1 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 experimantel 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 timepoints 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 peripheral 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.

2 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 1). Still, for means in the lower-midrange, the variances are not equal.

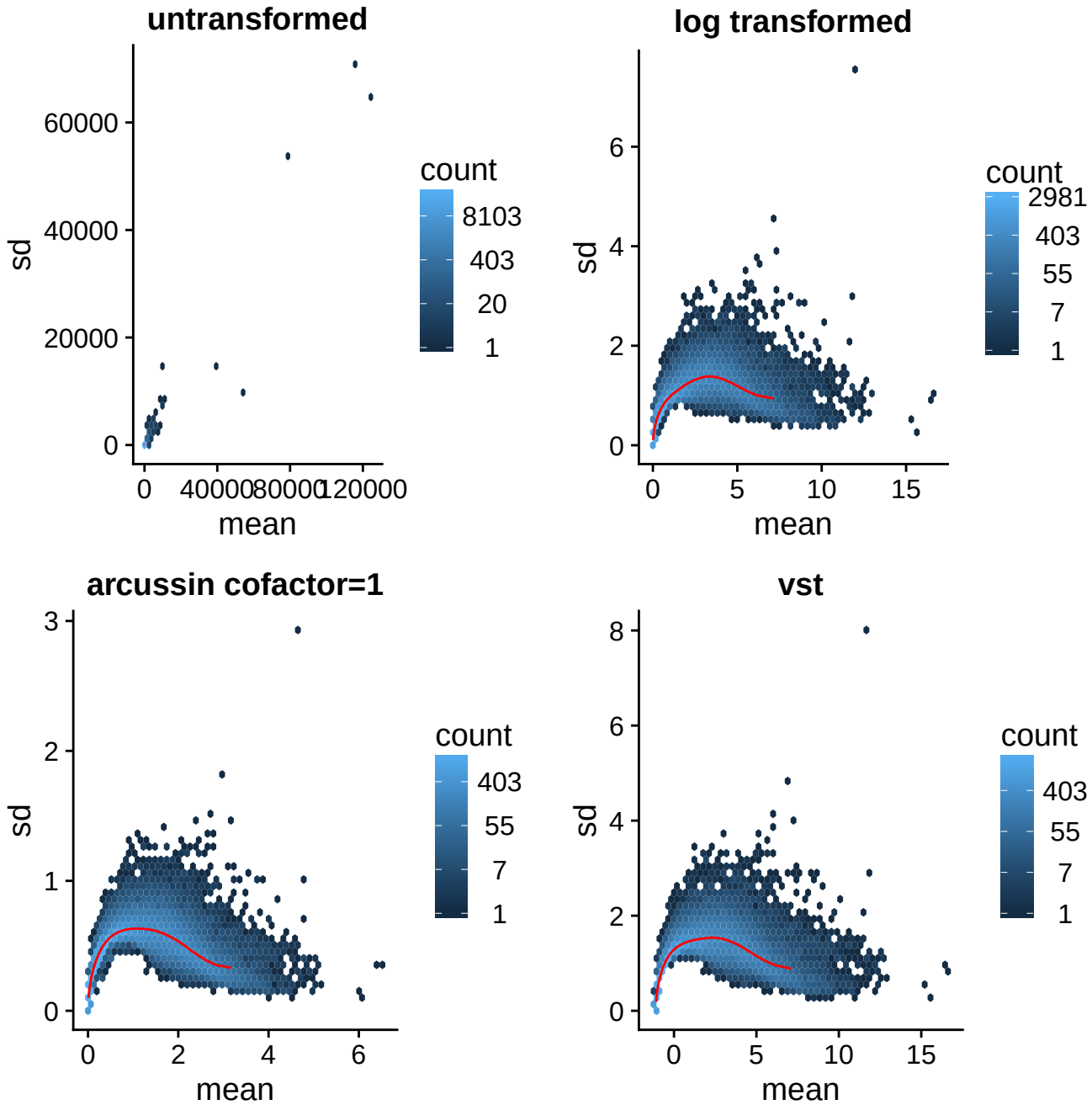

Figure 1: transformations Kumar2015.

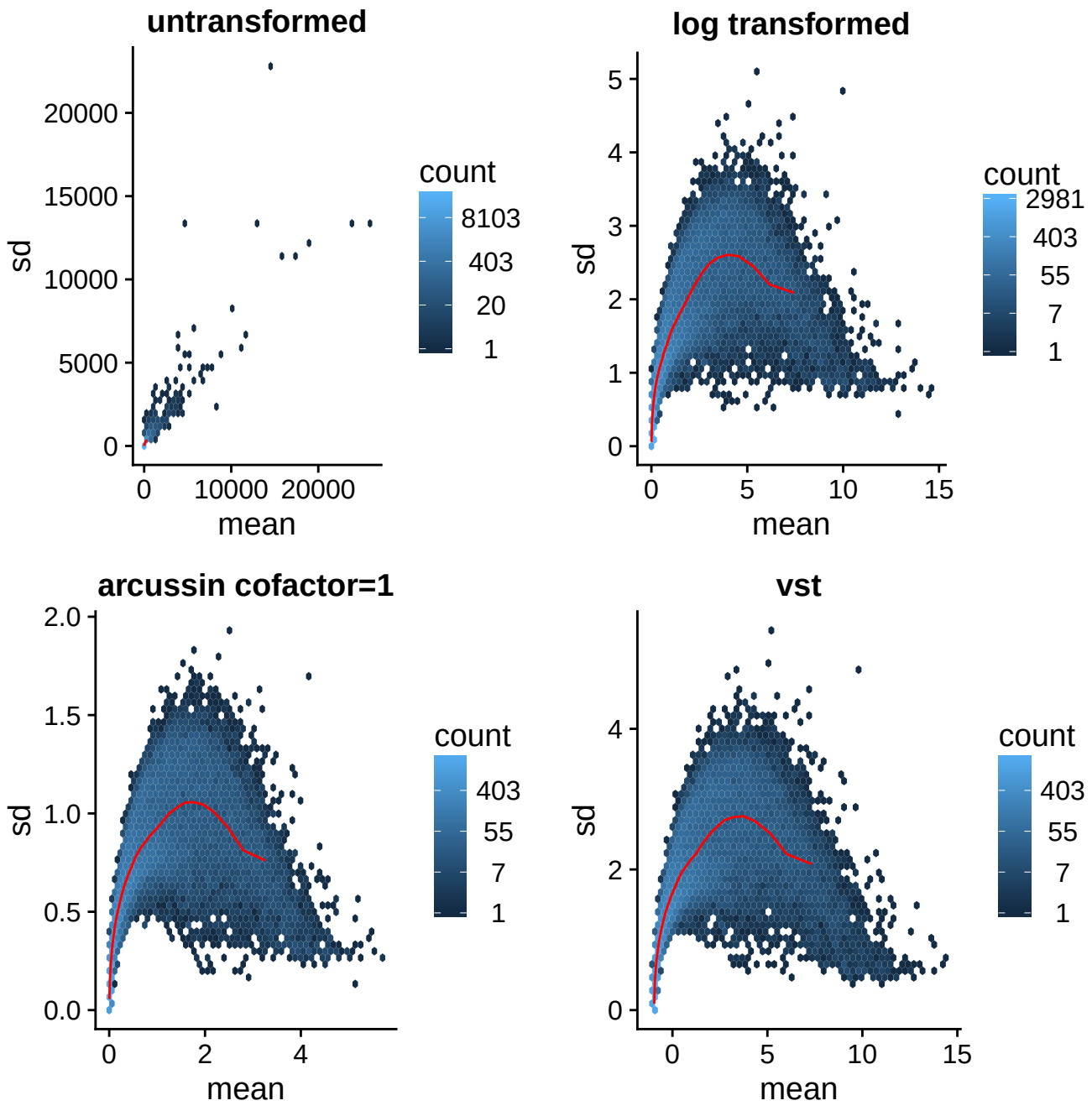

Figure 2: transformations Trapnell 2014.

