1 Intro

2 Datasets

Kumar et al. 2014 Kumar et al. used *Dgcr*8-knockout and V6.5 variotypes from mouse embryonic stem cells (mESCs). Cells were cultured on serum plus leukaemia inhibitory factor (LIF) or under Erk and GSK3 signalling inhibition (2Li). The authors investigated the expression of pluripotency factors and their involvement in heterogeneity of pluripotent stem cells. Sequencing was done using a Fluidigm C1 system and following a SMARTer protocol. The experimental design is completely confounded, as the conditions and batches are identical.

Trapnell et al. 2015 Trapnell et al used human skeletal muscle myoblast cells to investigate temporal differentiation. Cells were expanded under high-mitogen conditions. Differentiation is induced by switching to low-serum medium. Cells were captured before switching to low-serum medium (T0) , 24 h (T24) and after 48h (T48). Cell lines were harvested on the start of the experiment and after one and two days. Between 49 and 77 cells were isolated at each time point and used for single mRNA-Seq library preparation. Libraries were sequenced with paired-end sequencing on a HiSeq 2500 (Illumina) platform. Sequencing depth was 4 million reads per library. Note: Authors excluded libraries that contained fewer than 1 million reads. As the three different time points are on different plates the experiment confounded and no difference between biolgical and batch effects can be made.

Koh et al. 2016 H7 human embryonic stem cells (hESCs) were used to study human mesoderm developement. Starting from undifferentiated H7 hESCs several differentation stages, sorted by time point and further refined by fluorescence activated cell sorting (FACS) were isolated. Finally ten different cell lines were obtained namely anterior primitive streak populations , mid primitive streak populations, lateral mesoderm , FACS-purified GARP+ cardiac mesoderm, FACS-purified DLL1+ paraxial mesoderm populations, early somite progenitor populations , dermomyotome populations and PDGFR+ sclerotome populations. Before library preparation cells were checked for degradation and containing only a single cell. In total 10 different cell types were then sequenced on Fluidigm C1 and following SMARTer protocol. Sequencing depth was 1 to 2 million reads per cell. Note that the authors discarded libraries with less than one million reads during the quality control process, finally using 498 out of 651 cells.

Zheng et al. 2017 FACS-purified fresh periphelar blood mononuclear cells (PBMCS) sub-populations were sequenced using a droplet-based system. Gene filtering was done using only genes that showed at least one UMI count in at least one cell. FACS purity was 98 - 99 percent. Four of these purified filterd cell populations; namely CD19+ B, CD8+CD45RA+ naive cytotoxic, CD14+ monocytes and CD4+/CD25+ regulatory t cells were selected and used for the clustering analyses. CD19+ B and CD14+ monocytes cells are distinct cell populations. Whereas CD8+CD45RA+ naive cytotoxic cells and CD4+/CD25+ regulatory t cells are not distinguishable by tSNE dimension reduction and kmeans clustering. To construct an artificial population 200 cells were sampled from these libraries and merged to obtain a single expression matrix.

Simulated dataset Using the Splatter package (Oshlack et al. 2017) expression data were simulated. Parameters for the simulation were estimated from the Kumar dataset and a expression matrix with n cells and j features were simulated. Three subpopulation with a propabilty of 0.25, 0.5 and 0.25 and a proportion of 0.1,0.2 and 0.4 differentially expressed genes were simulated.

3 Transformation

RNA-seq data may suffer from heteroscedasticity, skewness and mean-variance dependancy. Genes with higher mean have on average a higher variance across cells leading to unequal variances between different genes. Count data are known to have a skewed distribution. Shows some examples.... To account for that, different transformation were considered. Logarithmic, arcus sin and a variance-stabilizing transformation (VST) of the data are used. Log

transformations will have an impact on extreme values and after transformation, the distribution should be more normally distributed. However, log transformations do not address the problem of heteroscedasticity. Arcus sin transformation should deal with extreme values and equalize the variances. After transformation the mean and the variances should be independent. VST address the problem of extreme values and unequal variances across genes. After such transformation, the mean and the variances of the genes should be independent. Box-Cox transformations adress the problem of extreme values, also the data should be less skewed... Using log transformation and VST the mean-variance dependence is less extreme (see Figure). Still, for means in the lower-midrange, the variances are not equal.

4 Filtering and Normalization

The quality control of the datasets follows Lun et al. Length scaled, count scaled transcript per million (TPM) were used for the datasets Kumar, Trapnell and Koh.For the Koh dataset UMI counts are used. To find potential outliers PCA on the pheno type characteristic (example) of each cell can be used (Figure). Kumar and Trapnell show some potential ouliers.

