0.1 Evaluation of the clustering methods using different run modes

Using filtered and normalised data the methods were operated under the default mode, with the number of clusters given by the ground truth and under a range of parameters. Additionally, the methods were also run with the unfiltered datasets.

An overview of the filtering and normalisation steps is given in Table **??**.

Many users will use the methods on the default mode, hence it was seen important to provide results without any fine tuning of parameters. The run parameters in the default mode were given either by the default setting of the packages or by examples in the different package vignettes. If the method was able to detect the number of subpopulations, this auto-detection function was used to infer the number of clusters. When the number of clusters had to be provided, the number of clusters given by the authors annotation were used.

Clustering result have to be evaluated by some sort of "ground truth". Here, the cell annotation provided by the authors or the the given truth from the simulations were used.

Seurat, TSCAN, RaceID, SC3 and Linnorm have their own filtering and/or normalisation procedures. To test these methods preprocessing capabilities the methods were tested with the unprocessed raw counts. The methods tSNEkmeans, pcaReduce, SC3, SNN-Cliq, SIMLR, ZinbWaVE and CIDR do not include filtering and normalisation steps. For these methods filtered, normalised and log-transformed counts were used.

In a further analysis, the clustering methods were tested for different values for the number of clusters k. Seurat does not allow the setting of the number of clusters. Hence, Seurat was run under a range of the parameters KNN and the resolution parameter.

Based on the evaluation by the ARI metric, the parameter k which maximizes the ARI score was used to compare the methods in an optimal setting.

To asses the stability of the clustering methods a random subsample without replacement was drawn from the Kumar dataset. The size of the subsample was 100 and the subsampling was repeated 30 times. The clusterings were then compared using the overlapping samples and the ARI scores.

0.2 Parameter settings

The number of cluster k is the main parameter for most of the methods. Except for Seurat, the clustering functions allow for the direct controlling of the number of subpopulations. Seurat allows the setting of k only indirectly trough a resolution parameter. Other important parameters were the number of kNN, the number or the type of latent space dimensions used for the clustering algorithms and the settings of the filtering and normalisation steps. pcaReduce, SC3, Linnorm, RACEID, TSCAN can be run under an unsupervised mode, and no parameters have to be provided. Although its possible to run these the methods unsupervised, fine-tuning of the parameters is highly recommended. CIDR, RtSNEkmeans, SIMLR need the specification of the number of clusters k. For SEURAT and the number PCs or the number of the kNN have to be defined. A full listing of the parameter settings for each method and run mode is provided in table .... Next, a brief overview of the chosen parameter setting and the rationale behind it is given.

RtSNEkmeans To reduce the run time the Barnes-Hut tSNE implementation from the R package Rtsne is used. Perplexity was set to 30 for all data sets. We note that the value of the perplexity can give different tSNE representations, however here the default setting with the perplexity parameter set to 30 was chosen. tSNE is performed on the first 20 dimensions in the PCA latent space.

pcaReduce For pcaReduce, the range of clusters cannot be specified. Instead, the number of dimension q in the PCA latent space are to be specified. The results are $q-1$ different clustering solutions, with $k-2$ clusters. For all data sets, 30 dimensions were chosen, and evaluation was based on the respective number of clusters in the subsequent analysis. The method is stochastic and has to be run several times for stable results. Here 100 samples were chosen. Merging of clusters was done by sampling proportional to the joint probabilities.

SC3 A gene filtering step is implemented in the method. Based on the dropout distribution genes that are below the 10th and above and 90th percentile are filtered out. However, for the Koh and Zheng data set the upper threshold is set to the 99th percentile, otherwise it was not possible to run the method. When running under the default, a range of cluster from 2 to 10 is given, and the number of subpopulations is automatically inferred by the method. Otherwise, k is set to the number of annotated subpopulations.

SIMLR Implemented by the method is a gene-wise data scaling step. When running under the default mode no scaling was used. In the other run modes scaling was performed. The tuning parameter k was set to the default value of 10 on all runs. The number of clusters is set accordingly to the run mode.

CIDR CIDR uses three parameter settings; the number of clusters, the number of PCs (nPCs) and the method for hierarchical clustering. By default, Ward distances are used in the hierarchical clustering. CIDR can infer the number of from 2 to n clusters. By default n is set to $nPC * 2 + 2$ and the parameter nPC is 4 by default. When not run in the default mode we choose the nPC according to a variation of the scree-plot and set the number of clusters accordingly to the respective data set. The number of used PCs for the data sets Kumar, Trapnell, Koh, Zhengmix and simDataKumar and simDataKumar2 are 5, 10, 8, 8, 3 and 3, respectively.

