1 Introduction

Cells are one of the fundamental units of life. They show an immense complexity and diversity. Their identity and function is determined by environmental stimuli, the physical environment, the cell cycle and neighbouring cells (Wagner et al., 2016). Only recently has it been possible to investigate the transcriptome of a single cell. Single-cell RNA sequencing (scRNA-seq) was first published by Tang et al. (2009). This method addresses new biological issues, such as the identification of rare cell populations, and allows us to measure the frequency of cell types in tissues, characterise differences in similar cell types and investigate the heterogeneity of cell states or cell lineages (Andrews and Hemberg, 2017).

A typical scRNA-seq workflow consists of the isolation of single cells, the extraction of RNA, cDNA library preparation, and the amplification and sequencing of the libraries. A wide variety of scRNA-seq protocols exists, differing in throughput, full transcript or 3' sequencing, costs and automatisation. Small-scale protocols are standard PCR plate-based methods or methods in which cell isolation and library preparation are combined into one protocol. A typical small-scale method is the PCR plate-based SMART-seq2 (Picelli et al., 2013). Libraries are full-transcript sequenced using a standard Illumina sequencing approach. Typically, hundreds of cells are processed and ERCC is used for normalisation. On the other side of the spectrum are droplet-based methods such as Drop-seq and 10xChromium, which allow the processing of thousands of cells.

The differences between scRNA-seq and bulk experiments are the lower sequencing depth (100,000–5 million reads per cell), higher variability and more outliers. scRNA-seq data suffers from technical noise, batch effects and low capture efficiency. Batch effects occur when different biological conditions are processed in different batches, making the deconvolution of technical noise and biological effect impossible. This should be avoided by an appropriate experimental design that allows for the statistical deconvolution between unwanted and wanted variation. In scRNA-seq the single experimental unit is the cell, making it not always possible to use this approach. Different cells in an experiment may need different sample processing, or their biological differences may affect the downstream analyses.

The starting amounts of the library preparation can be as low as ten picogrammes of total RNA (Picelli et al., 2013). Two main issues are arising due to the low starting amount are overamplification and low capture efficiency. Low and moderate expressed genes are not captured during the reverse transcription, which leads to dropouts of genes and a zero-inflated gene expression. scRNA-seq data has an excess of zero counts, which can be split into systematic, semi-systematic and stochastic zeros (Lun et al., 2016a). Systematic zeros and semi-systematic zeros come from genes that are silent in a subpopulation or across all cells, respectively. Stochastic zeros are zero counts that have been obtained due to sampling. They affect genes with a count distribution near zero and have to be dealt with the normalisation steps.

To deal with technical noise Unique Molecular Identifiers (UMI) or spike-in from the External RNA Control Consortium (ERCC) are used. UMI are short random barcodes attached to the single-stranded cDNA in the reverse transcriptase process. By counting the unique UMI reads aligned to the genome an estimated tag count is obtained. Spike-in are added before amplification. Under the assumption that the amplification of the endogenous and exogenous RNA is similar, they can be used for library size normalisation and to remove technical noise.

In general, scRNA-seq experiments consist of high-dimensional data. High-dimensional data suffers from the curse of dimensionality (Wagner et al., 2016). Distances in high-dimensional data become unstable and subpopulations cannot be separated (Andrews and Hemberg, 2017). Additionally, computational requirements are high. Reduction of the dimension is made by two approaches. Using linear or non-linear projections of the data from the original high-dimensional to a lower-dimensional space, or by feature selection, in which uninformative genes are removed.

The detection of new cell types is one of the most common aims for scRNA-seq data. Lately, a broad spectrum of clustering methods have been specifically developed for the clustering of single cells. The aim of this study is to evaluate clustering methods for scRNA-seq data.

2 Methods

2.1 Dimension reduction

All methods require a dimension reduction step before clustering. Commonly used methods are either Principal Component Analysis (PCA) or t-distributed Stochastic Neighbour Embedding (tSNE). PCA finds a rotation of the original data in which the newly obtained first coordinates have the highest possible variance, the second coordinates the second-greatest variance etc. Practically, PCA is computed by spectral decomposition of the correlation matrix. Dimension reduction is achieved either by selecting only the first few PCs whose eigenvalues are less than average, or by determineing the number graphically by plotting the eigenvalues. PCA is deterministic and relatively fast but restricted to linear spaces.