Cells with log10-library sizes that are more than 3 median absolute deviations (MADs) below the median log-library size were filtered out. The same filtered 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 a ERCC proportion above 3 MADs are as well removed. The same filter was used for mitochondrial gene expressionin in the Zheng data. For the Trapnell 2014 data set information about the cell quality was available. In this dataset cells that were marked as debris, or if a single library consist of more than one cell were as well filtered out. Leaving 531 cells in the Koh dataset, 246 in the Kumar dataset and 222 in the Trapnell dataset. The filtering was less strict in the Koh data set compared to the original analysis 2016 were they retained 498 cells.

Low-abundance genes were filterd out by removing features that have an average count below 0.0001. For the Zheng data features which are not expressed in at least two cells are removed.

To find batch effects a linear model regressing the PC values against the total features was used (Lun et al, 2016). No correlation can be seen in the dataset Trapnell and Zheng. Wheras for Kumar and Koh PC1 has a high correalation with the number of features.

Another examination of the technical factors that have an influence on the variances can bo done using the marginal variances. For that a linear model with the expresion values per gene as response variables and a chosen explanatory variable is fitted. The correlation coefficant can than be seen as the marginal explained variance for the explanatory variables. In Kumar similar amount of the variance is explained by the total number of genes, the proportion of ERCC and the phenotype. This indicates that the data set is heavily influenced by batch effects. They same holds in the Trapnell data, but on a lower scale. Variance in Koh data is largely influenced by the phenotype and to a lesser extend by total number of genes and the top 200 features. Zheng is largely dominated by the biological variation with the other explantory factors contributiong only marginally to the complete variances.

scRNA-seq data has an excess of zero counts. These can be split into systematic, semi-systematic and stochastic zeros (Lun, 2016). Systematic zeros are silent across all cells. These features were removed prior to the analysis. Stochastic zeros are zero counts that were obtained due to sampling. It affects genes with a count distribution near zero. Semi-systematic zeros come from genes that are silent in a subpopulation of cells. Different methods exist to normalize RNA-seq data like TMM normalization, DEseq normalization and by library size. However, none of these methods are designed to deal specifically with the zero-inflated nature of scRNA-seq data. Another approach is the normalization by spike-inn. This approach is not feasible as no or only a limited number of spike-inn counts were present. Here normalization through pooled cells was used (Lun et al., 2016). Counts from different cells were pooled together. The summed count size was then used to estimate size factor. The size factors for the pooled cells were then "deconvoluted" into cell-based factors (Lun et al., 2016). By default the expression values are log transformed.

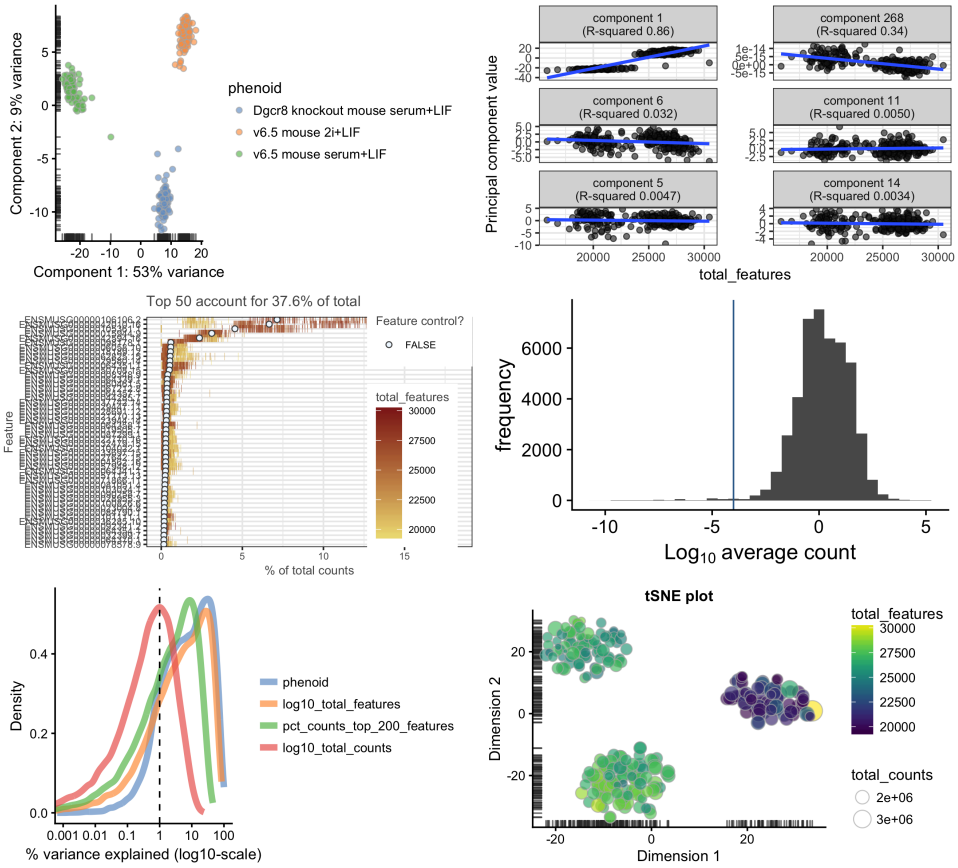

Figure 1: QC summary of Kumar 2015.