3 Filtering and Normalization

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. 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

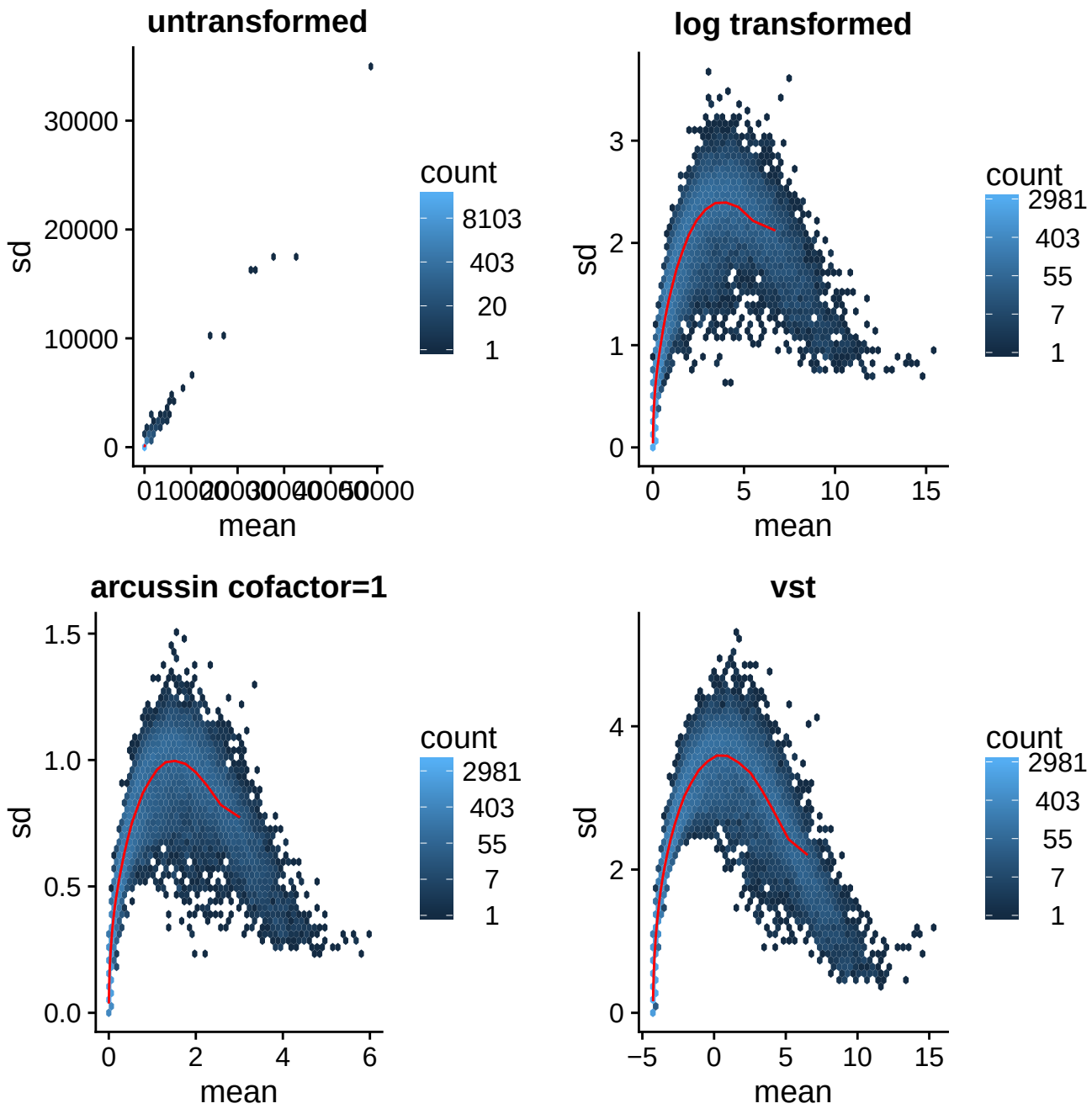

Figure 3: transformations Koh 2016.

number of genes per cell. Similarly datasets that contains ERCC spike-inn features. 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. Low abundance genes were filterd out by removing features that have an avergae count below 1. For the Zheng data features which are not expressed in at least two cells are removed. 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 therefore less strict in the Koh data set

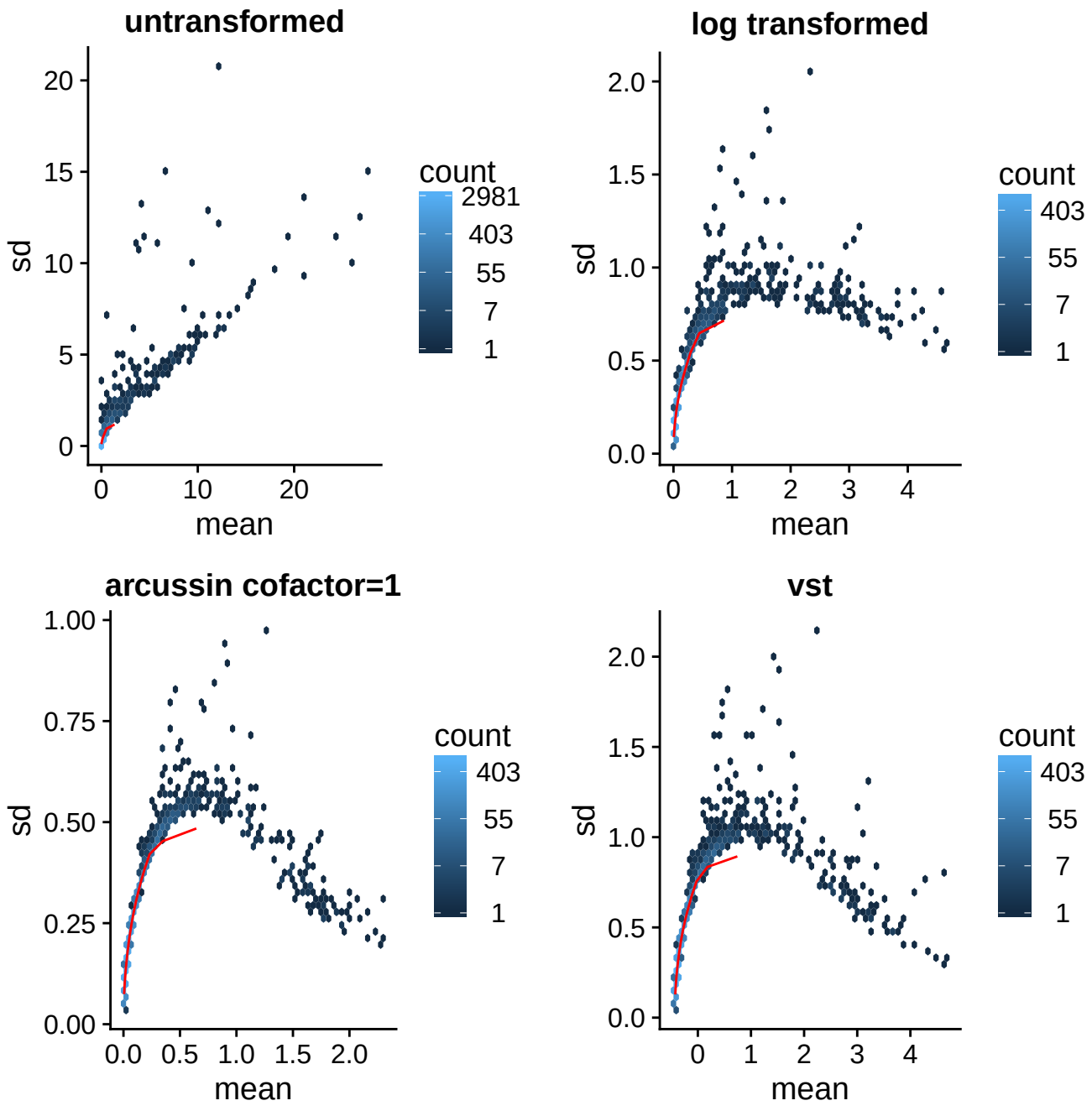

Figure 4: transformations Zheng 2016.

compared to the original analysis 2016 were they retained 498 cells. To find potential outliers PCA on the pheno type characteristic (example) of each cell can be used (Figure 10,8,9,11). Kumar and Trapnell show some potential ouliers. 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

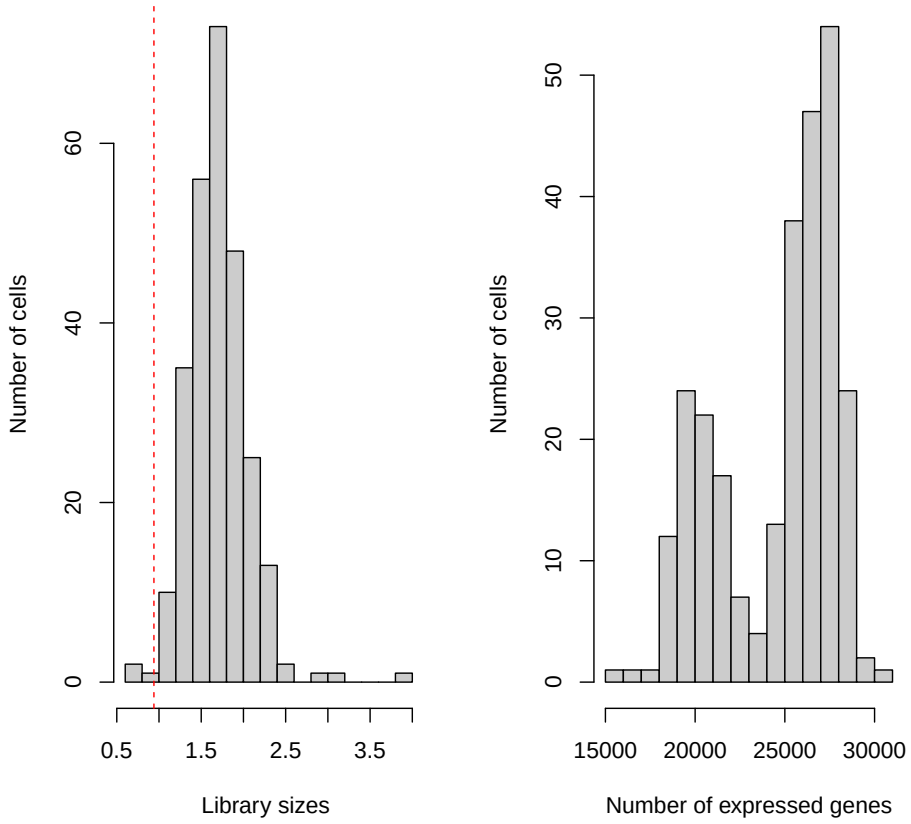

Figure 5: Histogramm of Kumar 2014.

seen as the marginal explained variance for the explanatory variables. In Kumar a big part of the variance is explained by the total number of genes, the expression of ERCC but also by biological variation. Around 0.1 to 10 percent of the varation is explained by the library size, number of genes and the putative cell type. Koh and Zheng is largely dominated by the biological variation.