In contrast, tSNE is a non-linear mapping (Van Der Maaten, 2013). Stochastic neighbour embedding (SNE) transforms Euclidean distances to conditional probabilities $p_{j|i}$. That is the probability of x_j is the nearest neighbour of x_i under a Gaussian centred at x_i. The low-dimensional counterpart $q_{i|j}$ is similar with a Gaussian centered at y_i and variance of $1/\sqrt(2)$. SNE minimises the divergence between $p_{j|i}$ and $q_{j|i}$ using the Kullback-Leiber divergence. tSNE implements a Student's t-distribution for the low dimensional space and a symmetric version of the cost function to simplify optimisation and to overcome the crowding problem. In tSNE, the cost function uses the joint probabilities p_{ij} and q_{ij} instead of conditional probabilities. To deal with large datasets, the Barnes-Hut implementation uses random walks on the nearest neighbour network with a PCA step to reduce the dimensionality of the high-dimensional data. tSNE is stochastic, depends on a perplexity parameter and distances between clusters are not preserved.

2.2 Clustering methods

Identifying unknown cell populations is one of the main uses of scRNA-seq data (Andrews and Hemberg, 2017). Eleven clustering methods have been evaluated for this study. These methods and algorithms can be roughly classified into three groups: K-means, graphclustering and hierarchical-based clustering. SC3, SIMLR, Linnorm and RaceID use k-means in different fashions, while pcaReduce and CIDR are based on hierarchical clustering. The graph-based methods are SNN-Cliq and Seurat. An overview of the methods is given in Table ...

K-means K-means clustering finds a predefined number of centres k and cell assignments, such that their within-group sum of squares is minimised (Hartigan and Wong, 1979). k cluster centres are randomly assigned. Each data point is then assigned to the nearest centre using Euclidean distances. The centres are then recomputed using the average of the data points that are assigned to each of the k centres. This procedure is iterated until the algorithm converges. The assigning of the centres is random. Also, it's not guaranteed to find the global minimum. The drawbacks of the method are that it assumes spherical cluster and the sensitivity to scaling.

pcaReduce pcaReduce uses k-means clustering to find the number of clusters in the reduced dimension given by PCA (Yau et al., 2016). The main assumption is that large classes of cells are contained in low-dimension PC representation and more refined subsets of these cells types are contained in higher-dimensional PC representations. Given a gene expression matrix, the clustering algorithm starts with a k-means clustering on the PCA projections $Y_{n \times q}$ with $q + 1$ clusters. Where n are the cells and q are the number of PCs. The number of initial clusters k is typically around 30, guaranteeing that most cell types are captured. For all pairs of clusters, the joint probabilities are computed. Two clusters are merged by selecting the pair with the highest joint probability or by sampling proportionally by the joint probabilities. The number of clusters is now decreased to $K - 1$. Next, the PC with the lowest variance is deleted and a k-means clustering with $K - 2$ centres is performed. This process is repeated until only one single cluster remains. When using pcaReduce q cluster partitions with $k - 1$ clusters are obtained.

SC3 Implemented in the SC3 method is a gene- and cell-filtering, as well as a log transformation step of the expression matrix (Kiselev et al., 2017). The filtered expression matrix is then used to compute Euclidean, Pearson and Spearman dissimilarity measures. By PCA or Laplacian graphs a lower-dimensional representation of the data is obtained. K-means clustering is then performed on the different dimensions. Next, a consensus matrix of the different clustering results is computed. The consensus matrix is a binary similarity matrix with entry one if two cells belong to the same cluster and zero otherwise. The consensus matrix is obtained by averaging the individual clustering. The last step is a hierarchical clustering step with complete linkage. The cluster is inferred by the k level of hierarchy, where k is supplied by the user. To reduce runtime SC3 changes the clustering method when supplied with more than 5'000 cells. Randomly selected cells are then used for the clustering approach described above. These subpopulations are then used to train a support vector machine to infer the remaining cells.

