

iScience, Volume 23

Supplemental Information

scID Uses Discriminant Analysis to Identify Transcriptionally Equivalent Cell Types across Single-Cell RNA-Seq Data with Batch Effect

Katerina Boufea, Sohan Seth, and Nizar N. Batada

Supplemental Figures

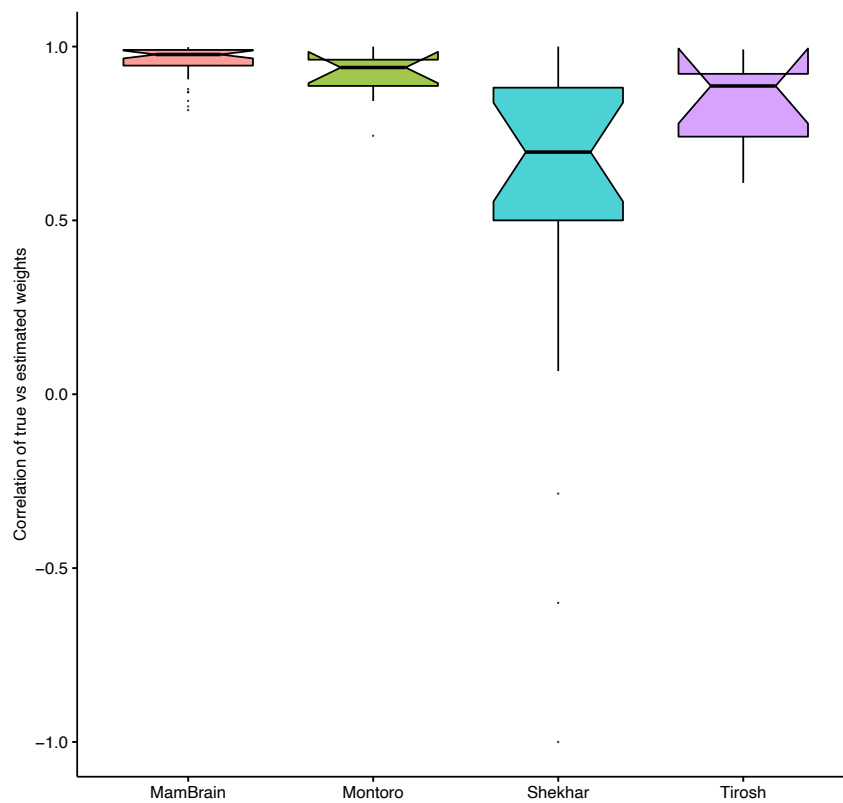


Figure S1. **Assessment of accuracy of the DPR strategy for cell selection, related to Figure 1B.** Y-axis shows Spearman rank correlation of weights estimated from the DPR strategy to weights estimated from the known labels from the reference data. Dataset sources used are listed on the x-axis labels.