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.

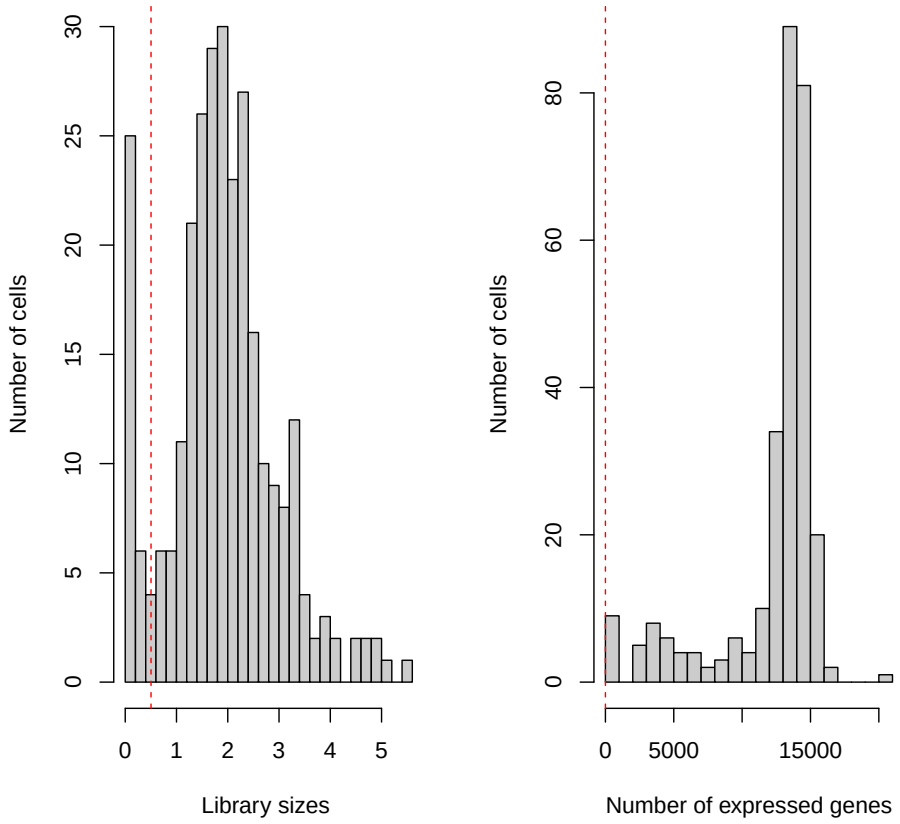

Figure 6: Histogramm of Trapnell 2014.

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 1). 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.

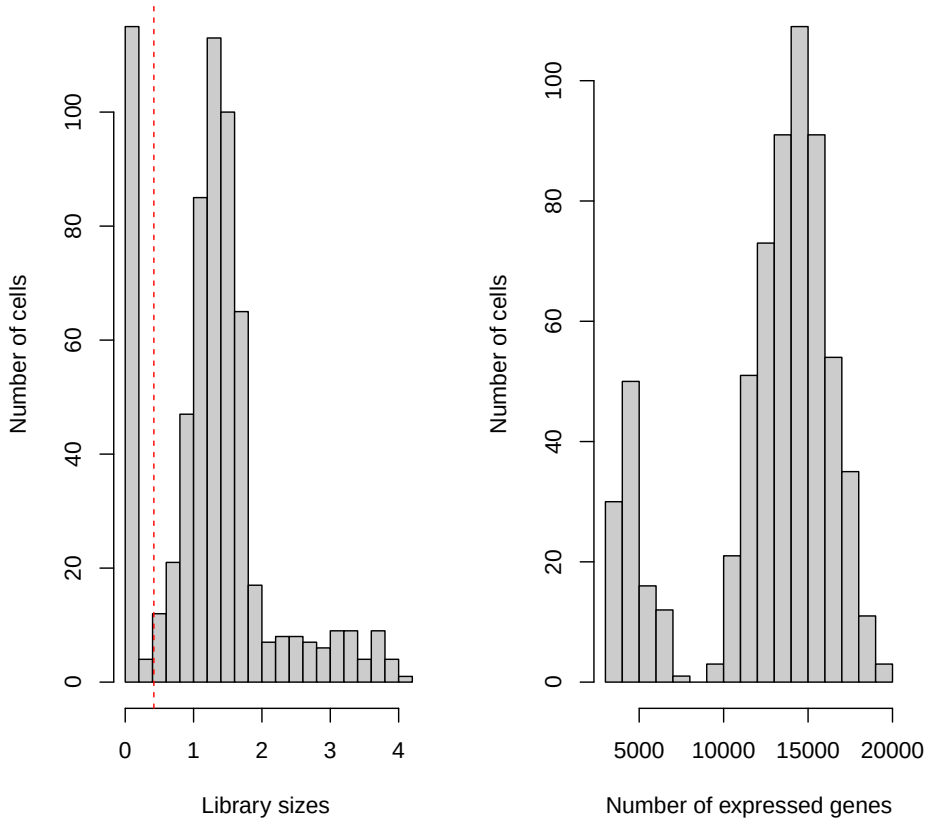

Figure 7: Histogramm of Koh 2016.

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. formula... 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 tries to minimize the within-group sum of squares with a predefined number of clusters k. With the within-group sum of squares formula.... K-means clustering uses K

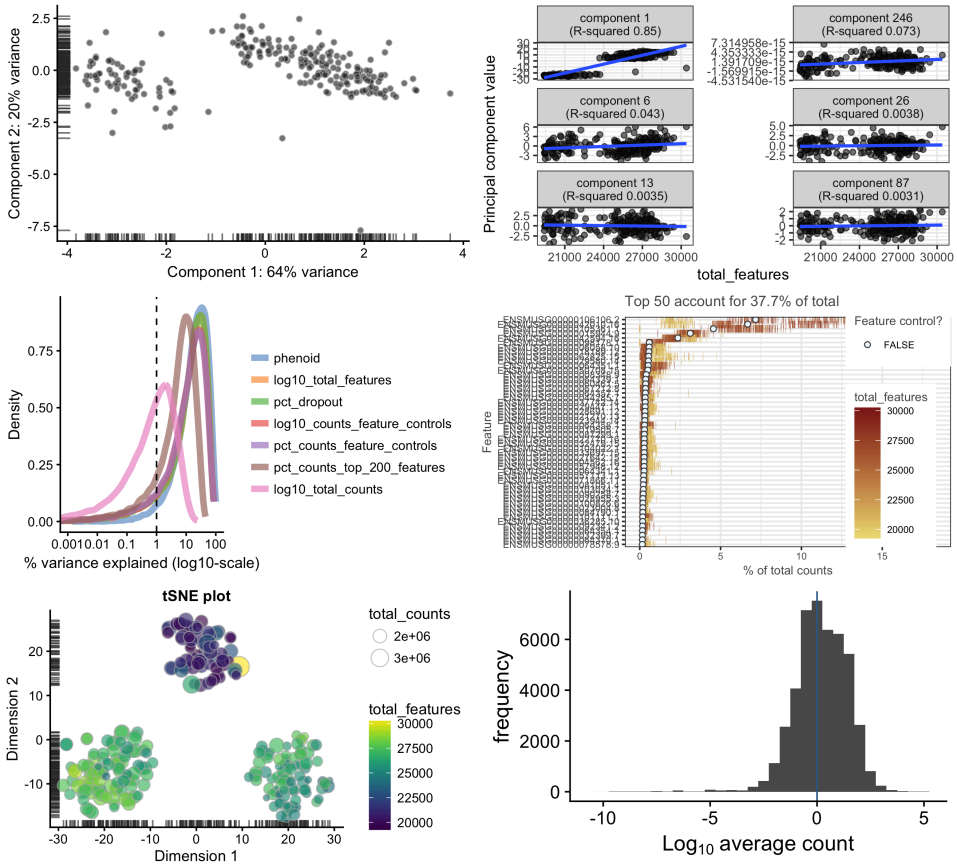

Figure 8: QC summary of Kumar 2015.

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 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.