SNNcliq SNNcliq computes a shared nearest-neighbour-graph based on the high-dimensional data (Xu and Su, 2015). The nodes are the data points and the weighted edges are the similarities between the data points. Cells are defined as a cluster if they have a defined number of edges between them, forming a "clique".

A similarity matrix using Euclidean or other similarity measures is then computed. Using this similarity matrix, the k-nearest-neighbours (KNN) for each data point are listed. The number of nearest neighbours has to be supplied by the user. Edges between data points are assigned if they share at least one KNN. The weights of the edges are defined by a function of the number of nearest neighbours and their respective ranks. Identification of clusters is made by finding quasi-cliques associated with each node and merging them to unique clusters by the use of a greedy algorithm. A node induces a subgraph, which consists of all neighbour nodes and edges. For each node, a local degree is computed and a node is removed from the subgraph if its degree is lower thana given threshold which is proportional to the size of the clique. The threshold is supplied by the user and is typically set to 0.7. Next, the degrees between the nodes are recomputed and the process is repeated until no more nodes can be removed. A subgraph is assigned to a quasi-clique if it contains more than three nodes. To reduce redundancy, quasi-cliques that are completely included in other cliques are removed as well. Clusters are then identified by merging the quasi-cliques. For each pair, an overlapping rate is computed. If it exceeds a predefined threshold m the sub graphs are merged. Merging in different orders leads to different results, so pairs with larger sizes are prioritized.

SIMLR Most clustering methods rely on standard similarity metrics like Euclidean distances (Wang et al., 2017). SIMLR uses a weighted function of multiple kernels to compute a distance matrix. The assumptions is that the matrix has a block-diagonal structure, where the blocks represent the clusters c. The kernels are Gaussian kernels with a range of hyperparameters defining the variance of each kernel. The similarities are then used for data visualisation with tSNE or clustering using k-means on the latent space representations of the similarities.

CIDR Clustering through Imputation and Dimensionality Reduction (CIDR) takes the high dropout rate in scRNA-seq data into account (Lin et al., 2017). The method splits the squared Euclidean distance into three terms. These consist of one term in which both genes k for the cell pairs i an j are non-zero, a second term in which one gene is zero and a third where both are zero. The authors state that only the cases where one gene is zero have a strong influence on the distances and the subsequent dimension reduction and clustering. To reduce the dropout-induced zero inflation, the method imputes the third term by its expected value given the distribution of the dropouts. The algorithm works basically in five steps: (i) Find features that are dropout candidates. That is genes that show an expression level below a threshold T. (ii) Find the empirical dropout probability $\hat{P}(u)$ using the whole data set. (iii) Calculate the dissimilarity using Euclidean distances together with pairwise a imputation process. Features that fall below the threshold T are imputed using a weighting function. The weighting is based on the probability of being a dropout. (iv) Perform dimension reduction using PCA on the imputed distance matrix. (v) Perform hierarchical clustering using the first few PCs. The number of PCs can be determined by several methods. Here, we use an implemented variation of the scree method.

Linnorm Linnorm is a normalisation and transformation method for count data (Yip et al., 2017). The main assumption is that a homogeneously expressed gene-set exists. Using this gene subset, and by ignoring zero counts, the normalisation and transformation parameters are calculated. After normalisation, the expression values should show both homoscedasticity and normality. Linnorm includes functions for subpopulation analysis by t-SNE or PCA dimension reduction and subsequent k-means or hierarchical clustering.

First, the values are scaled according to library size. Low count genes and genes that show high technical noise are filtered out. By default, genes showing a non-zero expression in at least 75 % of the cells are retained. Note that this threshold is set to maintain at least three non-zero cells per gene, in order to calculate the skewness of the gene distributions. By gradually increasing this threshold only genes which show a negative correlation between the mean and the standard deviation (SD) is assured. A locally weighted scatterplot smoothing (LOWESS) curve is fitted on the mean versus SD relationship. The SD is scaled and outliers based on the SD are removed. Next, genes that show a high skewness are filtered out. The data is then transformed using a modified log transformation.