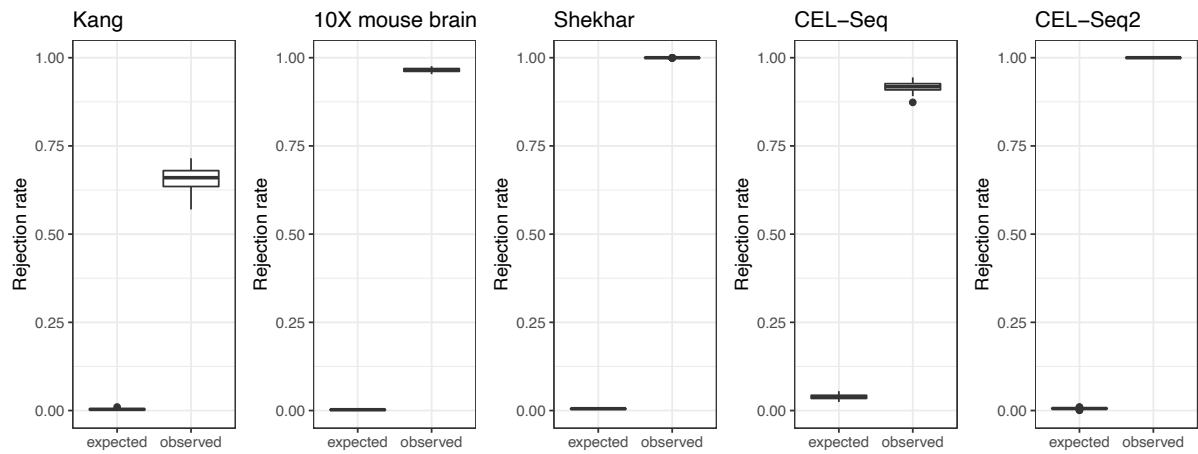


Figure S2. Quantification of batch effect in the reference/target dataset pairs used in the manuscript, related to Figures 1, 2 and 3. Extent of batch effect as measured by kBET (Buttner, Miao et al., 2019) is shown on the y-axis (normalized units). The larger the difference between the observed and expected rejection rate, the bigger the batch effect. Target datasets used are listed in the title of the panel.

