scID: Identification of equivalent transcriptional cell populations across single cell RNA-seq data using discriminant analysis

*Katerina Boufea, Sohan Seth and Nizar N. Batada**

Institute for Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK EH4 2XU
School of Informatics, University of Edinburgh, Edinburgh, UK Eh8 9AB

* Correspondence should be addressed to N.B. (nizar.batada@igmm.ed.ac.uk)

Abstract

The power of single cell RNA sequencing (scRNA-seq) stems from its ability to uncover cell type-dependent phenotypes, which rests on the accuracy of cell type identification. However, resolving cell types within and, thus, comparison of scRNA-seq data across conditions is challenging due to technical factors such as sparsity, low number of cells and batch effect. To address these challenges we developed scID (*S*ingle *C*ell *ID*entification), which uses the framework of Fisher's Linear Discriminant Analysis to identify transcriptionally related cell types between scRNA-seq datasets. We demonstrate the accuracy and performance of scID relative to existing methods on several published datasets. By increasing power to identify transcriptionally similar cell types across datasets, scID enhances investigator's ability to extract biological insights from scRNA-seq data.

Introduction

Single cell RNA-sequencing (scRNA-seq) (Hashimshony et al., 2012; Jaitin et al., 2014; Macosko et al., 2015; Picelli et al., 2014; Ramskold et al., 2012; Zheng et al., 2017) is now being routinely used to characterize cell type specific changes in development, disease and perturbations (Regev et al., 2017) which require analysis of cross-data comparison(Butler et al., 2018; Haghverdi et al., 2018; Kiselev et al., 2018). However, there are technical and biological variations between the different datasets which may hamper the joint analysis of these data (Butler et al., 2018; Haghverdi et al., 2018; Kiselev et al., 2019).

Batch correction approaches such as CCA (Butler et al., 2018) and MNN (Haghverdi et al., 2018) have been designed to integrate scRNA-seq datasets. The advantage of these methods is that batch correction allows merging of datasets which provide more power to resolve cell types (Heimberg et al., 2016), particularly those that are rare. In both CCA and MNN, aligning cells is based on preserving the local correlation or neighbourhood structure of the data. However, when the data are significantly imbalanced in the number of cells and library size per cell, both of which influences accuracy of clustering, it is unclear how these methods perform. Moreover, depending on the extent of technical bias or extent of overlap in cell types between the data, batch correction between data may not be possible and runs the risk of over-correcting i.e. the biological variance between the data may be lost.

As deeply characterized, extensively validated and annotated tissue, organ and organism level atlases are being generated (Regev et al., 2017), it is worthwhile to reuse such high-quality information to identify known populations of cells in the target data. In addition to the propagation of cell labels and metadata across datasets, reference data can also be used to group cells in the target data in which the unsupervised clustering does not resolve clusters consistent with literature or with reference data. In such a case, a biomarker-based approach for identifying transcriptionally equivalent cells across datasets involves determining whether the top marker genes derived from the reference clusters are present in the target cells. However, this biomarker-based approach does not work effectively due to the dropout problem with scRNA-seq data i.e. for a large proportion of genes, the expression values are not sampled (Bacher and Kendziorski, 2016; Stegle et al., 2015; Vallejos et al., 2017). To map cells from target scRNA-seq data to a reference data, scmap (Kiselev et al., 2018) extracts features from each reference cluster and uses a distance and correlation based metric to quantify the closeness between each reference cluster's centroid and the cell in the target data to assign target cells to reference clusters. However, distance metrics in high dimensional space may not work as expected (Aggarwal, 2001).

Here we present a novel method called scID for identifying transcriptionally related groups of cells across datasets. Rather than extracting features from the reference dataset, scID extracts a list of cluster-specific genes from the reference dataset, and weighs their relevance in the target dataset by learning a classifier with respect to a putative population of cells either expressing or not expressing those genes. Through extensive analysis of published datasets with strong asymmetry in cell number and library sizes per cell in which the independent clustering of the target data via unsupervised method is not obviously similar to that in the reference, we show that scID is more accurate than existing methods. Given that scRNA-seq datasets to be compared can vary broadly in quality (i.e. different number of cells, dropout levels and dynamic range of gene expression), mapping approach, such as scID, can be more suitable than batch correction in certain cases and thus help uncover hidden biological variation present within the datasets.

Results

Identification of transcriptionally similar cells across scRNA-seq datasets

Identifying target cells that matches a reference cluster can be seen as a classification problem by treating the reference cluster-of-interest as one class (c) and the rest of the reference cells as another class (c^-). Assuming linear separation, one can for example, use Fisher's linear discriminant analysis to learn a direction, $\boldsymbol{w^c} = (w_1^c, w_2^c, \ldots, w_k^c)$, that maximises the separation between c and c^- population where k is the number of genes or features. Thus, to each target cell we can assign an 'scID score'

$$s_j^c = \frac{\sum_{i=1}^{k} w_i^c \, \tilde{g}_i^j}{\sum_{i=1}^{k} w_i^c} \qquad \textbf{(Eq. 1)}$$

where $\tilde{g}_i^j$ is the normalized gene expression of feature i in cell j and w_i^c is the weight that reflects the discriminatory power of the i-th gene, such that in the projected space, the mean expression or centroids of the c and the c^- populations are maximally separated. The weights are given as

$$w_i^c = \frac{\mu_i^c - \mu_i^{c^-}}{\sigma_i^{c2} + \sigma_i^{c^-2}} \qquad \textbf{(Eq. 2)}$$

where μ_i^c ($\mu_i^{c^-}$) represents mean expression of the cluster c (c^-) and σ_i^c ($\sigma_i^{c^-}$) represents standard deviation of gene expression within the cluster c (c^-). A target cell can then be assigned to a reference cluster by choosing a suitable threshold.