5 Optimal number of clusters

Methods to determine the optimal number of clusters are subjective methods as elbow or silhouette plots. In the Elbow plots, the within-cluster sum of square is plotted against a range of clusters. The silhouette plot is a standardized measure of distances between each point inside and outside of the respective cluster. Less subjective is the gap statistic. Here the log within sum of squares is compared to its expectation. The null distribution is expected to be uniformly distributed, it is not clear if this is correct for high dimensional data. Other possible methods are the calinsky criterion, hierarchical clustering.... The elbow plots suggest three clusters for the Kumar dataset, 2 - 5 in the Trapnell data and 3 in Koh 2016 (see Figure **??**). Minimization of within sum of squares was also done in the tSNE latent space with 30 dimensions. Here the optimal number of clusters are 3,4, and 5 in the Kumar, Trapnell and the Koh datasets.

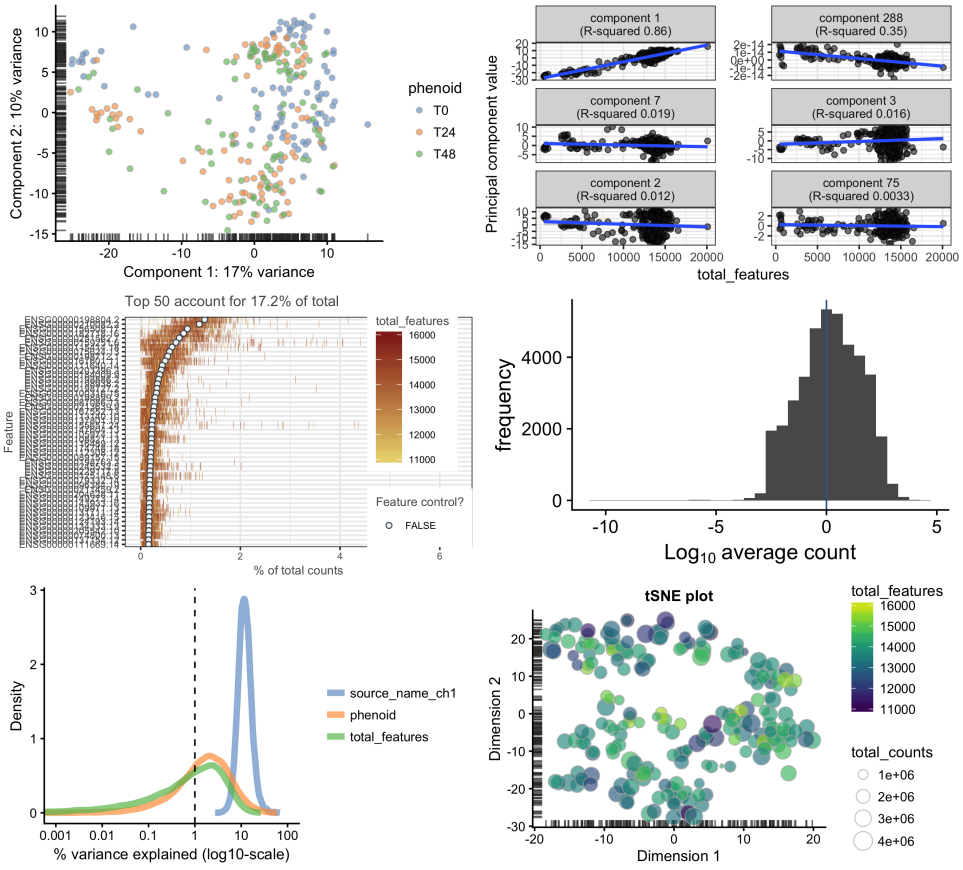

Figure 2: QC summary of Trapnell 2014.

6 Methods

tSNE In contrast to other dimensionality reduction techniques like multidimensional scaling (MDS; Torgerson, 1952) tSNE (t-distributed stochastic neighbourhood embedding) is a non-linear mapping. Stochastic neighbour embedding (SNE) transforms euclidean distances to conditional probabilities pj—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 centred at y_i and variance $1/sqrt(2)$. SNE minimizes 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 optimization and to overcome the crowding problem. crowding problem explains.... In tSNE the cost function uses joint propabilities p_{ij} and q_{ij} instead of conditional probability. Where p_{ij} is formula... To deal with large data sets the Barnes-Hut implementation uses random walks on the nearest neighbour network with PCA step to reduce the dimensionality of the high dimensional data.