Gaussian mixture models Gaussian mixture models are defined as formula.... with k different distributions, the prior that any point belongs to cluster m and k different Gaussian distribution is x given that x lies in cluster m. Using a Expectation -maximization algorithm the parameters θ and priors p_j were found. The observation x are assigned to cluster j such that P(x element of j—x) = pjgj(x, thetaj) / f (x;p,theta) is maximal.

pcaReduce pcaReduce uses a 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 ar contained in low dimension PC representation and more refined (subsets) of these cells types are contained in higher dimensional PC representations. Given an original gene expression matrix Xnxg, 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. In an iterative process subsets of the Clusters Yi and Yj were merged together according to their probabilities that they belong to the same cluster. The clusters pairs with the highest probabilities P(i,j) are merged together. 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

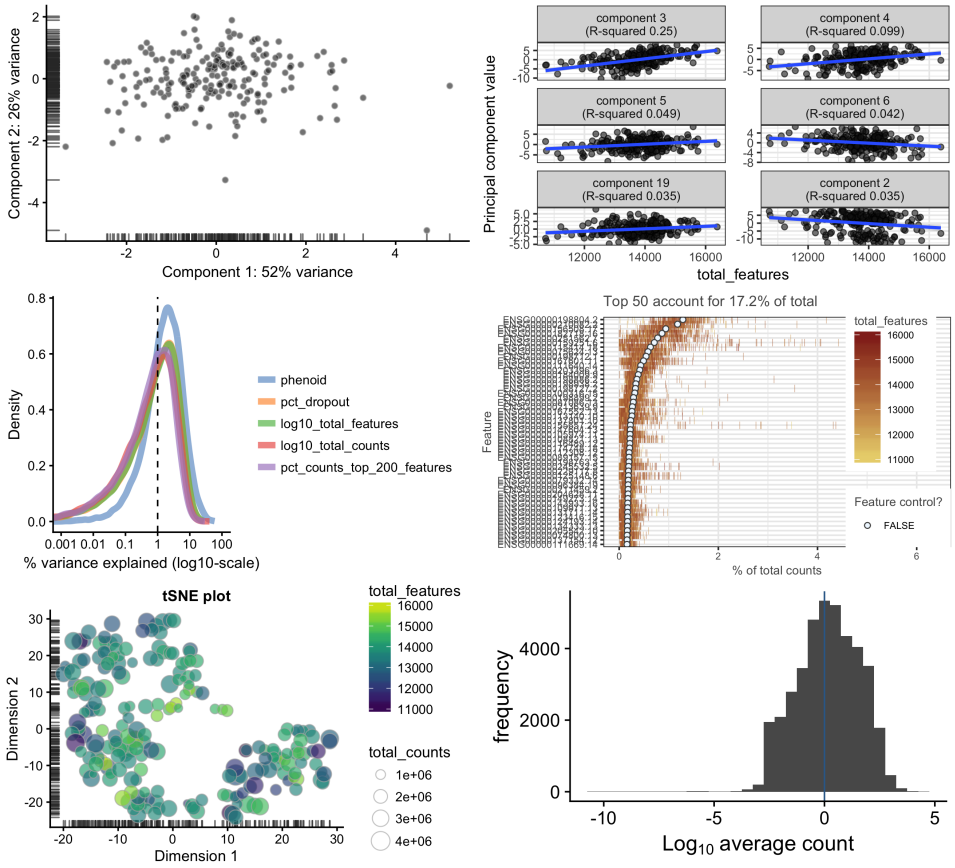

Figure 9: QC summary of Trapnell 2014.

K-2 centres is performed. This process is repeated until only one single cluster remains.

SC3 SC3 uses distance measures of the filtered and log transformed expression matrix and then uses PCA or Laplacian graph for a lower dimensional representation of the data. The distance measures can be Euclidean, Pearson or Spearman. 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 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 xi are listed, with xi as its first entry. The parameter KNN has to be supplied by the user. An edge e(xi,xj) to data point xi and xj is assigned if they share at least one KNN. The weights of the edges e(xi,xj) are defined as followed: formula.... 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 v induces a sub graph S which consists of all its neighbour nodes and edges. The degrees di are computed and the node si is removed from the sub graph if di/S ¡ r. The threshold r is supplied by the user and is typically set

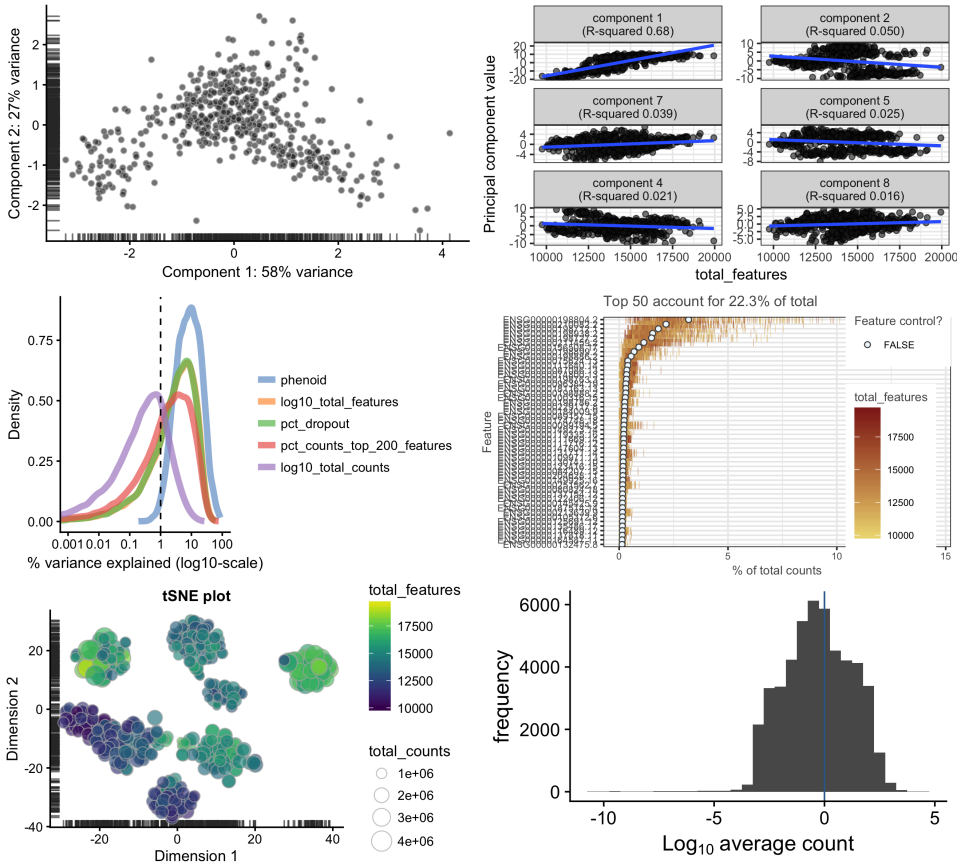

Figure 10: QC summary of Koh 2016.

to 0.7. Next the degrees between the nodes are recomputed and the process is repeated until di/S ¿ r. If the Sub graph S has more than three nodes the quasi-clique is assigned to v. To reduce redundancy quasi-cliques that are completely included in other cliques are removed. Clusters are identified by merging the quasi-cliques. For the Sub graphs Si and Sj 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 the pair Si and Sj with the largest size are prioritized.

SIMLR SIMLR uses a gene expression matrix (normalized) to solve for a similarity matrix S. Assumptions are that S should have a block-diagonal structure with C blocks, where C represents the clusters. Using an optimization framework it minimizes S,L and w. Where S is the similarity matrix, L is a low-dimensional matrix (NxC) and w is the weights vector for the multiple kernels. the Kernels are Gaussian kernels with with a range of hyper parameters defining the variance of each kernel.The similarities are then used for data visualization with tSNE (Barnes hut implementation) or clustering using k means and the latent space representations of the similarities.

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

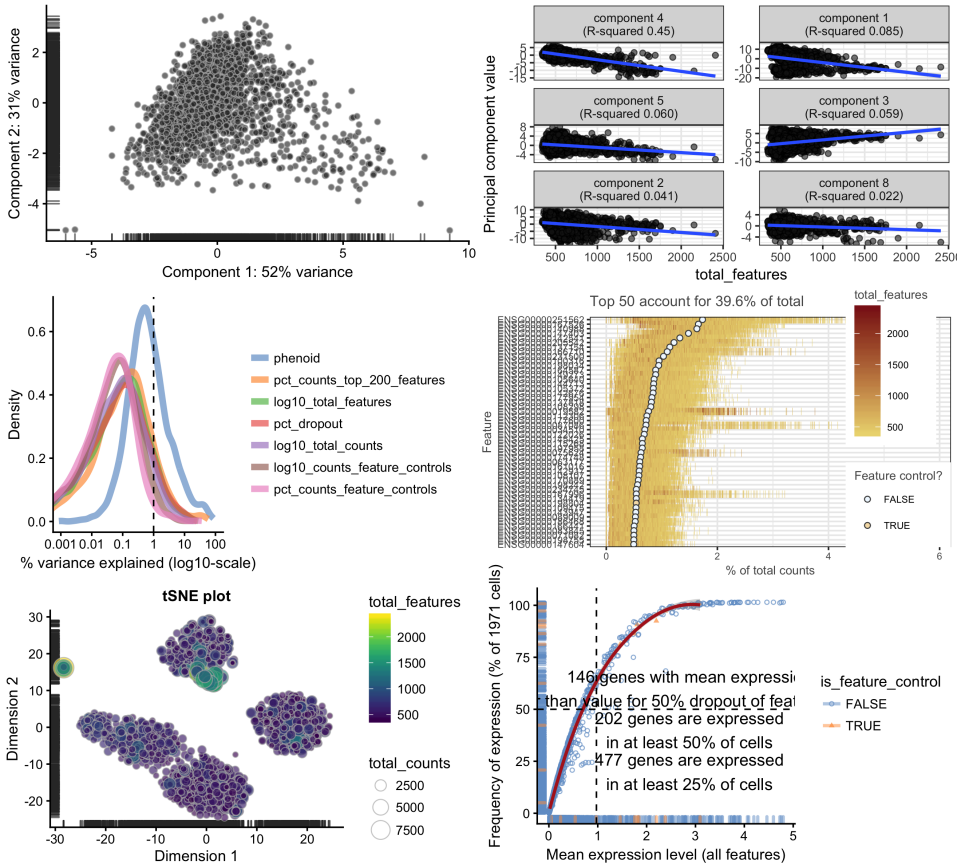

Figure 11: QC summary of Zheng 2016.