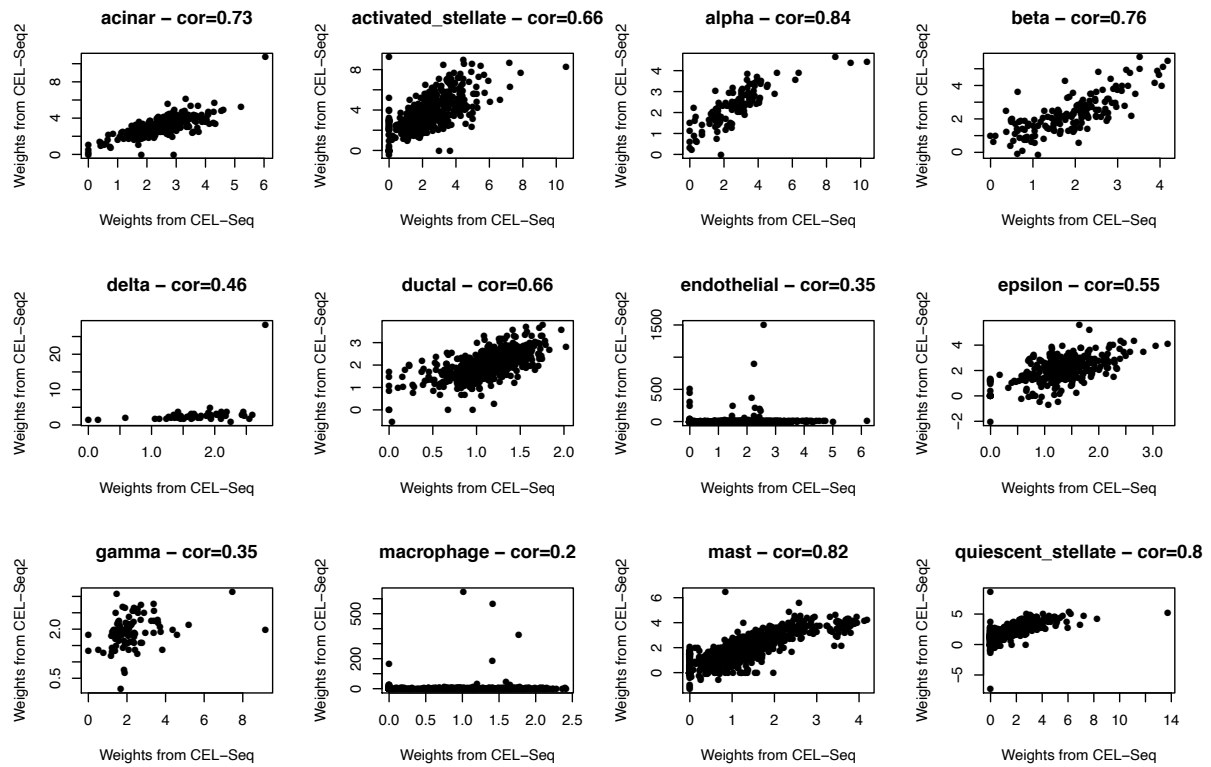


Figure S3. Different target datasets with the same cell types have different weights, related to Figure 1E. Scatter of weights estimated from pancreas scRNA-seq data from CEL-Seq (x-axis) (GSE81076) and weights estimated from pancreas scRNA-seq data from CEL-Seq2 (y-axis) (GSE85241) using the pancreas Smart-seq2 scRNA-seq data (Segerstolpe, Palasantza et al., 2016) (E-MTAB-5061) as reference. For each of the cell types in the reference dataset, weights were computed using DPR positive and negative set selection in the two target datasets. Spearman rank correlation is shown in the title of each panel. Divergence of the correlation from $r=1$ suggests that the weights are not exactly the same in the two datasets for the same cell type.

Transparent Methods

scID workflow

Given a set of $k^C = k_p^C + k_n^C$ features (genes) that are positive (k_p^C) and negative (k_n^C) markers for a reference cluster C , scID selects target cells equivalent to reference cluster C based on their score s_j^C , which is a weighted linear sum given by

$$s_j^C = \frac{\sum_{i=1}^{k^C} w_i^C \tilde{g}_i^j}{\sqrt{\sum_{i=1}^{k^C} w_i^{C^2}}} \quad (\text{Eq. 1})$$

for $j = 1, \dots, n$ where n is the number cells in the target data, $\tilde{g}_i^j \in [0,1]$ is the normalized gene expression value of the i -th gene in the j -th cell, and w_i^C is a weighting factor that represents the discriminative power of gene i to identify target cells equivalent to reference cluster C . To reduce sensitivity to outliers, the gene expression values are normalized to $[0,1]$ by the 99th percentile instead of the maximum i.e. $\tilde{g}_i^j = \min\left(\frac{g_i^j}{P_{99}}, 1\right)$, where g_i^j is the library depth normalized gene expression of gene i in cell j and P_{99} is the 99th percentile of the expression of gene i across all target cells.

The weights can be computed from the reference data as follows

$$w_i^C = \frac{\mu_i^C - \mu_i^{C^-}}{\sigma_i^{C^2} + \sigma_i^{C^-2}} \quad (\text{Eq. 2})$$

where μ_i^C, σ_i^C ($\mu_i^{C^-}, \sigma_i^{C^-2}$) represent the mean and standard deviation, respectively, of gene i in the cluster C (all clusters except C). Each term of the weight of gene i is in turn calculated as follows:

$$\begin{aligned} \mu_i^C &= \frac{1}{l^C} \sum_{j \in C} \tilde{r}_i^j \\ \mu_i^{C^-} &= \frac{1}{N - l^C} \sum_{j \in C^-} \tilde{r}_i^j \\ \sigma_i^{C^2} &= \frac{1}{l^C} \sum_{j \in C} (\tilde{r}_i^j - \mu_i^C)^2 \\ \sigma_i^{C^-2} &= \frac{1}{N - l^C} \sum_{j \in C^-} (\tilde{r}_i^j - \mu_i^{C^-})^2 \end{aligned}$$

where l^C is the number of cells in cluster C , N is the total number of cells in the reference data, and $\tilde{r}_i^j \in [0,1]$ is the normalized gene expression value of the i -th gene in the j -th cell in the reference data.

This definition of the weights (**Eq. 2**) is a solution to Fisher's linear discriminant analysis assuming diagonal covariance, where the separation between the two classes is defined as the ratio of the variance between the classes to the variance within the classes. Ideally, we would like to include the full covariance matrix in order to better capture the relationships among clusters. However, we have chosen to

approximate the covariance matrix as diagonal for computational efficiency but also due to limitations posed by the nature of scRNA-seq data. When datasets have sparse coverage, the covariance matrix is not full rank and cannot be inverted. Additionally, when the list of genes is long, computing the full covariance matrix is error prone due to insufficient number of samples.