K-means K-means clustering minimizes the within-group sum of squares with a predefined number of clusters k. K-means clustering uses K centres for the K clusters. The data points are 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. As the variable with the largest range can dominate

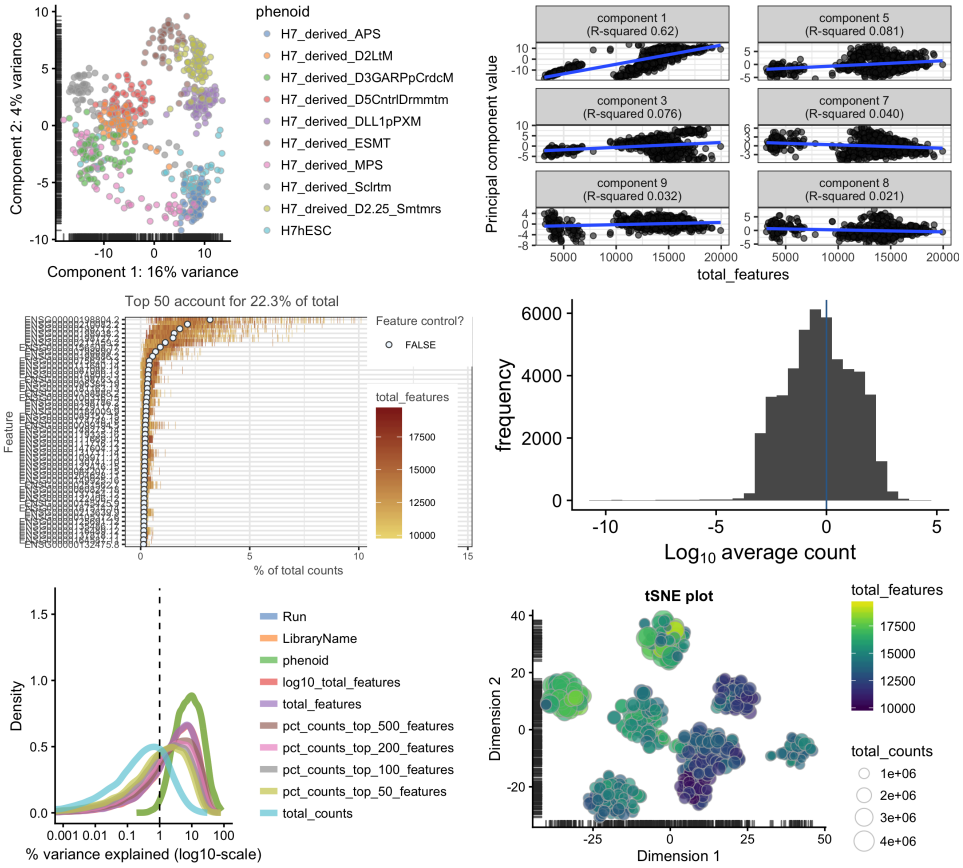

Figure 3: QC summary of Koh 2016.

the other others, it is often advised to use scaled data. PAM is a similar method with the cluster centres defined as a data point in the respective cluster.

pcaReduce pcaReduce uses PCA and kmeans clustering to find the number of clusters in the reduced dimension given by PCA. The method expects 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 an gene expression matrix, the clustering algorithm starts with a K-means clustering on the projections Ynxq with q+1 clusters. The number of initial clusters K is typically around 30, guraenteering that most cell types are captured. For all pairs of clusters the joint probabilties are computed. two cluster are merged togeter by selecting the pair with the highset 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.

SC3 Implemented in the SC3 method is a gene and cell filtering and log transformation step of the expression matrix. 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 d different dimensions. Next, a consensus matrix of the different clustering results is computed. The consensus matrix is a binary similarity matrix with 1 if two cells belong to the same cluster and 0 otherwise. The consensus matrix is obtained by averaging the individual clustering(how?). The last step is a hierarchical clustering step with complete linkage. The cluster is inferred by the k level of hierarchy, where