Seurat Implemented in the method is a normalisation and a gene filtering step. The filtering criteria are the minimal number of gene expression in the cells and the number of total features per cell. By default, no cell filtering step is included when preprocessed datasets are used.

When running with unfiltered data, genes which are expressed in less than two cells were filtered out. For the Zhengmix data, the threshold is set to one, according to the filtering used in the QC and normalisation steps.

The default log normalisation is used, which is currently the only option. The scale factor for cell-level normalisation was set to the default of 10000. As a default, no explanatory variables were chosen to be regressed out. The experimental batch would be a natural choice as a covariate, but cannot be used as the datasets containing this information are completely confounded

The clustering parameters to be defined were a resolution parameter and the number of PCs for the clustering. The resolution parameter was set to 0.7 for the Koh data and 0.6 for the other datasets. The number of PC was determined by the methods recommended by the authors. By the use of a scree plot and a jackknife permutation test the number of PCs was determined. For the datasets, Kumar, Trapnell, Zheng and Koh of 9, 12, 10, 15, 10 and 10 PCs were used for the dataset, respectively. Ten percent of the total cells were used for the number of neighbours in the k-nearest neighbour algorithm. A range from 0.5 % to 40 % of the total number of cells is used to infer the optimal number for the kNN parameter.

TSCAN TSCAN adds a pseudo-count of one, and the data is log-transformed, this setting is used for all run modes. In the default run mode, genes that show-zero expression in at least half of cells are filtered out. When running with unfiltered data, this threshold was changed to 0.1 for the Zheng, simDataKumar and simDataKumar2 data to be able to run the method for these methods. This filter was switched-off when working with the pre-filtered datasets. By default, the method infers the number of clusters from a range of 2 to 9 clusters. Here, a range from 2 to 10 was used in the default mode. If run semi-supervised the respective number of clusters is given. By default "ellipsoidal, varying volume, shape, and orientation" is used for the model.

SNNCliq The connectivity of the quasi-cliques was set to the default value 0.7. Likewise, the merging threshold parameter was set to the default of 0.5. The method was run with normalised, filtered data and the number of clusters was set to a range from 3 to 10 in all datasets. SNNclique works on different distance metrics, here the default euclidean distances are used.

RaceID For the default run mode, cells with a minimum total library size of 3000 are retained. The gene filter is set to filter out all genes with less than five transcripts in at least one cell. Oversaturated genes that have over 500 transcripts in one cell are as well filtered out. Here, we use this filter only for the Zheng data as it is the only set that contains UMI counts. Otherwise, the filter is turned off.

For the run mode with unfiltered datasets the minimum total library size was set to 1000, 500, 400, 420, 1200 and 1200 for the datasets Kumar, Trapnell, Koh, Zheng, simDatkumar and simDatakumar2, respectively. These thresholds were chosen, such that they correspond according to the thresholds based on the MADs.

The filters for oversaturated genes and minimum gene expression were turned off as well, corresponding to the filtering steps in chapter 2.3. Filtering based on the original analysis was done using mean counts. To set the gene filter, we retain genes that show at least one transcript for two cells in the count based datasets. Except for the Zheng data, where genes are retained that show at least five counts in two cells.

If the method is run with already prefiltered datasets, the filters are turned off, and the appropriate number of clusters is provided. In all run modes, the Pearson metric as a distance measure is used. The gap statistic is used to determine the number of clusters. The default setting from the clusterboot function is on a range from 2 to 20 clusters.

Linnorm Except for the Zheng data, the filtering thresholds are set to the default in all run methods, Due to the low sequencing depth in the Zheng data the minimum non-zero expression had to be set to a proportion of 0.1. tSNE and k–means are used for clustering and dimension reduction. By default, a range of k from 2 to 20 was provided to infer k. In the other run modes, the respective k per dataset is supplied.

0.3 Evaluation metrics

One evaluation criteria was the Hubert - Arabje Adjusted Rand Index (ARI) for comparing two partitions (Hubert and Arabie, 1985). The measure is adjusted for chance. The uncorrected Index is subtracted by its expectance divided by the maximum minus the expetance. Independent clusterings have an expected value of zero and are one if there is full agreement between the partitions. The index can take on negative values when the partitions are worse than chance.

Another metric is the F1 score. It is the weighted average mean between precision and recall. The weights are defined as the inverse of the precision and recall. F1 scores can take on values between zero and one. The predicted clusters and the "ground truth" were matched by the Hungarian algorithm. Some of the clustering methods are unsupervised and the partitions does not need to have the same sizes (non-bipartite). This causes problems with Hungarian algorithm. As a solution the assignment matrix is augmented with dummy columns with the maximum matrix value as its entries.