (2016) Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat Rev Immunol* **16**, 449–462.
14 Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G & Sandberg R (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096–1098.
15 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.
16 Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049.
17 Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, Fortune S, Love JC & Shalek AK (2017) Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods* **14**, 395–398.
18 Rosenberg AB, Roco C, Muscat RA, Kuchina A, Mukherjee S, Chen W, Peeler DJ, Yao Z, Tasic B, Sellers DL *et al.* (2017) Scaling single cell transcriptomics through split pool barcoding. *bioRxiv* [preprint].
19 Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M *et al.* (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science (New York, NY)* **353**, 78–82.
20 Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T,

Raychowdhury R *et al*. (2016) Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 e17.
21 Jaitin DA, Weiner A, Yofe I, Lara-Astiaso D, Keren-Shaul H, David E, Meir Salame T, Tanay A, van Oudenaarden A & Amit I (2016) Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-Seq. *Cell* **167**, 1883–1896.e15.
22 Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R & Smibert P (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865–868.
23 Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, Moore R, McClanahan TK, Sadekova S & Klappenbach JA (2017) Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol* **35**, 936–939.
24 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.
25 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.
26 Macaulay IC, Ponting CP & Voet T (2017) Single-cell multiomics: multiple measurements from single cells. *TIG* **33**, 155–168.
27 Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, Guertin DA, Chang J, Lindquist RA, Moffat J *et al*. (2006) Cell Profiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol* **7**, R100.
28 Peng T, Thorn K, Schroeder T, Wang L, Theis FJ, Marr C & Navab N (2017) A BaSiC tool for background and shading correction of optical microscopy images. *Nat Commun* **8**, 14836.
29 Smith K, Li Y, Piccinini F, Csucs G, Balazs C, Bevilacqua A & Horvath P (2015) CIDRE: an illumination-correction method for optical microscopy. *Nat Methods* **12**, 404–406.
30 Wählby C (2003) Algorithms for applied digital image cytometry. Acta Universitatis Upsaliensis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 896, 75 pp., Uppsala. ISBN 91-554-5759-2.
31 Malpica N, de Solórzano CO, Vaquero JJ, Santos A, Vallcorba I, García-Sagredo JM & del Pozo F (1998) Applyingwatershed algorithms to the segmentation of clustered nuclei. *Cytometry* **28**, 289–297.
32 Wahlby C, Sintorn IM, Erlandsson F, Borgefors G & Bengtsson E (2004) Combining intensity, edge and shape information for 2D and 3D segmentation of cell nuclei in tissue sections. *J Microsc* **215**, 67–76.
33 Ortiz de Solórzano C, García Rodriguez E, Jones A, Pinkel D, Gray JW, Sudar D & Lockett SJ. (1999) Segmentation of confocal microscope images of cell nuclei in thick tissue sections. *J Microsc* **193**, 212–26.
34 Meyer F & Beucher S (1990) Morphological segmentation. *J Vis Commun Image Represent* **1**, 21–46.
35 Leipold MD (2015) Another step on the path to mass cytometry standardization. *Cytometry Part A* **87**, 380–382.
36 Takahashi C, Au-Yeung A, Fuh F, Ramirez-Montagut T, Bolen C, Mathews W & O'Gorman WE (2017) Mass cytometry panel optimization through the designed distribution of signal interference. *Cytometry Part A* **91**, 39–47.
37 Fletez-Brant K, Špidlen J, Brinkman RR, Roederer M & Chattopadhyay PK (2016) flowClean: automated identification and removal of fluorescence anomalies in flow cytometry data. *Cytometry Part A* **89**, 461–471.
38 Bashashati A & Brinkman RR (2009) A survey of flow cytometry data analysis methods. *Adv Bioinform* **2009**, 584603.
39 Monaco G, Chen H, Poidinger M, Chen J, deMagalhães JP & Larbi A (2016) flowAI: automatic and interactive anomaly discerning tools for flow cytometry data. *Bioinformatics* **32**, 2473–2480.
40 Hahne F, Khodabakhshi AH, Bashashati A, Wong CJ, Gascoyne RD, Weng AP, Seyfert-Margolis V, Bourcier K, Asare A, Lumley T *et al*. (2010) Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A* **77**, 121–131.