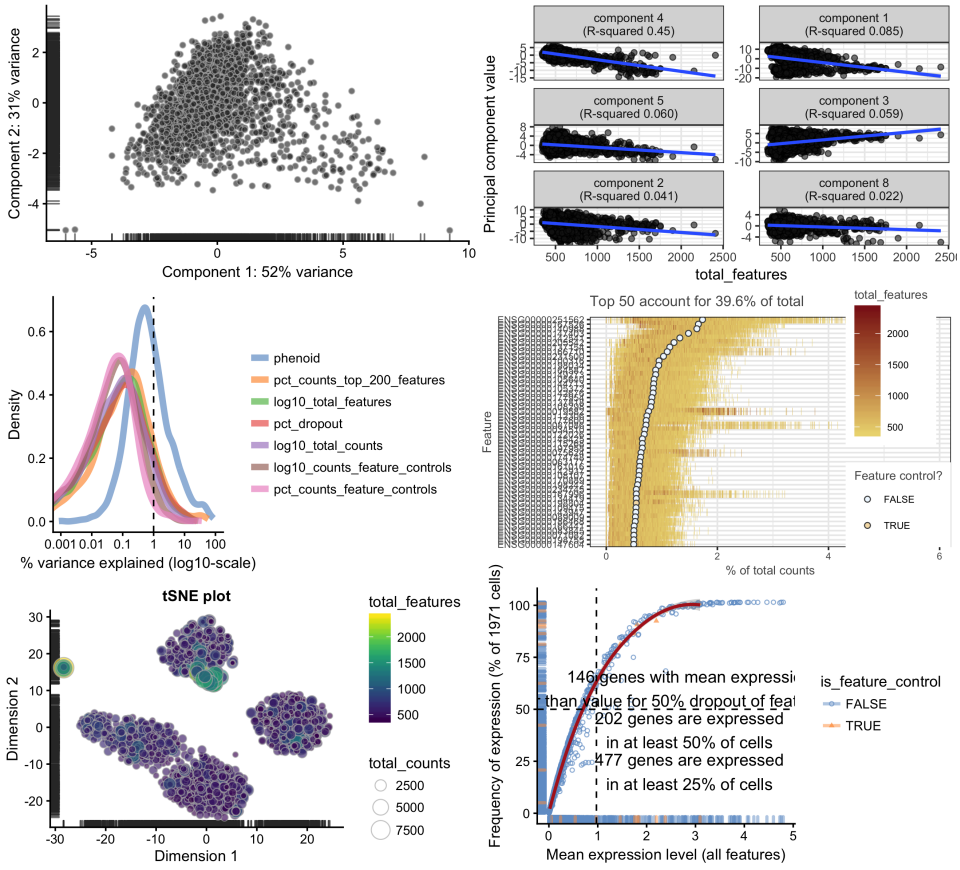

Figure 4: QC summary of Zheng 2016.

k is supplied by the user.

SNNcliq In SNNcliq the high-dimensional data is modelled as a shared nearest neighbour graph. 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-neighbors (KNN) for each data point are listed. The parameter KNN has to be supplied by the user. Edges between datapoints are assigned if they share at least one KNN. The weights of the edges are defined by a function of the number of neirest neighbors and their respective ranks. Identification of clusters is done by finding quasi-cliques associated with each node and merging them to unique clusters. To find maximal quasi-cliques a greedy algorithm is used. A node induces a sub graph which consists of all its neighbour nodes and edges. For each node a local degree is computed and a node removed from the sub graph 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 sub graph 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. 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.