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 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 is determined by a variation of the scree method.

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.

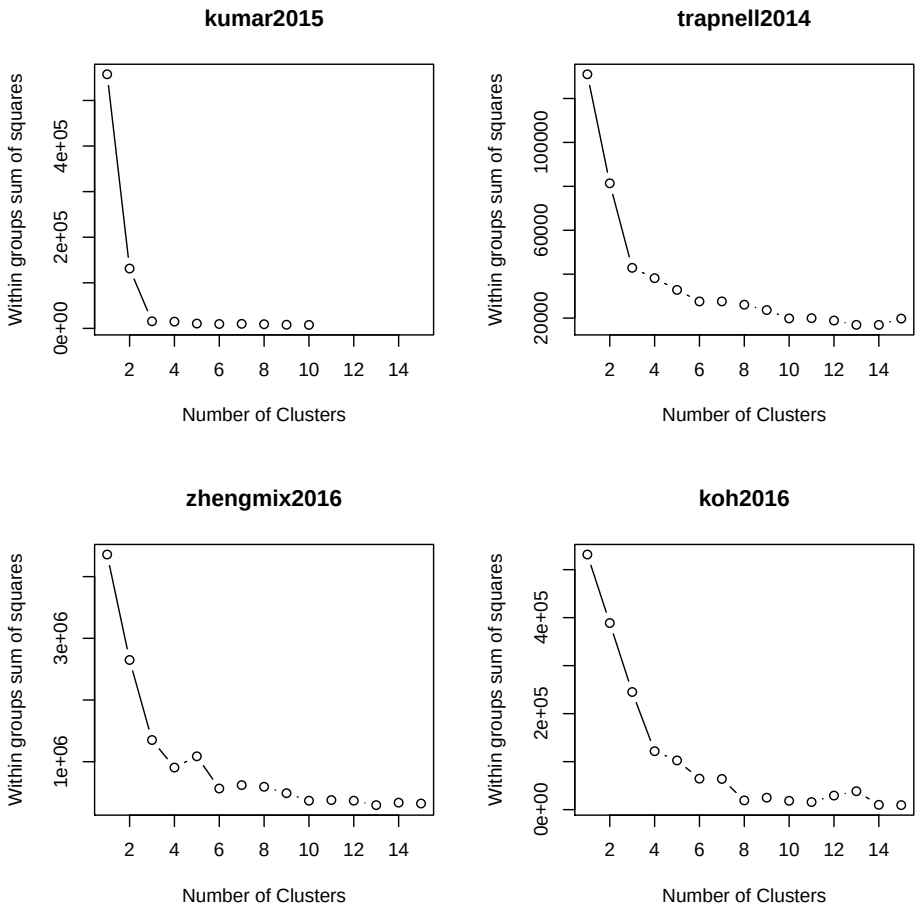

Figure 12: Optimal number of clusters by minimizing within sum of squares based on the latent space of tSNE (30 dimensions)

7 Results

8 Methods

tSNEkmeans To reduce dimensionality the Barnes-Hut tSNE implemenation is used. Normalized and filtered counts were used as a input. Perplexity was set to 30 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 using normalized , filtered counts. The range of clusters cannot be specified, instead number of starting dimension q are to be specified. Resulting in q-1 different clustering solutions. For all datasets 30 dimensions were chosen and the results for 10 to 2 clusters were used in the subsequent analysis. The method is based on kmeans clustering and has to be run several times for stable results. Here 50 samples were chosen.

SC3 Clustering on 2 to 10 clusters 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. A range of 2 to 10 clusterd were used.

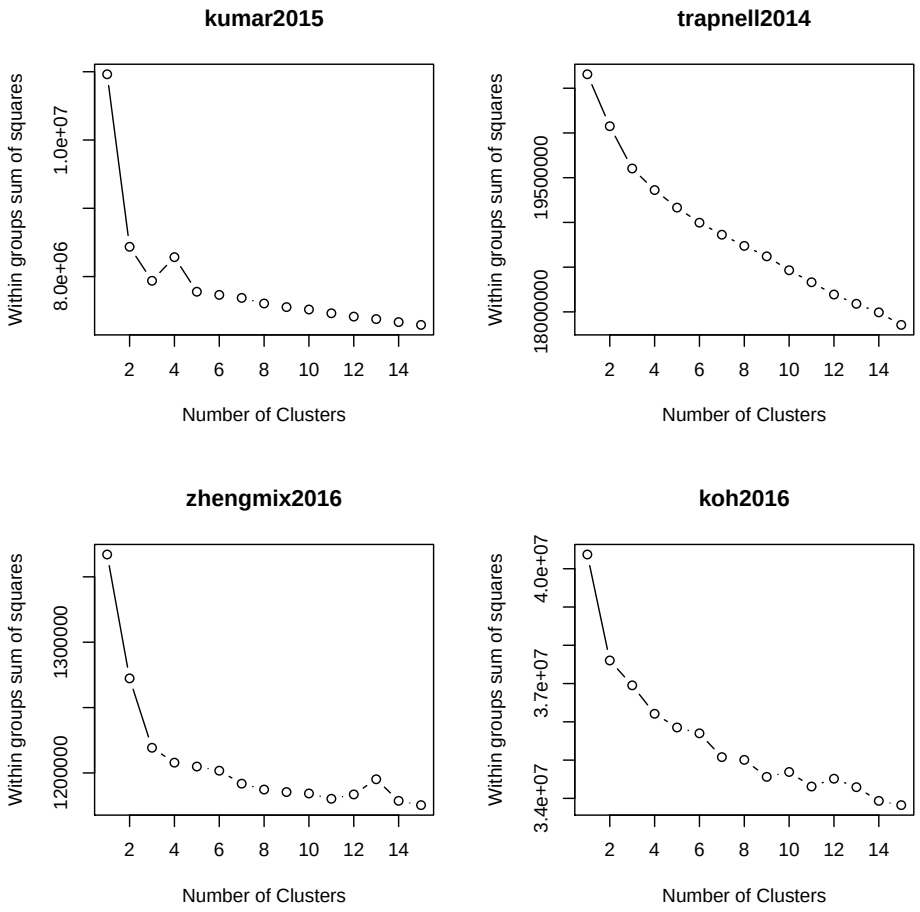

Figure 13: Optimal number of clusters by within sum of squares based on the full dimensions

SNNCliq The connectivity of the quasicliques was set to the default value 0.7 , like wise the merging threshold parameter was set to 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.

SIMLR

9 Clustering

As it is not possible to determine the true subpopulations in the datasets clustering of the dataset was done using a range of parameters. For the semi-supervised methods cidr, pcaReduce, tSNE and kmeans, SC3 and SIMLR a range of number of clusters were used. To compare the results the Adjusted Rand Index (ARI) was computed. For all three datasets the scores were realtively stable. In the Kumar dataset all methods show a maximum with three clusters. CIDR, pcaReduce, tSNE and kmeans show a maximum score with three clusters in the Trapnell dataset. Where as for SIMLR the optimal number of clusters is 4 and 2 for SC3. In the Koh dataset the highest score is with 9 clusters.