TSCAN TSCAN uses a pseudo-time algorithm for cell ordering (Ji and Ji, 2015). PCA dimension reduction on the preprocessed gene expression data is performed. Preprocessing is done by log transformation and by adding a pseudo-count. Low expressed genes are filtered out based on the zero-proportion and their covariance. Clustering is done by model-based clustering. The travelling salesman problem (TSP) is then solved by a minimum spanning tree. The user can then define the start/end point through the available biological information and compute a pseudo-time ordering score.

2.3 Datasets

The datasets Kumar, Trapnell and Koh were downloaded from the *conquer* repository `http://imlspenticton.uzh.ch:3838/conquer`. The Zheng data was downloaded from the 10xChromium repository: `https://support.10xgenomics.com/single-cell-gene-expression/datasets`. For the Kumar, Trapnell and Koh data count-scale, length-scaled Transcripts Per kilobase Million (TPM) are used. These are on a count-scale and are independent from the average length of the transcript (Soneson et al., 2015). The Zheng data consists of UMI counts.

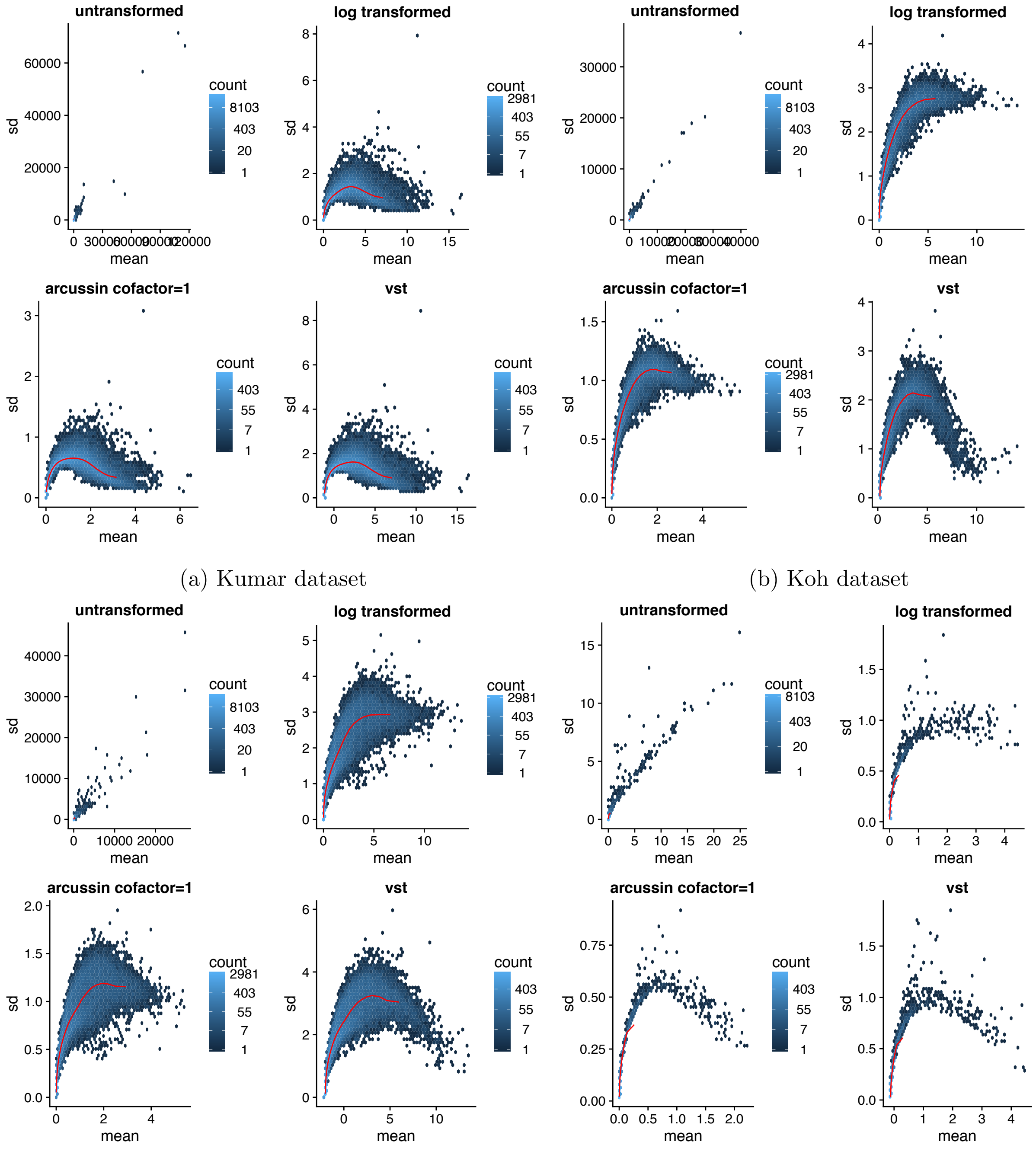
Kumar et al. (2014) The Kumar dataset consists of *Dgcr*8-knockout and V6.5 variotypes from mouse embryonic stem cells (mESCs). Cells were cultured on a serum with a Leukaemia Inhibitory Factor (LIF) or under Erk and GSK3 signalling inhibition (2Li). The authors investigated the expression of pluripotency factors and their involvement in the heterogeneity of pluripotent stem cells. Sample preparation and whole transcriptome amplification was done using a Fluidigm C1 system and following a SMARTer protocol. Sequencing was done using the Illumina system with paired-end reads. Sequencing depth was 1 million reads per cell.