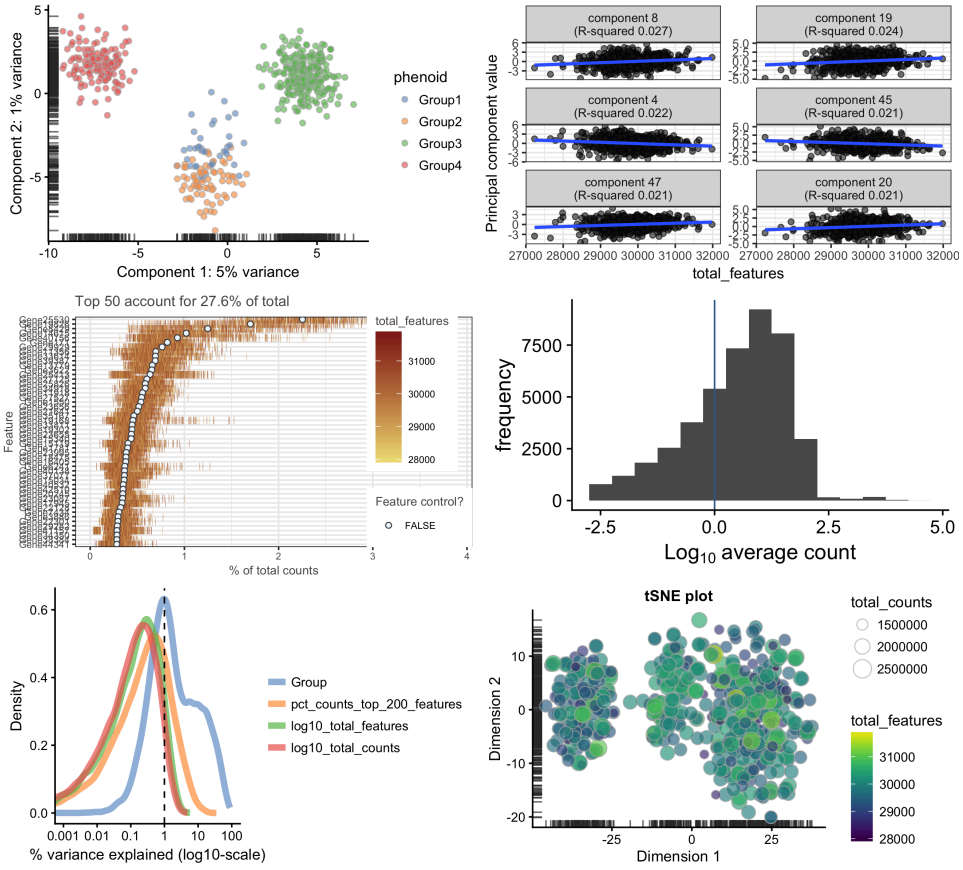

Figure 5: QC summary of simDataKumar.

SIMLR Most clustering method rely on standart similarity metrics like euclidean distances. 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 represents the clusters. the Kernels are Gaussian kernels with a range of hyper parameters defining the variance of each kernel.The similarities are then used for data visualization with tSNE or clustering using k means and the latent space representations of the similarities.

ZINB-WaVE The method is based on a zero-inflated negative binomial (ZINB) model that accounts for the zero inflated, overdispersed nature of scRNAseq count data. Based on the preprocessing steps given by the authors genes with less than 5 counts per feature were filtered out. Also it is recommended to use only high variable genes. ZINB-Wave is based on Factor analysis.

dbscan dbscan is a density based clustering method. A general assumption is that high density areas are well separated by low-density areas. The methods work with euclidean distances, as well as other distant measures. Data points are defined as core points, border points and noise points. A core point is defined as point that lies in a neighbourhood of a neighbourhood with a predefined number of other points. Border points are in the neighbourhood of core points. Noise points are all other points. Each of the points were labeled as core, noise or border points. Edges between all core points that lie inside a neighborhood ϵ were assigned. Connected core points belong to the same cluster. Border points are then assigned to the cluster of the respective core points. The border points can belong to different clusters so there's no unique solution. The number of cluster is not predefined and the cluster can have different forms (but not densities). A disadvantages is that the method performs badly with