Method	Description	dimension reduction	clustering	zero inflation	normalization	unsupervised
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	no	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 on-imputed distances	hierarchical clustering	yes	no	yes
Seurat v1.4	Neirest neighbor graph based on PCA latent space	HVG and PCA		no	yes	yes

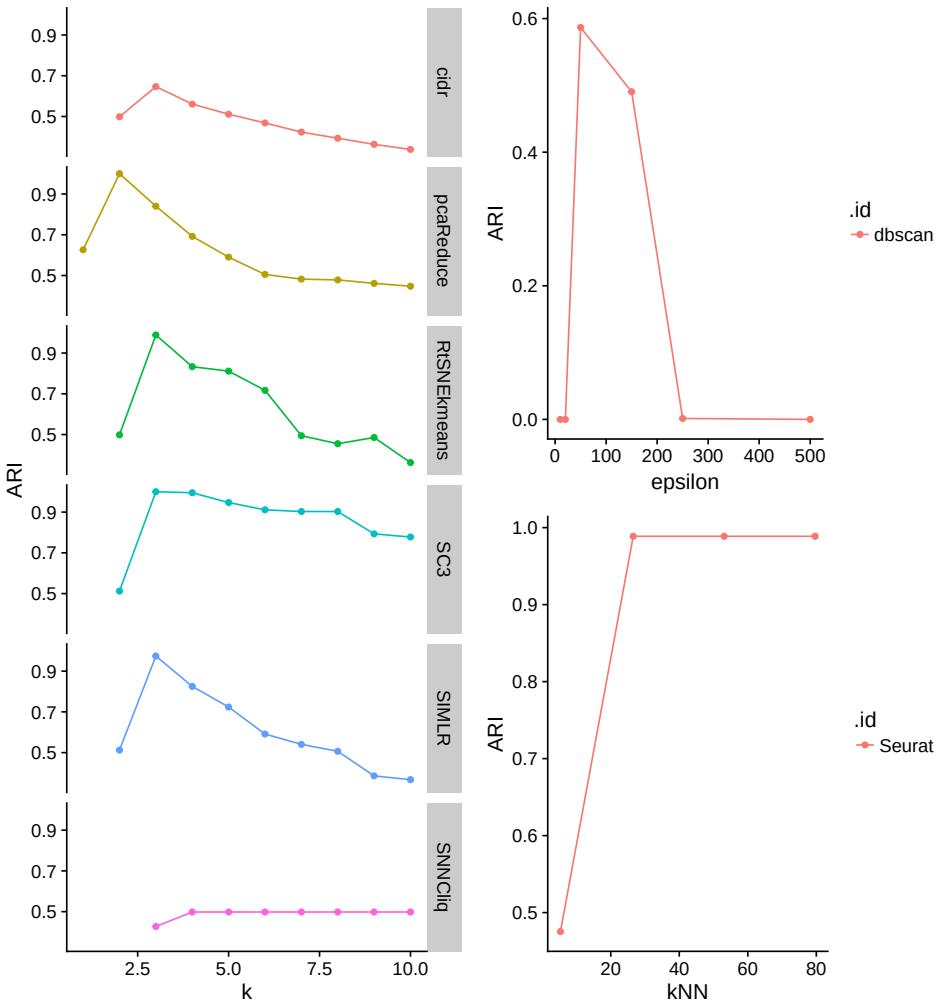

Figure 14: ARI scores for range of parameters, kumar2015 .