Trapnell et al. (2014) Trapnell et al. (2014) used human skeletal muscle myoblast cells to investigate temporal differentiation. Cells were expanded under high–mitogen conditions. Differentiation was induced by switching to low-serum medium. Cells were captured before switching to low-serum medium (T0), after 24 h (T24) and 48h (T48). Between 49 and 77 cells were isolated at each time point and used for single mRNA-Seq library preparation by the use of a Fluidigm C1 system and following a SMARTer protocol. Libraries were sequenced with paired-end sequencing on a HiSeq 2500 (Illumina) platform. Sequencing depth was 4 million reads per library. The authors excluded libraries that contained fewer than 1 million reads.

Koh et al. (2016) H7 human embryonic stem cells (hESCs) were used to study human mesoderm development. Starting from undifferentiated stem cells, several differentiation stages, sorted by time point and further refined by fluorescence-activated cell sorting (FACS) were isolated. Finally, nine different cell lines were obtained: undifferentiated H7 hESCs (H7hESC), anterior primitive streak populations (APS), mid primitive streak populations (MPS), lateral mesoderm (D2LtM), FACS-purified DLL1+ paraxial mesoderm populations (DLL1pPXM), early somite progenitor populations (ESMT), PDGFR+ sclerotome populations (Sclrtm) and two different dermomyotome populations (D5CntrlDRmmtm). In total, ten different cell types were then sequenced on a Fluidigm C1 system and following a SMARTer protocol. Libraries were sequenced through paired-end sequencing on a HiSeq 2500 (Illumina) platform. Sequencing depth was 1 – 2 million reads per cell.

Zheng et al. (2017) FACS-purified fresh peripheral blood mononuclear cells (PBMCs) were used to assess the performance of the 10xChromium system. Sample preparation and library construction were done with a 10xChromium system. The libraries were sequenced using an Illumina system. For this study, the datasets for CD19+B, CD8+CD45RA+ naive cytotoxic, CD14+ monocytes and CD4+/CD25+ regulatory T cells were used to construct an artificial population. From each library, 200 cells were sampled before beeing merged to obtain a single expression matrix.

Simulated datasets Using the Splatter package, expression data were simulated (Oshlack et al., 2017). Parameters for the simulation were estimated from a subpopulation of the Kumar dataset. Embryonic stem cell variotypes V6.5 with signalling inhibition and LIF were used for the estimation. SimDataKumar consists of 500 cells with four subpopulations. The fractions per subpopulations were 0.1, 0.15, 0.5 and 0.25 of the total cell population. The probability that a gene is differentially expressed is 0.05, 0.1, 0.2 and 0.4 in the four different groups. Similarly, SimDataKumar2 consists of four subgroups with fractions of 0.2, 0.15, 0.4 and 0.25 of a total of 500 cells. The fractions of differentially-expressed genes were lower, with a probability of 0.01, 0.05, 0.05 and 0.08. The spike-in RNA is excluded before the parameter estimation.