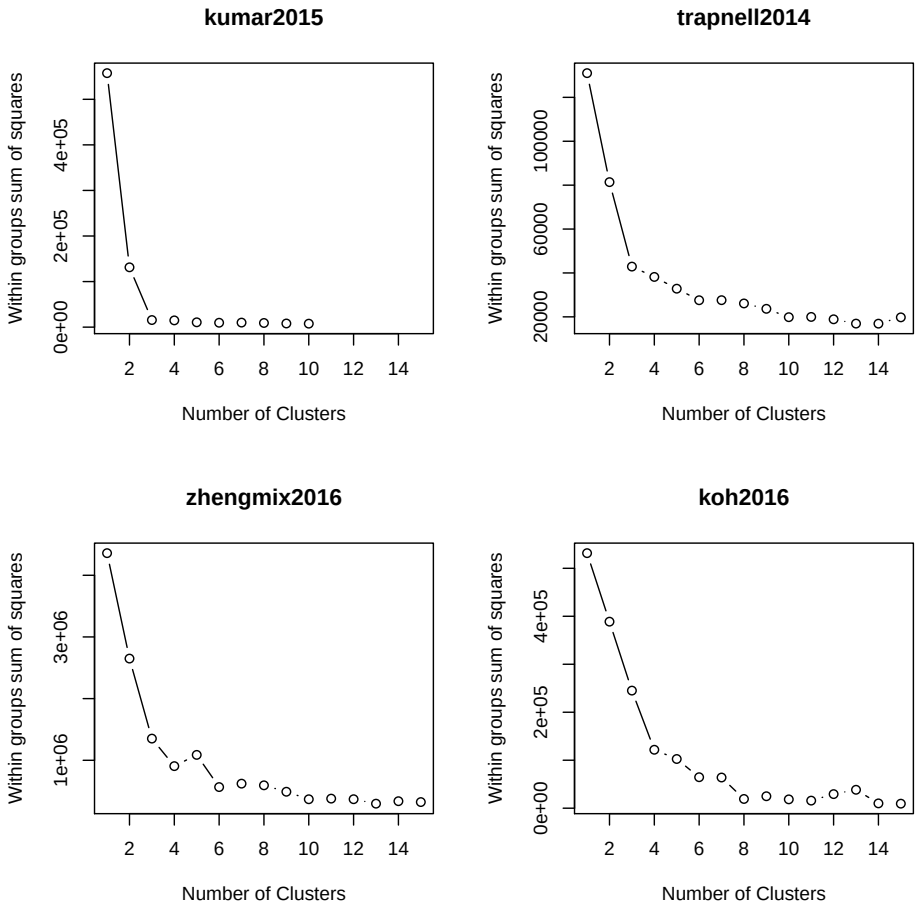

Figure 6: Optimal number of clusters by minimizing within sum of squares based on the latent space of tSNE (30 dimensions)

high dimensional data. So a dimensional reduction step is recommended.

CIDR Clustering through Imputation and Dimensionality Reduction (CIDR) takes the high dropout rate in scRNA seq data into account. CIDR uses TPMs as expression data. The method splits the squared euclidean distance in three terms. One in which both genes k for the pairs i an j are non-zero, one in which one gene is zero and both are zero. The authors state that only the cases were one gene is zero has an 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. CIDR works basicallz in five steps. (i) Find feature that are dropout candidates. That is genes that show a 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 beeing a drop-out. (iv) Dimension reduction using PCA on the imputed distance matrix. (v) Hierarchical clustering using the first few PC. The number of PC can be determined by several methods. Here we use a variation of the scree method.

Seurat Seurat uses raw counts, filtering is done gene- and cellwise. A user specified threshold for the minium number of expressed features per cell and minimum number of genewise expression per cell. Scaling, log transfor-

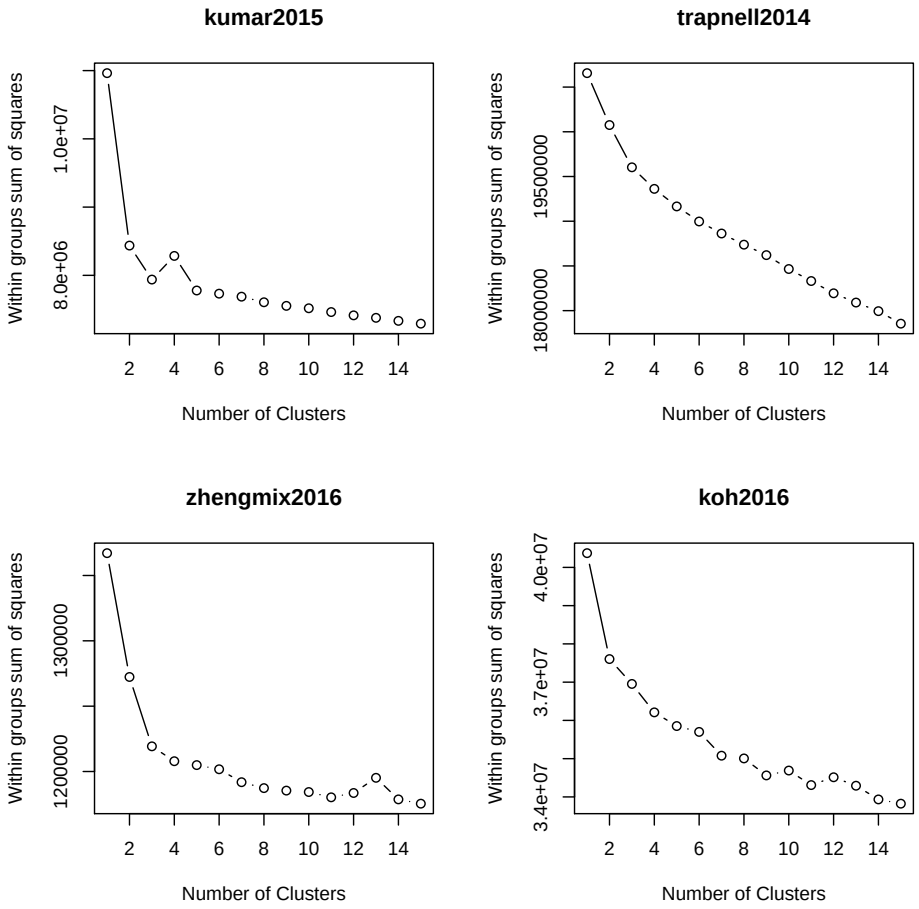

Figure 7: Optimal number of clusters by within sum of squares based on the full dimensions

mation and normalization of the counts is done with a scale factor of 10000, a log2 transformation.... In second gene filtering step low varaince genes are filter out. Clustering is finally done using PCA and a smart local moving algorithm (SLM). Here a resolution parameter defines the number of clusters.