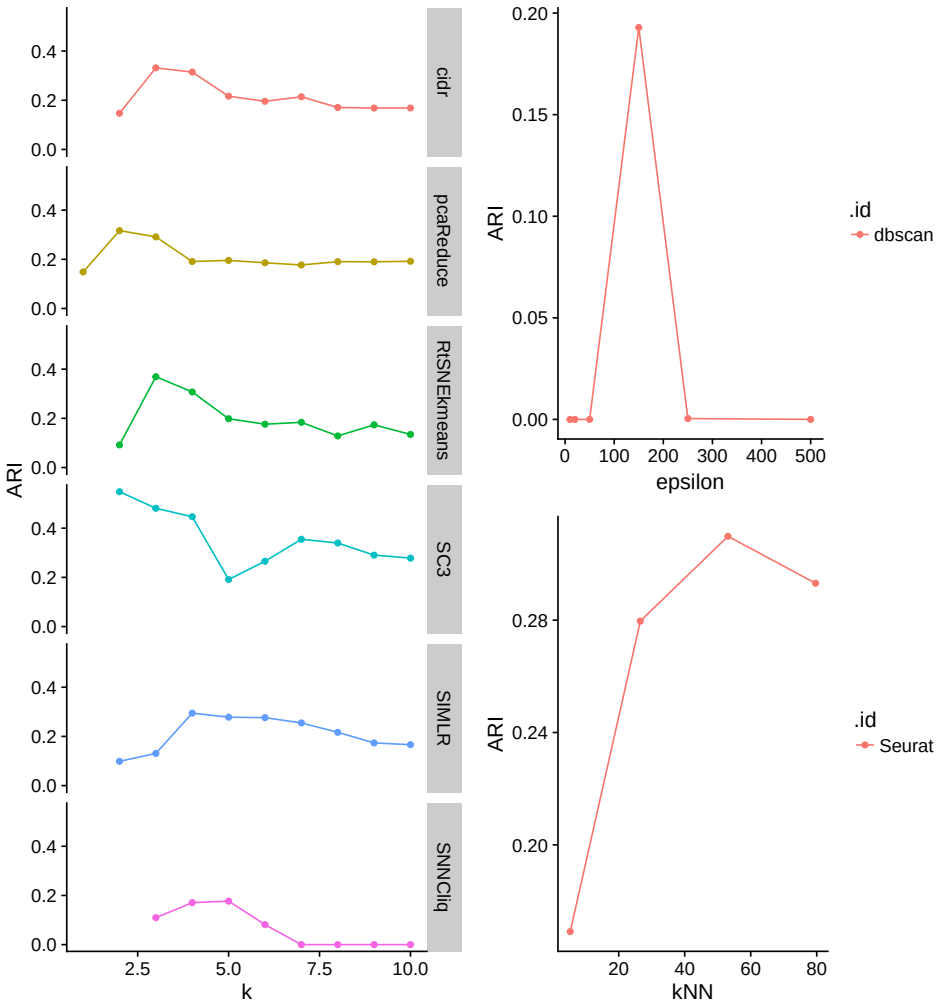

Figure 15: ARI scores for range of parameters,trapnell2014

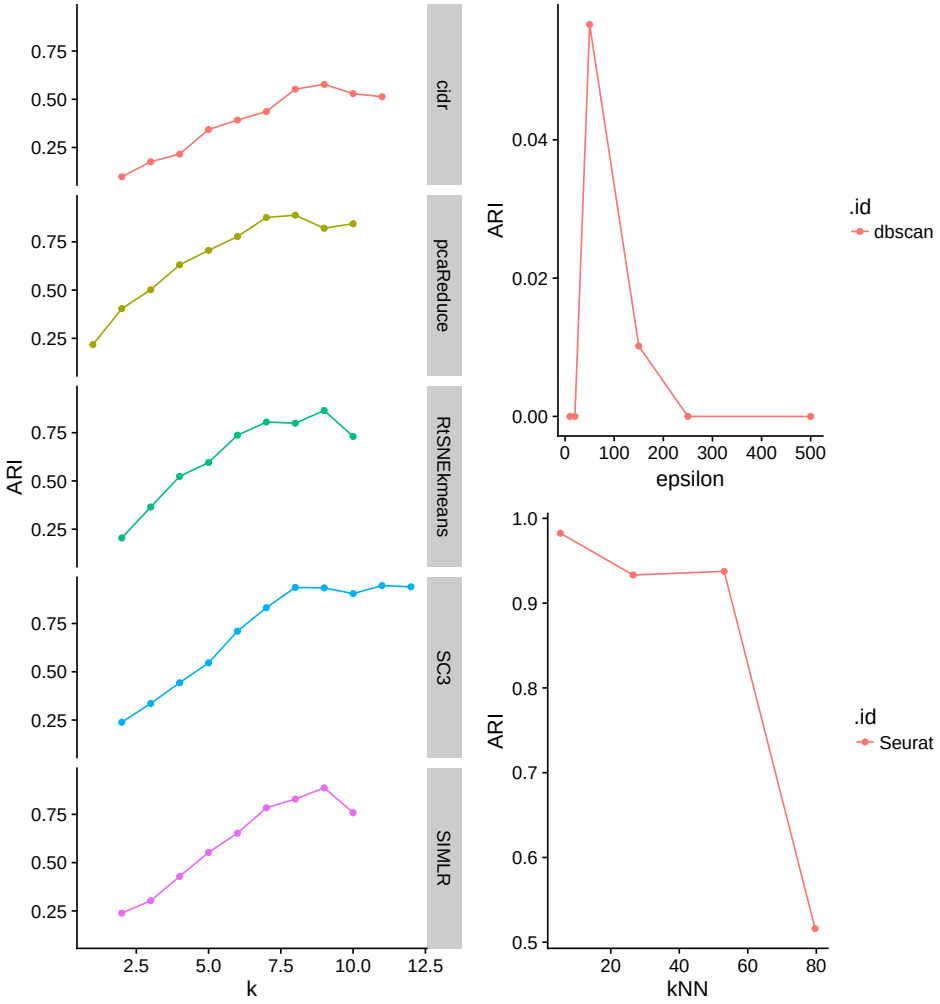

Figure 16: ARI scores for range of parameters,koh2016

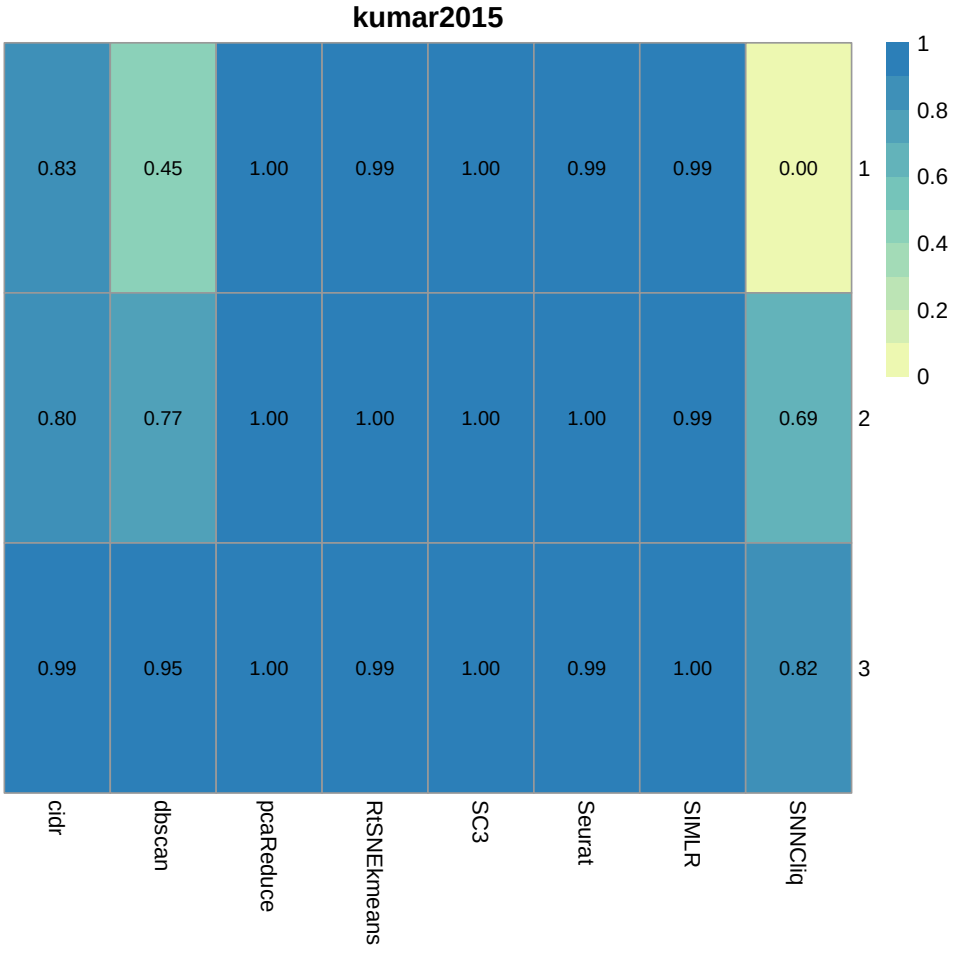

Figure 17: F1 scores for Kumar dataset

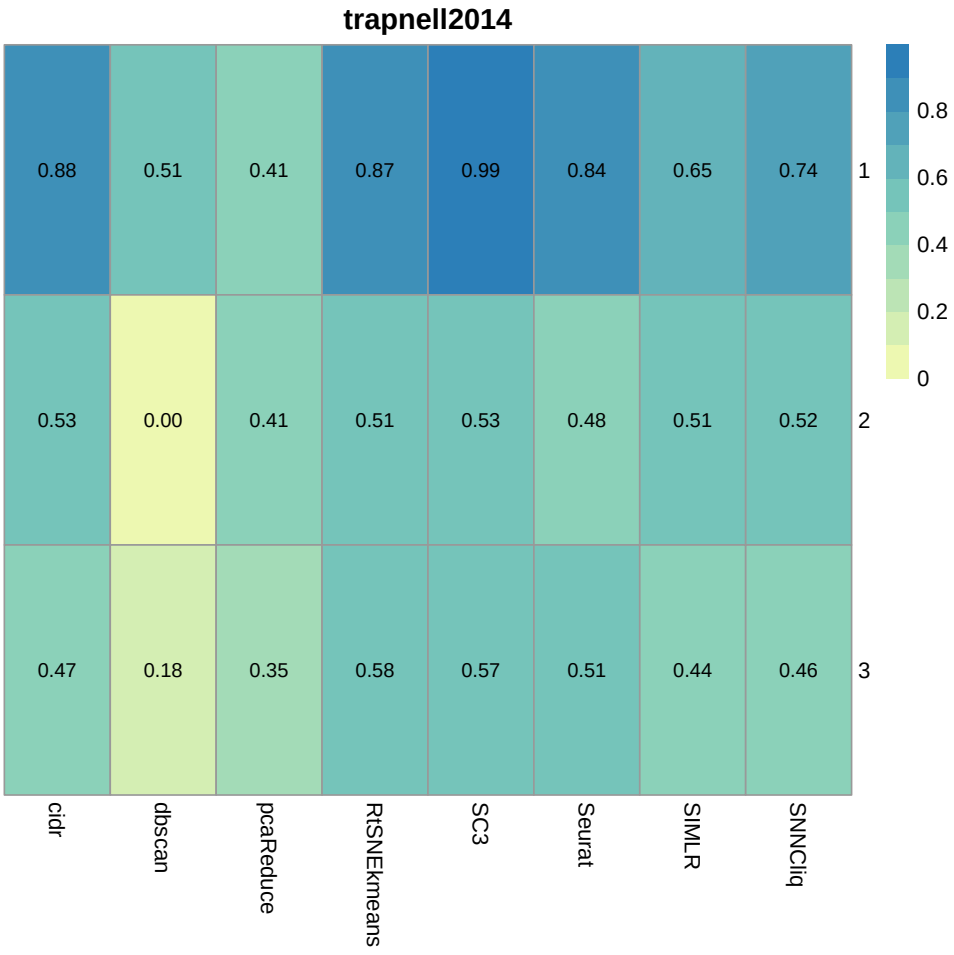

Figure 18: F1 scores for Trapnell dataset

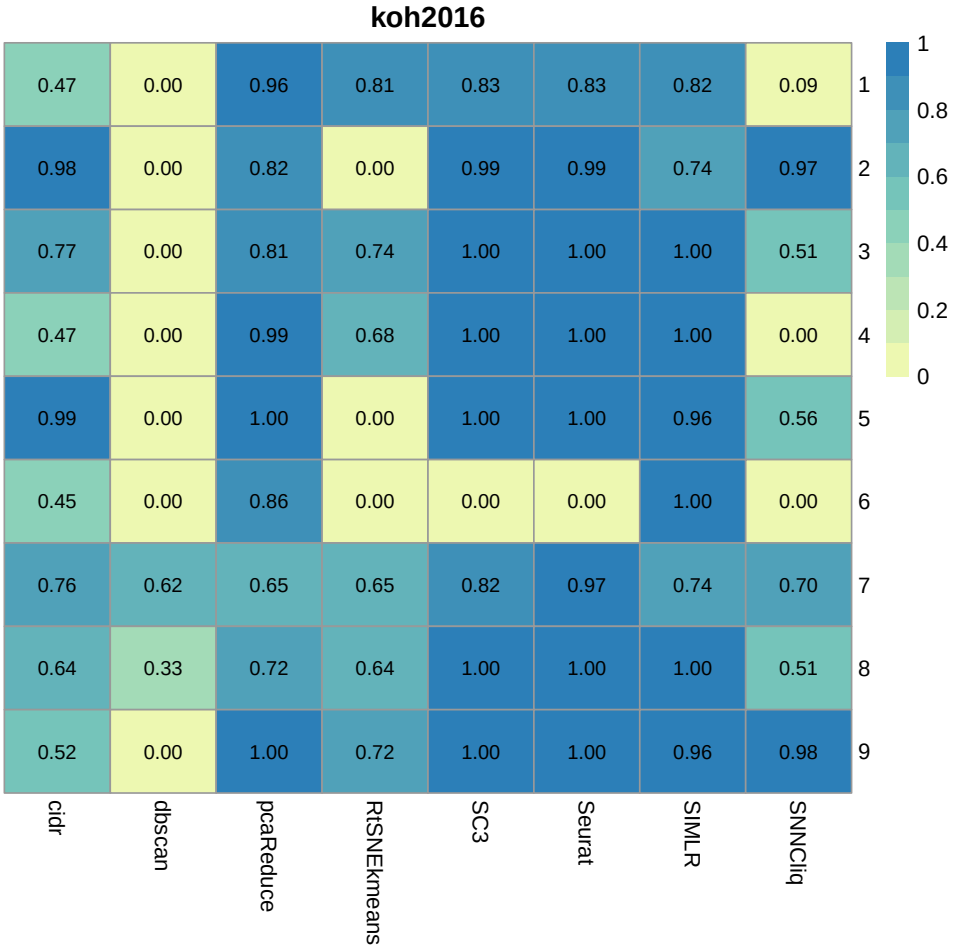