41 Finak G, Frelinger J, Jiang W, Newell EW, Ramey J, Davis MM, Kalams SA, De Rosa SC & Gottardo R (2014) OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS Comput Biol* **10**, e1003806.
42 Malek M, Taghiyar MJ, Chong L, Finak G, Gottardo R & Brinkman RR (2015) flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics* **31**, 606–607.
43 Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I & Enard W (2017) Comparative analysis of single-cell RNA sequencing methods. *Mol Cell* **65**, 631–643.
44 Stegle O, Teichmann SA & Marioni JC (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133–145.
45 Poirion OB, Zhu X, Ching T & Garmire L (2016) Single-cell transcriptomics bioinformatics and computational challenges. *Front Genet* **7**, 163.

46 Bacher R & Kendziorski C (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol* **17**, 63.
47 McCarthy DJ, Campbell KR, Lun ATL & Wills QF (2016) scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R. *bioRxiv* [preprint].
48 Butler A, Hoffman P, Smibert P, Papalexi E & Satija R (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420.
49 Haghverdi L, Lun ATL, Morgan MD & Marioni JC (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421–427.
50 Lun ATL, McCarthy DJ & Marioni JC (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. *F1000Research* **5**, 2122.
51 Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann SA, Stegle O, Marioni JC & Buettner F (2015) Computational assignment of cellcycle stage from single-cell transcriptome data. *Methods* **85**, 54–61.
52 Buettner F, Pratanwanich N, McCarthy DJ, Marioni JC & Stegle O (2017) f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol* **18**, 212.
53 Mortazavi A, Williams BA, McCue K, Schaeffer L & Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628.
54 Wagner GP, Kin K & Lynch VJ (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* **131**, 281–285.
55 Bullard JH, Purdom E, Hansen KD & Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94.
56 Vallejos CA, Risso D, Scialdone A, Dudoit S & Marioni JC (2017) Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods* **14**, 565–571.
57 Pierson E & Yau C (2015) ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol* **16**, 241.
58 Risso D, Perraudeau F, Gribkova S, Dudoit S & Vert JP (2017) ZINB-WaVE: a general and flexible method for signal extraction from single-cell RNA-seq data. *bioRxiv* [preprint].
59 Lun ATL, Richard AC & Marioni JC (2017) Testing for differential abundance in mass cytometry data. *Nat Methods* **14**, 707–709.
60 Vallejos CA, Marioni JC & Richardson S (2015) BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol* **11**, e1004333.
61 Ding B, Zheng L, Zhu Y, Li N, Jia H, Ai R, Wildberg A & Wang W (2015) Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics* **31**, 2225–2227.
62 Katayama S, Töhönen V, Linnarsson S & Kere J (2013) SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics* **29**, 2943–2945.
63 Reid LH (2005) Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genom* **6**, 150.
64 Baran-Gale J, Chandra T & Kirschner K (2017) Experimental design for single-cell RNA sequencing. *Brief Funct Genomics* **17**, 233–239.
65 Tung PY, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK & Gilad Y (2017) Batch effects and the effective design of single-cell gene expression studies. *Sci Rep* **7**, 39921. https://doi.org/10.1038/srep39921
66 Lun AT, Calero-Nieto FJ, Haim-Vilmovsky L, Gottgens B & Marioni JC (2017) Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *bioRxiv* [preprint].
67 Bacher R, Chu LF, Leng N, Gasch AP, Thomson JA, Stewart RM, Newton M & Kendziorski C (2017) SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods* **14**, 584–586.
68 van der Maaten L & Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* **9**, 2579–2605.
69 Kruskal JB (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**, 1–27.
70 Haghverdi L, Buettner F & Theis FJ (2015) Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998.
71 Weinreb C, Wolock S & Klein A (2017) SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *bioRxiv* [preprint].
72 Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA & Trapnell C (2017) Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979–982.
73 Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T & Saeys Y (2015) FlowSOM: using selforganizing maps for visualization and interpretation of cytometry data. *Cytometry Part A* **87**, 636–645.