(a) Kumar dataset

(b) Koh dataset

(c) Trapnell dataset

(d) Zheng dataset

Figure 1: Shown is the genewise standart deviation versus the mean for the datasets Kumar (a), Koh (b), Trapnell (c) and Zheng (d). Different transformations were considered; log, arcus sin and VST transformations.

2.4 Data transformation and normalisation

Data transformation RNA-seq data may suffer from heteroscedasticity and skewness (Zwiener et al., 2014). Genes with higher mean have on average a higher variance across cells leading to unequal variances between different genes. To handle this property different transformation were considered. Namely, a binary logarithmic transformation with a pseudo-count of one, arcus sinus transformations and a variance-stabilising transformation (VST) from the DESeq package (Huber et al., 2002). Log transformations will have an impact on extreme values. However, they will not address the problem of heteroscedasticity. Arcus sinus transformation should deal with extreme values and equalise the variances. After transformation, the mean and the variances should be independent. VST addresses the problem of extreme values and unequal variances across genes. After such transformation, the mean and the variances of the genes should be independent. For the study, a binary logarithmic transformation plus a pseudo-count of one is used. The mean-SD dependence for different transformations is shown in Figure 1.

Filtering and normalisation The quality control of the data sets follows Lun et al. (2016b). In the first step, genes that are not expressed in any cell (systematic–zeros) are removed in order to reduce the size of the expression matrix. To find potential outliers, PCA can be used on the phenotype characteristic of each cell can be used (Figure 7, 8, 9, 10; a). Cells were filtered based on the library size and the total number of genes. Cells with log10 library sizes that are more than three median absolute deviations (MADs) below the median log-library size were filtered out (Figure 2, 3, 4 and 5). The same filter was used with respect to the total number of genes per cell. For the Kumar and the Zheng dataset, ERCCs and MT counts were available. Cells with large proportions of ERCC or mitochondrial RNA are seen as low-quality cells. In the Kumar dataset, cells with an ERCC proportion above three MADs are as well removed. The same filter was used for mitochondrial RNA in the Zheng data.

The metadata for the Trapnell dataset contained information about the cell quality. In this dataset, cells that were marked as debris and any single libraries consisting of more than one cell were filtered out. After filtering, 531 cells in the Koh dataset, 246 in the Kumar dataset and 222 in the Trapnell dataset were retained. The filtering was less strict in the Koh dataset compared to the original analysis where they retained 498 cells.

Low-abundance genes influence the mean-variance trend. Here low-abundance genes are filtered by their average counts (see Figures 7, 8, 9 and 10; d). For the Kumar, Trapnell, simDataKumar and Zheng data genes with average counts less than one are removed. The Zheng data set had a shallower sequencing depth. A different filter is used, and features which are not expressed in at least two cells are excluded. To find batch effects a linear model regressing the PC values against the total features was used (Lun et al., 2016b).

Another examination of the technical variation was done using the marginal variances(Lun et al., 2016b). For that, a linear model with the expression values per gene as response variables and a chosen explanatory variable is fitted. The correlation coefficient can then be seen as the marginal explained variance for the explanatory variables.

A wide variety of normalisation methods exist based on bulk RNA methods. These methods are usually not designed for dealing with the zero-inflated nature of scRNA-seq dataLun et al. (2016a). Methods for normalisation of scRNA-seq data are based on spike-ins or RNA counts. Spike-in RNA is added before the library preparation. Any changes in the spike-in coverage are assumed to be due to technical factors. The normalisation is done by scaling the counts to level the spike-in. However, this approach is not feasible as none or only a limited number of spike-in counts were present in the datasets.

Here, normalisation through pooled cells is used, where the problem of excess zero counts is reduced by the pooling of multiple cells Lun et al. (2016a). The normalisation procedure can briefly be described as follows: (i) Different pools of cells are defined. (ii) The expression values are summed across the cell pools. (iii) The cell pool is normalised against an average of the summed expression values. (iv) This step is repeated several times to construct a linear system. The summed count size is then used to estimate the corrected size factor. The size factors for the pooled cells are then deconvoluted" into cell-based factors.