Figure 19: F1 scores for Koh dataset

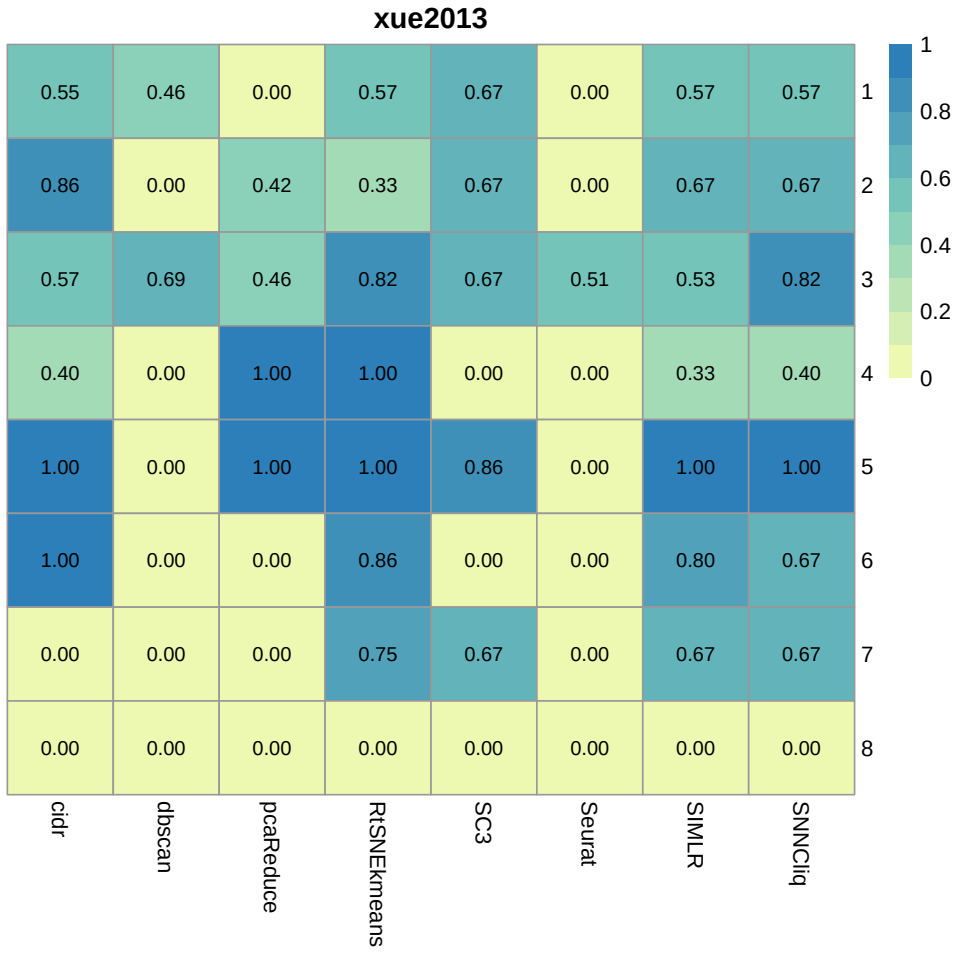

Figure 20: F1 scores for Xue dataset

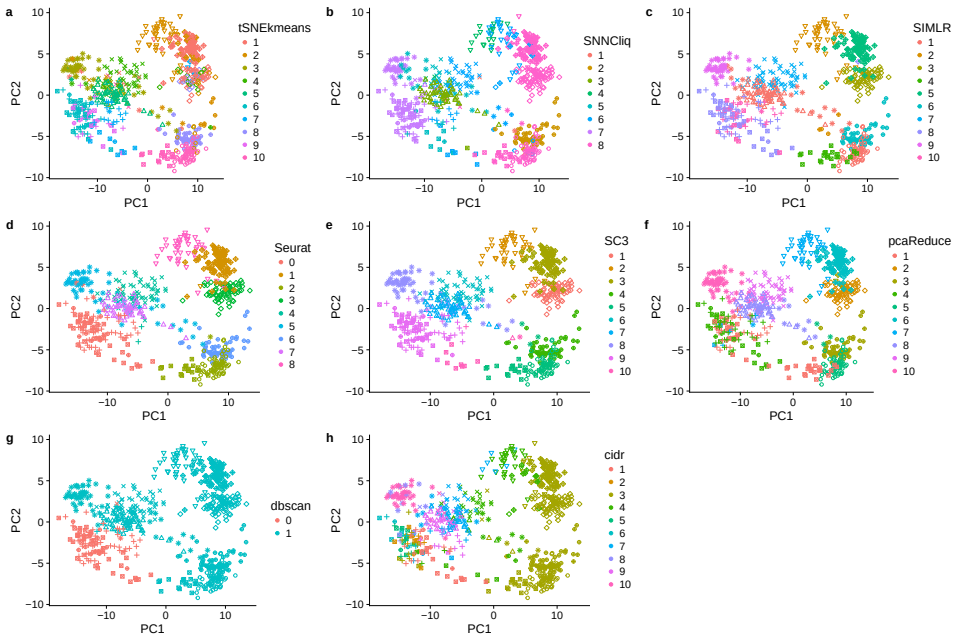

Figure 21: Clusters koh2016 on PC representations.

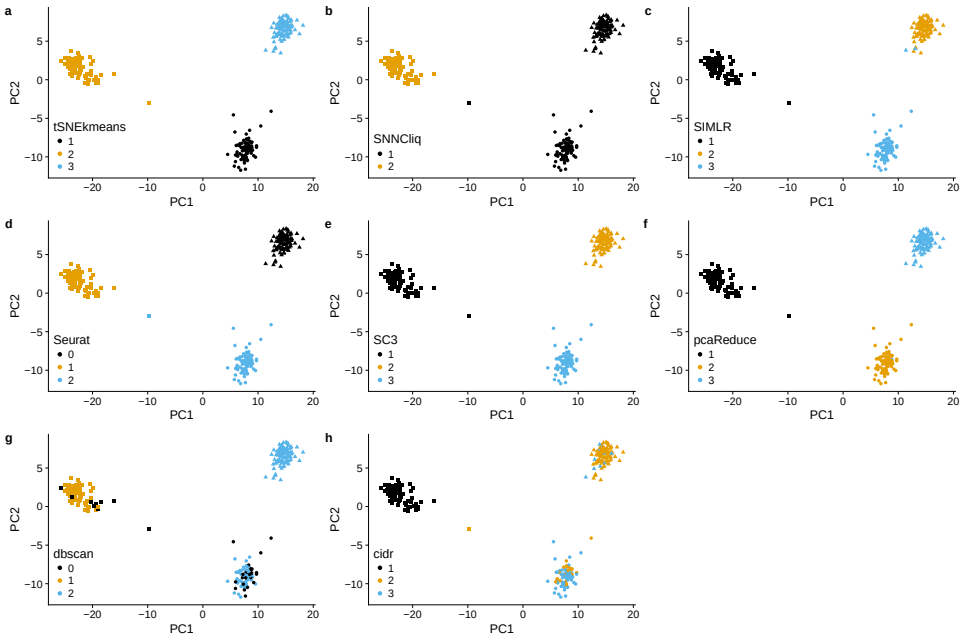

Figure 22: Clusters Kumar 2015 on PC representations.

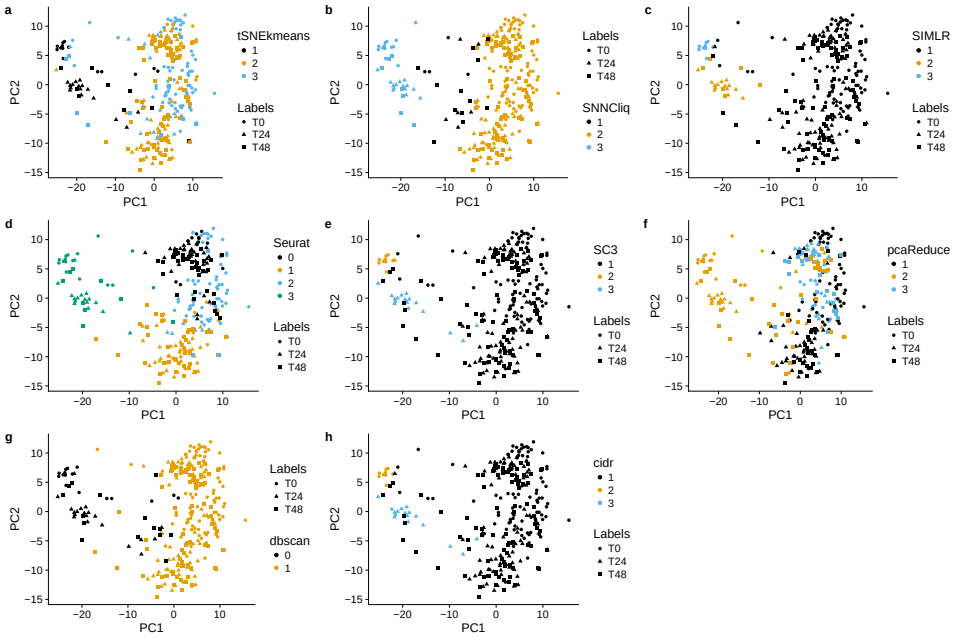

Figure 23: Clusters Trapnell 2014 on PC representations.