Choosing the weights in this way penalizes high variability and low mean expression in order to account for the following cases:

- a. When a positive marker is not expressed uniquely in the C population, μ_i^C will increase, reducing the weight as it does not provide sufficient evidence of belonging to the cells of interest.
- b. When a positive marker is expressed only within a subpopulation of C , $\sigma_i^{C^2}$ will increase, reducing the weight so that a cell's score does not drop sharply when the gene is missing. This also accounts for genes with high dropout rate even though they might be specific and sensitive markers.
- c. Finally, non-discriminative genes will be down-weighted, as the numerator $\mu_i^C - \mu_i^{C^-}$ will be low.

Although we can compute these weights from the reference, computing them from the target data can lead to improved accuracy, due to adjustments to the data quality and cell type composition in the target data. However, to do this we need to select target cells that are likely equivalent to the reference cluster C . The target cells (denoted as c) that express the k_p^C signature genes precisely and specifically and do not express the k_n^C signature genes are selected as equivalent to the reference cluster C by clustering the data in the differential precision–differential recall domain which we refer to as differential precision recall or DPR approach.

In general, precision of a cell expressing a set of genes is defined as the total number of expressed genes that belong to the given set divided by the total number of expressed genes and recall is defined as the number of expressed genes that belong to the set divided by the total number of genes in the set.

So, for the set of positive markers (k_p^C)

$$precision_p = \frac{n_{pme}}{n_e}$$

$$recall_p = \frac{n_{pme}}{n_{pm}}$$

where

n_{pme} : number of upregulated signature genes expressed

n_{pm} : total number of upregulated signature genes

n_e : total number of expressed genes

A cell expressing all positive markers will have $recall_p = 1$ and a cell expressing only positive markers (and no other genes) will have $precision_p = 1$. Thus, cells equivalent to cluster C will be in the first quadrant of the $(precision_p, recall_p)$ space close to $(1,1)$ and other cells not expressing the positive markers will be close to $(0,0)$.

For the set of negative markers (k_n^C)

$$precision_n = \frac{n_{nme}}{n_e}$$

$$recall_n = \frac{n_{nme}}{n_{nm}}$$

where

n_{nme} : total number of downregulated signature genes expressed

n_{nm} : total number of downregulated signature genes

n_e : total number of expressed genes

Similar to precision-recall for positive markers, a cell expressing all negative markers will have $recall_n = 1$ and a cell expressing only negative markers will have $precision_n = 1$. Thus, cells equivalent to cluster C should be in the third quadrant of the $(precision_n, recall_n)$ space close to $(0,0)$.

To identify target cells equivalent to reference cluster C , we have combined the precision and recall for positive and negative markers via a differential precision (DP) and differential recall (DR) metric as follows:

$$DP = precision_m - precision_p = \frac{n_{pme} - n_{nme}}{n_e}$$

$$DR = recall_m - recall_p = \frac{n_{pme}}{n_{pm}} - \frac{n_{nme}}{n_{nm}}$$

Thus, target cells equivalent to cluster C will be close to $(1,1)$, cells very different from cluster C will be close to $(-1,-1)$, and cells belonging to clusters that are similar to cluster C and share markers will be around $(0,0)$. This will help separate cells equivalent to cluster C from cells belonging to very similar clusters that could be otherwise grouped together if we were only using the positive markers.

To select putative positive and negative training populations from the DP-DR space, we cluster the cells using different Gaussian finite mixture models (Scrucca, Fop et al., 2016) and select the model with the lowest Bayesian Information Criterion (BIC). The clusters with highest DP and/or DR are selected as candidate matching clusters. From the candidate clusters, we select as c the one with the closest to $(1,1)$ centroid as measured by Euclidean distance. All other candidate clusters are discarded from the training set and the remaining clusters are used as training non-matching clusters (c^-).