74 Zunder ER, Lujan E, Goltsev Y, Wernig M & Nolan GP (2015) A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. *Cell Stem Cell* **16**, 323–337.
75 Weber LM & Robinson MD (2016) Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A* **89**, 1084–1096.

76 Spitzer MH, Gherardini PF, Fragiadakis GK, Bhattacharya N, Yuan RT, Hotson AN, Finck R, Carmi Y, Zunder ER, Fantl WJ *et al*. (2015) An interactive reference framework for modeling a dynamic immune system. *Science* **349**, 1259425.
77 Levine JH, Simonds EF, Bendall SC, Davis KL, EaD A, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER *et al*. (2015) Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197.
78 Klein A, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA & Kirschner MW (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201.
79 Satija R, Farrell JA, Gennert D, Schier AF & Regev A (2015) Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495–502.
80 Campbell KR & Yau C (2016) Order under uncertainty: robust differential expression analysis using probabilistic models for pseudotime inference. *PLoS Comput Biol* **12**, e1005212.
81 Anchang B, Hart TDP, Bendall SC, Qiu P, Bjornson Z, Linderman M, Nolan GP & Plevritis SK (2016) Visualization and cellular hierarchy inference of single-cell data using SPADE. *Nat Protoc* **11**, 1264–1279.
82 Shekhar K, Brodin P, Davis MM & Chakraborty AK (2014) Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE). *Proc Natl Acad Sci USA* **111**, 202–207.
83 Aghaeepour N, Finak G, Hoos H, Mosmann TR, Brinkman R, Gottardo R & Scheuermann RH (2013) Critical assessment of automated flow cytometry data analysis techniques. *Nat Methods* **10**, 228–238.
84 Newell EW & Cheng Y (2016) Mass cytometry: blessed with the curse of dimensionality. *Nat Immunol* **17**, 890–895.
85 Platon L, Pejoski D, Gautreau G, Targat B, Le Grand R & Beignon AS (2018) A computational approach for phenotypic comparisons of cell populations in high-dimensional cytometry data. *Methods* **132**, 66–75.
86 Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ & Nolan GP (2014) Automated identification of stratifying signatures in cellular subpopulations. *Proc Natl Acad Sci USA* **111**, E2770–E2777.
87 Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR *et al*. (2017) SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **14**, 483–486.
88 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 (New York, NY)* **347**, 1138–1142.
89 Sun Z, Wang T, Deng K, Wang XF, Lafyatis R, Ding Y, Hu M & Chen W (2018) DIMM-SC: a Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics* **34**, 139–146.
90 Lin P, Troup M & Ho JWK (2017) CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol* **18**, 59.
91 Wang B, Zhu J, Pierson E, Ramazzotti D & Batzoglou S (2017) Visualization and analysis of single-cell RNA-seq data by kernelbased similarity learning. *Nat Methods* **14**, 414–416.
92 Xu C & Su Z (2015) Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31**, 1974–1980.
93 Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine JC, Geur P & Aerts J (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083–1086.
94 Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, Kaper F, Fan J-B, Zhang K, Chun J *et al*. (2016) Characterizing transcriptional heterogeneity through pathway and gene set over dispersion analysis. *Nat Methods* **13**, 241–244.
95 Cannoodt R, Saelens W & Saeys Y (2016) Computational methods for trajectory inference from single-cell transcriptomics. *Eur J Immunol* **46**, 2496–2506.
96 Bendall SC, Davis KL, Amir EAD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP & Pe'er D (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725.
97 Saelens W, Cannoodt R, Todorov H & Saeys Y (2018) A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv* [preprint].
98 Cannoodt R, Saelens W, Sichien D, Tavernier S, Janssens S, Guilliams M, Lambrecht BN, De PK & Saeys Y (2016) SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development. *bioRxiv* [preprint].
99 Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N & Pe'er D (2016) Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* **34**, 1–14.
100 Liu Z, Lou H, Xie K, Wang H, Chen N, Aparicio OM, Zhang MQ, Jiang R & Chen T (2017) Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat Commun* **8**, 22.
