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 in 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 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 between 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 affect the downstream analyses.

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; 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 were obtained due to sampling. They affect genes with a count distribution near zero and have to be dealt with the normalisations steps.

To deal with technical noise Unique Molecular Identifiers (UMI) or spike-inns 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-inns 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 was specifically developed for the clustering of single cells. The aim of this study is the evaluation of 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 Neighborhood Embedding (tSNE). PCA finds a rotation of the original data in which the newly obtained first coordinates have the highest possible variance, the second coordinate the second greatest variance etc. Practically PCA is computed by spectral decomposition of the correlation matrix of the original data. Dimension reduction is achieved by selecting only the first few PCs whose eigenvalues are less than average or determine the number graphically by plotting the eigenvalues. PCA is deterministic, relatively fast but restricted to linear spaces.

In contrast, tSNE is a non-linear mapping (Van Der Maaten, 2013). Stochastic neighbor 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 $1/\sqrt{(2)}$. SNE minimises the divergence between $p_{j|i}$ and $q_{j|i}$ using the Kullback-Leiber divergence. tSNE implements a Student-t distribution for the low dimensional space and symmetric version of the cost function to simplify optimisation and to overcome the crowding problem. In tSNE the cost function uses joint probabilities p_{ij} and q_{ij} instead of conditional probabilities. To deal with large data sets 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). 11 clustering methods are evaluated for this study. These methods and algorithms can roughly be classified into three groups: K-means, graph or hierarchical based clustering. SC3, SIMLR, Linnorm, RaceID use k-means in different fashions. pcaReduce and CIDR are based on hierarchical clustering. 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. Drawbacks of the method are that is assuming spherical cluster and its sensitive 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. Using pcaReduce q cluster partitions with $k-1$ clusters are obtained.

SC3 Implemented in the SC3 method is a gene and cell filtering and 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 0 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 5000 cells. Randomly selected cells are then used for the clustering approach described before. 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). Nodes are the data points and 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 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 its neighbour nodes and edges. For each node, a local degree is computed and a node removed from the subgraph if the degree is lower as a 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 as well removed. 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 lead 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. Assumptions are 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 and 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. 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 were one gene is zero has 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 drop-out probability $\hat{P}(u)$ using the whole data set. (iii) Calculation of dissimilarity using Euclidean distances together with pairwise 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 drop-out. (iv) Dimension reduction using PCA on the imputed distance matrix. (v) Hierarchical clustering using the first few PCs. The number of PC can be determined by several methods. Here we use an implemented variation of the scree method.

Seurat Seurat uses raw counts, filtering is done gene- and cell wise. A user-specified threshold for the minimum number of expressed features per cell and a minimum number of gene-wise expression per cell. Scaling, log transformation and normalisation of the counts is done with a scale factor of 10000, a log2 transformation.... In second gene filtering step low variance genes are filtered out. Clustering is done using a PCA latent space representation for computing the Jaccard distances. The Louvain algorithm then is used for clustering.

and a smart local moving algorithm (SLM). Here a resolution parameter defines the number of clusters.

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

First, the values are scaled by 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 is retained. Note that this threshold is set to maintain at least three non-zero cells per gene 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 mean vs. 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 **??**. PCA dimension reduction on the Preprocessed gene expression data is performed. Preprocessing is done by log transformation and adding a pseudo count. Low expressed genes are filtered out based on the zero-proportion and their covariance. Clustering was done by model-based clustering. The travelling salesman problem (TSP) is then solved by a minim spanning tree. The user can then define the starting/ending point through the available biological information and compute a pseudo time ordering score.

ZINB-WaVE The method models the counts as an a zero-inflated negative binomial (ZINB) model that accounts for the zero-inflated, over-dispersed nature of scRNAseq count data(Risso et al., 2017). The model allows for the adjustment of gene and cell level confounders. Based on the preprocessing steps genes with less than five counts per feature were filtered out. Also, it is recommended to use only high variable genes.