Analogous to computing the weights in the reference data, they can be computed from the target data using the training sets of cells (c and c^-).

$$w_i^c = \frac{\mu_i^c - \mu_i^{c^-}}{\sigma_i^{c^2} + \sigma_i^{c^-2}} \quad (\text{Eq. 2})$$

where μ_i^c, σ_i^c represent the mean and standard deviation, respectively, of gene i in the cluster c and $\mu_i^{c^-}, \sigma_i^{c^-}$ represent the mean and standard deviation, respectively, of gene i in the cluster c^- . Each term of the weight of gene i is in turn calculated as follows:

$$\begin{aligned}\mu_i^c &= \frac{1}{l^c} \sum_{j \in c} \tilde{g}_i^j \\ \mu_i^{c^-} &= \frac{1}{l^{c^-}} \sum_{j \in c^-} \tilde{g}_i^j \\ \sigma_i^{c^2} &= \frac{1}{l^c} \sum_{j \in c} (\tilde{g}_i^j - \mu_i^c)^2 \\ \sigma_i^{c^-2} &= \frac{1}{l^{c^-}} \sum_{j \in c^-} (\tilde{g}_i^j - \mu_i^{c^-})^2\end{aligned}$$

where l^c, l^{c^-} are the number of cells in clusters c and c^- respectively.

The weights estimated from the target data reflect the cellular composition of the target data and quality (i.e. the distribution of dropout and dynamic range of gene expression) of the target data both of which influence the discriminatory power of the features.

Then, to identify target cells equivalent to reference cluster C , we fit different finite mixtures of Gaussians on the score s_j^C (**Eq. 1**) and assign cells that belong to the population with highest average score to reference cluster C .

When the reference data contains clusters that are highly similar transcriptionally, the features from a cluster can be correlated with the features from another cluster which in turn lead scID to assign target cells to multiple reference classes. To resolve this, the scores of target cells s_j^C is first z-score normalized and the ambiguous cell is assigned to the reference cluster with the highest normalized score over all other reference clusters it was assigned to.

Data source

Human Metastatic Melanoma immune cells from Tirosh et al. 2016 (Tirosh, Izar et al., 2016). This Smart-seq2 (Picelli, Faridani et al., 2014) data consists of malignant, immune and stromal cells from metastatic melanoma tumours from 19 patients, a total of 4,645 cells. We have used the 3,254 immune (CD45⁺) cells for our analysis, which on average had 3,925 genes per cell. Data was downloaded from the Broad Institute Single Cell Portal.

Mouse Retinal Bipolar Neurons from Shekhar et al. 2016 (Shekhar, Lapan et al., 2016): This study performed Drop-seq and Smart-seq2 experiments on Vsx2-GFP mouse retinal cells. The Drop-seq data had 27,499 cells with an average of 880 genes per cell. The Smart-seq2 data had 288 cells with an average of 4,556 genes per cell. Gene expression data was downloaded from the Broad Institute Single Cell Portal.

Brain cells from E18 mouse: This 10X data consists of brain cells from the cortex, hippocampus and subventricular zone of an E18 mouse. The scRNA-seq dataset had 9,128 cells with an average of ~2,500 genes per cell and the single nuclei RNA-seq data have 954 cells with an average of 2,832 genes per cell. Both these datasets

were downloaded from 10X Genomics (<https://support.10xgenomics.com/single-cell-gene-expression/datasets>).

Murine tracheal epithelium cells from Montoro et al. 2018 (Montoro, Haber et al., 2018): This combination of plate-based and droplet-based scRNA-seq data of murine airway epithelial cells consists of 7,193 cells with an average of 1,712 genes per cell. The cells were partitioned into seven clusters annotated post hoc using a biomarker approach by the authors. Data was downloaded from the Broad Institute Single Cell Portal.

Mouse brain cells from Hu et al. 2017 (Hu, Fabyanic et al., 2017): This Drop-seq single nuclei RNA-seq data from cortical tissues of adult mice consists of 18,194 cells with 1,649 genes per cell on average that were partitioned into 40 annotated clusters. Data was downloaded from the Broad Institute Single Cell Portal.