ZIFA ZIFA is a dimensionality reduction technique for scRNA-seq data. To reduce the dimensionality a probabilistic Principal Components Analysis (PCA) that includes a zero inflated model to account for dropout events.

7 Parameters

tSNEkmeans To reduce dimensionality the Barnes-Hut tSNE implemenation is used. Normalized and filtered counts were used as a input. Perplexity was set to 20 for all datsets. tSNE is performed on the default 30 dimension in the initial PCA step. Kmeans clustering was done using 2 to 10 initial cluster centers.

pcaReduce PcaReduce was run on the normalized, filtered counts. The range of clusters cannot be specified, instead the number of dimension q in the PCA latent spave are to be specified. The results are q-1 different clustering solutions. For all datasets 30 dimensions were chosen and the results for 2 to 10 clusters were used in the subsequent analysis. The method is based on kmeans clustering and has to be run several times for stable

Method	Description	dimension reduction	clustering	zero inflation	normalization	supervised
tSNEkmeans	tSNE dimension reduction and kmeans custering	tSNE	kmeans	no	no	no
pcaReduce	PCA dimension reduction and kmeans clustering through an iterative process. Step wise merging of cluster by joint probabilities and reducing the number of dimension by PC with lowest variance	PCA	kmeans, hierarchical clustering	no	no	
SC3	PCA dimension reduction or Laplacian graph. Kmeans clustering on different dimensions. Hierarchical clustering on conensus matrix obtained by kmeans.	PCA	repeated kmeans, hierarchical clustering on similarity matrix of kmeans results	no	no	yes
SNN-cliq	Shared nearest neighbor graph based on similarities. Clustering through forming of cliques and subsequent merging.	graph based	merging of cliques	no	no	
dbscan	Density based clustering	none	density based clustering	no	no	yes
SIMLR		tSNE	kmeans	yes	no	yes
CIDR	PCA dimension reduction based on zero imputed similarities. Hierarchical clustering on a number of PC determined by variation of scree method.	PCA onimputed distances	hierarchical clustering	yes	no	yes
Seurat v1.4	Neirest neighbor graph based on PCA latent space	HVG and PCA		no	yes	yes

results. Here 50 samples were chosen. Merging of clusters was done by default sampling proportional to the joint probablities.

SC3 Clustering was done using normalized, filtered counts. A gene filtering step is implenemented in the method but was not used as the data was already filtered in the previous steps. The number of cluster is to be supplied by the user and clustering on 2 to 10 clusters was done.

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

SIMLR The tuning parameter was set to the default value of 10. Clustering was done by estimating 2 to 10 clusters.

Seurat Implemented in the method are normalizations and a gene filtering step. For the clustering raw counts were used. Several parameters have to be defined. A cutoff value for the minimal expression and the number of total features per cell. Here we used the default value of zero for the anlysis. The expression threshold for a feature was as well set to the default value of zero. The default log normalization is used, currently the only option. scale factor for cell-level normalization was set to 10000. As a default no explanatory variables were choosen to be regressed out. In the clustering paramters to be defined were a resolution parameter and the number PCA dimension to use for the clustering. The resolution parameter was set to the default value of 0.8 . The number of PC was determined by the methods recommended by the authors. Using scree plots and a jacknife permutation test (more exact) to determine the number of principial components. 9, 12, 10 and 15 principal compoments were used for the Kumar,

Trapnell, Zheng and Koh dataset, respectively. 1, 5, 10 and 15 percent of cells were used for the number of neighbors in the k-nearest neighbor algorithm.

dbscan To choose appriobiate parameters for the size of neighborhood epsilon and the minimun number of points in neighborhood the k-nearest neighbor distance is used. As a initial number of neighbors 10 percent of cells is used. then for each point the k-NN distance is computed and plot by increasing order. The chosen values for the distances are 280, 410, 35 and 380 for the Kumar, Trapnell, Zheng and Koh datasets, respectively. The default value for the minimum number of points is 5. For each of the dataset a range of epsilon around the theoretical optimum was chosen to optimize the clustering results.

CIDR Log transformed TPMs were used to analyse the datasets KOH,Kumar and Trapnell. The Zheng data were excluded from analysis as the method is not impemented for UMIs. Parameters to define are the number of

8 Evaluation

One evaluation criteria was the Hubert - Arabje Adjusted Rand Index (ARI) for comparing two partitions. The measure is adjusted for chance and 0 if there's no agreement between pairs and 1 if there is full agreement between pairs. The other criteria is the F1 score. It is the wieghted average mean between precision and recall. With weights defined by the inverse of the precision and recall. F1 scores can take on values between 0 and 1.The predicted clusters and the "ground truth" were match 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 as entries.