101 Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS & Rinn JL (2014) The dynamics and regulators of cell

fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386.
102 Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E & Dudoit S (2017) Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *bioRxiv* [preprint].
103 Lin L, Finak G, Ushey K, Seshadri C, Hawn TR, Frahm N, Scriba TJ, Mahomed H, Hanekom W, Bart P-A *et al.* (2015) COMPASS identifies T-cell subsets correlated with clinical outcomes. *Nat Biotechnol* **33**, 610–616.
104 Arvaniti E & Claassen M (2017) Sensitive detection of rare disease-Associated cell subsets via representation learning. *Nat Commun* **8**, 1–10.
105 Nowicka M, Krieg C, Weber LM, Hartmann FJ, Guglietta S, Becher B, Levesque MP & Robinson MD (2017) CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research* **6**, 748.
106 Kharchenko PV, Silberstein L & Scadden DT (2014) Bayesian approach to single-cell differential expression analysis. *Nat Methods* **11**, 740–742.
107 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.
108 Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R & Kendziorski C (2016) A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol* **17**, 222.
109 Soneson C & Robinson MD (2018) Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods* **15**, 255–261.
110 Love MI, Huber W & Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550.
111 Robinson MD, McCarthy DJ & Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140.
112 Law CW, Chen Y, Shi W & Smyth GK (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29.
113 Padovan-Merhar O & Raj A (2013) Using variability in gene expression as a tool for studying gene regulation. *WIREs Syst Biol Med* **5**, 751–759.
114 Chan TE, Stumpf MPH & Babtie AC (2017) Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems* **5**, 251–267.e3.
115 Xu R, Nettleton D & Nordman DJ (2016) Case-specific random forests. *J Comput Graph Stat* **25**, 49–65.
116 Filippi S & Holmes CC (2017) A Bayesian nonparametric approach to testing for dependence between random variables. *Bayesian Anal* **12**, 919–938.
117 Castillo MS, Blanco D, Luna IMT, Carrion MC & Huang Y (2018) A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics* **34**, 964–970.
118 Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, Hayashi T & Nikaido I (2017) SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33**, 2314–2321.
119 Fiers MWEJ, Minnoye L, Aibar S, Bravo González-Blas C, Kalender Atak Z & Aerts S (2018) Mapping gene regulatory networks from single-cell omics data. *Brief Funct Genomics* **17**, 246–254.
120 Äijö T & Bonneau R (2017) Biophysically motivated regulatory network inference: progress and prospects. *Hum Hered* **81**, 62–77.
121 Dey SS, Kester L, Spanjaard B, Bienko M & Van Oudenaarden A (2015) Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol* **33**, 285–289.
122 Darmanis S, Gallant CJ, Marinescu VD, Niklasson M, Segerman A, Flamourakis G, Fredriksson S, Assarsson E, Lundberg M, Nelander S *et al.* (2016) Simultaneous multiplexed measurement of RNA and proteins in single cells. *Cell Rep* **14**, 380–389.
123 Albayrak C, Jordi CA, Zechner C, Lin J, Bichsel CA, Khammash M & Tay S (2016) Digital quantification of proteins and mRNA in single mammalian cells. *Mol Cell* **61**, 914–924.
124 Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W & Selbach M (2011) Global quantification of mammalian gene expression control. *Nature* **473**, 337–342.
125 Soh KT, Tario JD, Colligan S, Maguire O, Pan D, Minderman H & Wallace PK (2016) Simultaneous, single-cell measurement of messenger RNA, cell surface proteins, and intracellular proteins. *Curr Protoc Cytom* **75**, 7.45.1–7.45.33.
126 Kochan J, Wawro M & Kasza A (2015) Simultaneous detection of mRNA and protein in single cells using immunofluorescence combined single-molecule RNA FISH. *Biotechniques* **59**, 209–212, 214, 216.
127 Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY & Greenleaf WJ (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490.
128 Jin W, Tang Q, Wan M, Cui K, Zhang Y, Ren G, Ni B, Sklar J, Przytycka TM, Childs R *et al.* (2015) Genome-wide detection of DNase i hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**, 142–146.