Unstimulated and stimulated PBMCs from Kang et al. 2018 (Kang, Subramaniam et al., 2018): This 10X data consists of 14039 human PBMCs from eight patients, split into two groups; one control and one stimulated with interferon-beta (IFN- β). Seurat CCA was used to align and cluster the data in order to obtain gold standard cell identities as shown in Butler et al (Butler, Hoffman et al., 2018).

Human pancreatic islet cells from Segerstolpe et al. 2016 (Segerstolpe et al., 2016): This Smart-Seq2 data consists of pancreatic tissue and islets from six healthy individuals and four type 2 diabetes patients. RPKM-normalized gene expression data and cell labels were downloaded from ArrayExpress (E-MTAB-5061).

Human pancreatic islet cells from Grün et al. 2016 (Grun, Muraro et al., 2016): This CEL-seq data consists of pancreatic cells from deceased organ donors with and without type 2 diabetes. Gene expression data and cell labels were downloaded from NCBI GEO (GSE81076).

Human pancreatic islet cells from Muraro et al. 2016 (Muraro, Dharmadhikari et al., 2016): This CEL-seq2 data consists of islets from cadaveric pancreas. Gene expression data and cell labels were downloaded from NCBI GEO (GSE85241).

Data Normalization

When datasets were obtained as UMI counts, Counts Per Million (CPM) library-depth normalization was performed prior to the analysis. For the biomarker-based approach we further normalized the gene expression to [0,1] by the 99th percentile in order to use the same threshold between marker genes that can differ significantly in their expression level.

scID implementation

The R implementation and tutorial for scID is available on Github (<https://batadalab.github.io/scID/>).

scID has the following user-specified options:

1. \log_{FC} : Log-fold-change threshold for extracting cluster-specific genesets from the reference data. The \log_{FC} used for extracting gene signatures for the reference datasets used in the figures are as follows: For Hu et al. 2017

(Hu et al., 2017) \log_{FC} was set to 0.3; for Montoro et al. 2018 mouse tracheal epithelium data (Montoro et al., 2018) \log_{FC} was set to 0.5; for Shekhar et al. 2016 (Shekhar et al., 2016) \log_{FC} was set to 0.7; for Tirosh et al. 2016 (Tirosh et al., 2016) metastatic melanoma Smart-seq2 data \log_{FC} was set to 0.5; for Segerstolpe et al. 2016 (Segerstolpe et al., 2016) Smart-seq2 human pancreas data \log_{FC} was set to 0.5; for 10X E18 mammalian brain data (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/neuron_1k_v2) \log_{FC} was set to 0.6.

2. `estimate_weights_from_target`: Estimate weights using the target data by selecting training sets using the precision-recall-like approach. For all figures in the paper we set this to `TRUE`.
3. `only_pos`: Select only upregulated genes from each reference cluster. For all figures in the paper we set this to `FALSE` in order to select both upregulated and downregulated genes.

Software tools and parameter settings

We used the following R packages and parameters:

Seurat (Buttner et al., 2019) (version 3.0.1)

We used UMI count data when available and followed the standard Seurat workflow for clustering and data integration with default settings. When normalized data was provided instead, we skipped the `NormalizeData()` function of the workflow.

scran (version 1.10.2)

For running MNN (Haghverdi, Lun et al., 2018) we used log-transformed CPM-normalized gene expression values. Aligned and merged expression matrixes were clustered using Seurat with default parameters.

scmap (Kiselev, Yiu et al., 2018) (version 1.4.1)

We used log-transformed CPM-normalized gene expression values for both reference and target data. We selected number of highly variable genes to be used so that the maximum possible number of target cells can be classified. Specifically, for the data pairs of Figure 1E and Figure 3 we used 500 highly variable genes and for the data pairs of Figure 2 we used 150.

CaSTLe (Lieberman, Rokach et al., 2018)

For running CaSTLe we used the code provided in <https://github.com/yuvallb/CaSTLe/blob/master/CaSTLeMultiClass.R> with all predefined parameters. We used log-transformed CPM-normalized gene expression values for both reference and target data.

