1 Datasets

Kumar et al. 2014 Kumar et al. used mouse embryonic stem cells (mESCs) *Dgcr*8-knockout (Novus, NBA1-19349) and V6.5 variotypes. 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 involvemenet in heterogeneity of pluripotent stem cells.

Trapnell et al. 2015 Trapnell et al used human skeletal muscle myoblast cells(Lonza, catalog CC-2580) to investigate temporal differentiation. Cells were expanded under high-mitogen conditions. Differentation is induced by switching to low-serum medium. Cells were captured at the switch 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 timepoint 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: Libraries that contained fewer than 1 million reads were excluded.

Trapnell et al. 2015 H7 human embryonic stem cells were used to study.....

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 extrem values and after transformation the distribution should be more normally distributed. However, log transformations do not adress the problem of heteroscedasticity. Arcus sin transformation should deal with extreme values and equaleize the variances.After transformattion the mean and the variances should be independent. VST adress the problem of extrem values and unequal variances acrosse genes. After such transformation the mean and the variances of the genes should be independent. Using log transformation and vst the mean - varaince dependence is less extrem (see Figure **??**). Still for means in the lower - mid range the variances are not equal.

For the subsequent analysis log2 tansformed length scaled, count scaled transcript per million (TPM) were 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 filterd was used with respect to the total number of genes per cell. 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 filterd out.

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 normilazition and by library size. However, none of these methods are designed to deal specificially with the zero-inflated nature of scRNA-seq data. Another appraoch 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 were then used to estimate size factor. The size factors for the polled cells were then "deconvoluted" into cell-based factors (Lun et al., 2016).

3 Optimal number of clusters

Methods to determine the optimal number of clusters are subjective methods as elbow or sihlhoutte plots. In the Elbow plots the within-cluster sum of square is plotted against a range of clusters. The sihlhoutte plot is a

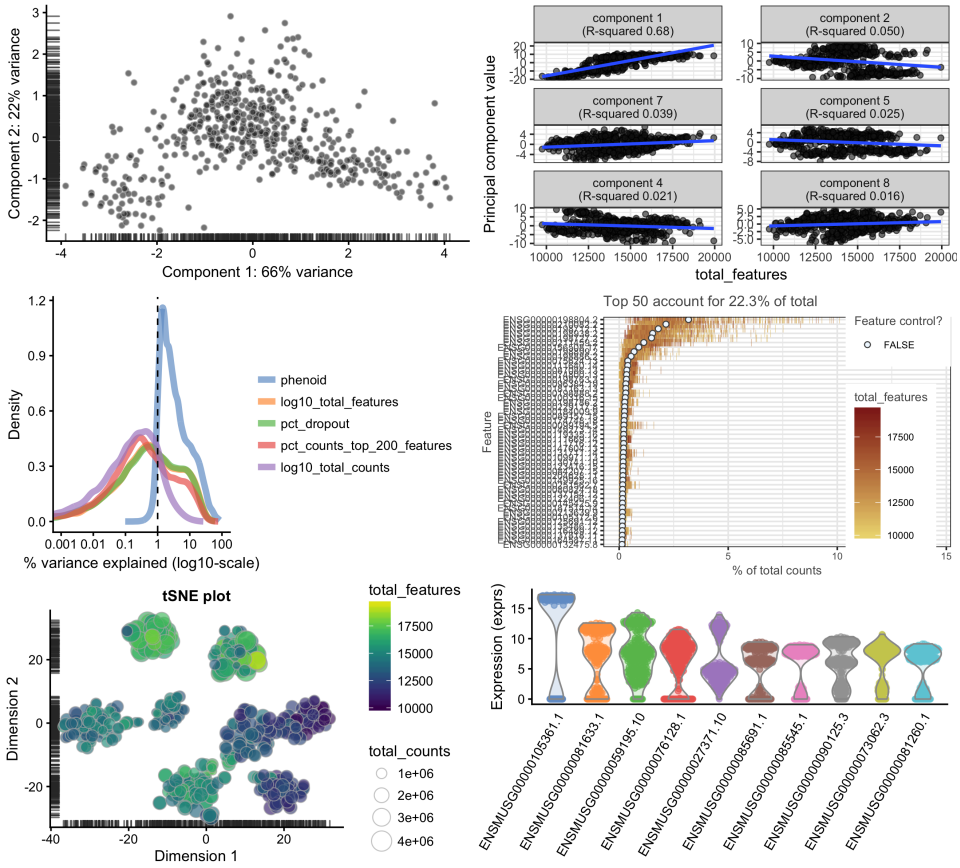

Figure 1: QC summary of Koh 2016.

standartized measures of distances between each point inside and outside their one 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 3 cluster for the Kumar dataset, 2 or 5 in the Trapnell data and 3 in Koh 2016 (see Figure **??**). Based on the plots on the tSNE latent space there 3 in Kumar, 4 in Trapnell and 5 clusters in Koh data.

4 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 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 hierarchical 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 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 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 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. 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.

5 Results