Biomarker-based classification of cells

To assign labels to target cells using the biomarker-based approach we first extracted the top two highly enriched markers (referred to as biomarkers) from each reference cluster, sorted by average log fold change. Then binarized the 99th-percentile-normalized gene expression data using different thresholds (0.10, 0.25, 0.50 and 0.75) and checked which marker genes are present in each cell. Cells that expressed biomarkers of different cell types were labelled as “ambiguous”. Cells that expressed biomarkers of a single cell type were assigned to the respective cell type

and cells that did not express any biomarker were labelled as “orphans”. Only the uniquely classified cells were used to assess the performance of other methods.

Quantification of batch effect between pairs of scRNA-seq data

To measure the extent of batch effect between the reference-target pairs of data used in the manuscript, we used kBET (Buttner et al., 2019)(version 0.99.5). High rejection rate indicates poor mixing of the data.

Statistical tests

For testing the improvement of scID Stage 3 versus scID Stage 2 (**Figure 1C**), we tested the difference in True Positive (TPR) and False Positive Rate (FPR) between the two stages for each reference cluster using two-sided paired Kruskal-Wallis test. For the comparison of the various methods classification of cells to their gold standard labels we have used the Adjusted Rand Index (ARI) and the Variation of Information (VI) metrics. High similarity between the testing method and the true classification results in high ARI and low VI values.

Supplemental References

- Butler A, Hoffman P, Smibert P, Papalexi E, Satija R (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 36: 411-420
- Buttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ (2019) A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods* 16: 43-49
- Grun D, Muraro MJ, Boisset JC, Wiebrands K, Lyubimova A, Dharmadhikari G, van den Born M, van Es J, Jansen E, Clevers H, de Koning EJP, van Oudenaarden A (2016) De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell* 19: 266-277
- Haghverdi L, Lun ATL, Morgan MD, Marioni JC (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 36: 421-427
- Hu P, Fabyanic E, Kwon DY, Tang S, Zhou Z, Wu H (2017) Dissecting Cell-Type Composition and Activity-Dependent Transcriptional State in Mammalian Brains by Massively Parallel Single-Nucleus RNA-Seq. *Mol Cell* 68: 1006-1015 e7
- Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, Wan E, Wong S, Byrnes L, Lanata CM, Gate RE, Mostafavi S, Marson A, Zaitlen N, Criswell LA, Ye CJ (2018) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* 36: 89-94
- Kiselev VY, Yiu A, Hemberg M (2018) scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods* 15: 359-362
- Lieberman Y, Rokach L, Shay T (2018) CaSTLe - Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS One* 13: e0205499
- Montoro DT, Haber AL, Biton M, Vinarsky V, Lin B, Birket SE, Yuan F, Chen S, Leung HM, Villoria J, Rogel N, Burgin G, Tsankov AM, Waghray A, Slyper M, Waldman J, Nguyen L, Dionne D, Rozenblatt-Rosen O, Tata PR et al. (2018) A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* 560: 319-324

Muraro MJ, Dharmadhikari G, Grun D, Groen N, Dielen T, Jansen E, van Gurp L, Engelse MA, Carlotti F, de Koning EJ, van Oudenaarden A (2016) A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* 3: 385-394 e3

Picelli S, Faridani OR, Bjorklund AK, Winberg G, Sagasser S, Sandberg R (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* 9: 171-81

Scrucca L, Fop M, Murphy TB, Raftery AE (2016) mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J* 8: 289-317

Segerstolpe A, Palasantza A, Eliasson P, Andersson EM, Andreasson AC, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK, Smith DM, Kasper M, Ammala C, Sandberg R (2016) Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab* 24: 593-607

Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M, Adiconis X, Levin JZ, Nemesh J, Goldman M, McCarroll SA, Cepko CL, Regev A, Sanes JR (2016) Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* 166: 1308-1323 e30

Tirosh I, Izar B, Prakadan SM, Wadsworth MH, 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, Fallahi-Sichani M, Dutton-Regester K, Lin JR, Cohen O, Shah P, Lu D, Genshaft AS, Hughes TK, Ziegler CG, Kazer SW et al. (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352: 189-96