scID maps cells in the target data that are equivalent to clusters in the reference data in three key stages (**Figure 1A**). In the first stage, genes that are differentially upregulated in each cluster (herein referred to as gene signatures) are extracted from each cluster of the reference data. Instead of using all the genes or highly variable genes to compute feature weights, scID uses only those genes that are highly upregulated in each cluster to speed up to computation. In general, the weights of features extracted from the reference may not be discriminative in the target due to significantly different technical features, biological treatment or cellular composition. Thus, in the second stage, for each reference cluster, the weights of gene signatures (which reflect their discriminability) are estimated from the target data by identifying putative c and c^- cells. In the last stage, the extracted features (i.e. the combination of gene signatures and their weights) are used to identify target cells that are most concordant with

extracted features. A mixture of two Gaussians is used to model the scores of the c and c^- cells. Cells with high likelihood of falling under the distribution that is closer to 1 are putatively assigned to the reference cluster from which the gene signature was extracted. Ideally a target cell should be assigned to only one reference cluster i.e. it should be in the c^- populations for the rest of the reference clusters, but when closely related cells are present in the reference data, it is possible that the features extracted are not able to highly discriminate between them leading to assignment of a cell to multiple clusters in the reference. Conflicts in assignments are resolved using a relative normalized score strategy. Further details are provided below and in the **Methods** section.

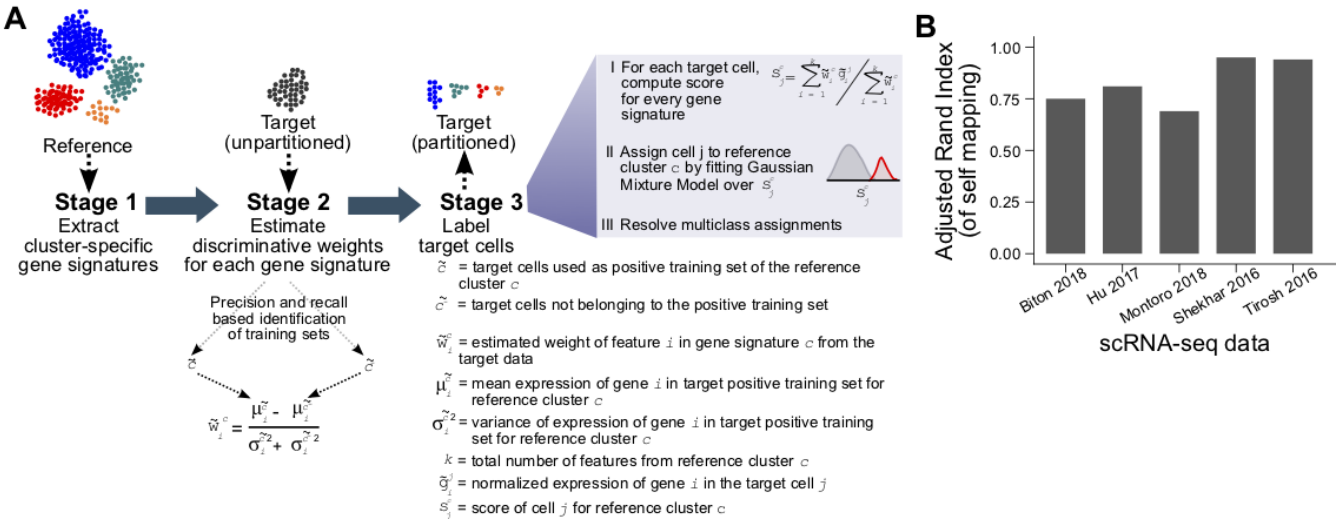

Figure 1. Overview and assessment of scID. (A) *The three main stages involved in mapping cells across scRNA-seq data with scID are as follows: In stage 1, gene signatures are extracted from the reference data (shown as clustered groups on a reduced dimension). For every reference cluster (represented by a different colour), the list of genes that are significantly enriched are extracted using MAST (Finak et al., 2015). In stage 2, discriminative weights are estimated from the target data for each reference cluster-specific gene signature. The combination of the gene signature and weights of the signature genes serves as the feature for scID. In stage 3, every target cell is scored for each feature and is assigned to the corresponding reference cluster if this score is sufficiently high. When necessary, conflicting assignments are resolved.* ***(B)*** *Assessment of accuracy scID. The indicated data (x-axis labels) were self-mapped by scID and the assigned labels were compared to the labels obtained from unsupervised clustering. Adjusted Rank Index is a commonly used measure of accuracy in cluster assignment.*

Through self-mapping (i.e. using same dataset as reference and target) assessment from 5 published datasets, we confirmed that scID has good overall accuracy (**Figure 1B**). Below, we assess the accuracy of scID in identifying equivalent cells across data relative to existing methods.

Reference based identification of equivalent cells by scID across the mouse retinal bipolar neurons scRNA-seq datasets.

A large number of cells despite shallow coverage per cell provide more power to resolve highly similar cells than low number of cells with deep coverage per cell (Kiselev et al., 2019). This is well illustrated by the mouse retinal bipolar neuronal Drop-seq (Macosko et al., 2015)-based scRNA-seq data (Shekhar et al., 2016) which has highly similar cell types. In that study (Shekhar et al., 2016), scRNA-seq data were generated from mouse retinal bipolar neurons using two different technologies: Drop-seq (Macosko et al., 2015) based had ~26,800 cells (with ~700 genes per cell) and the plate-based Smart-seq2 data had 288 cells (~5000 genes per cell).

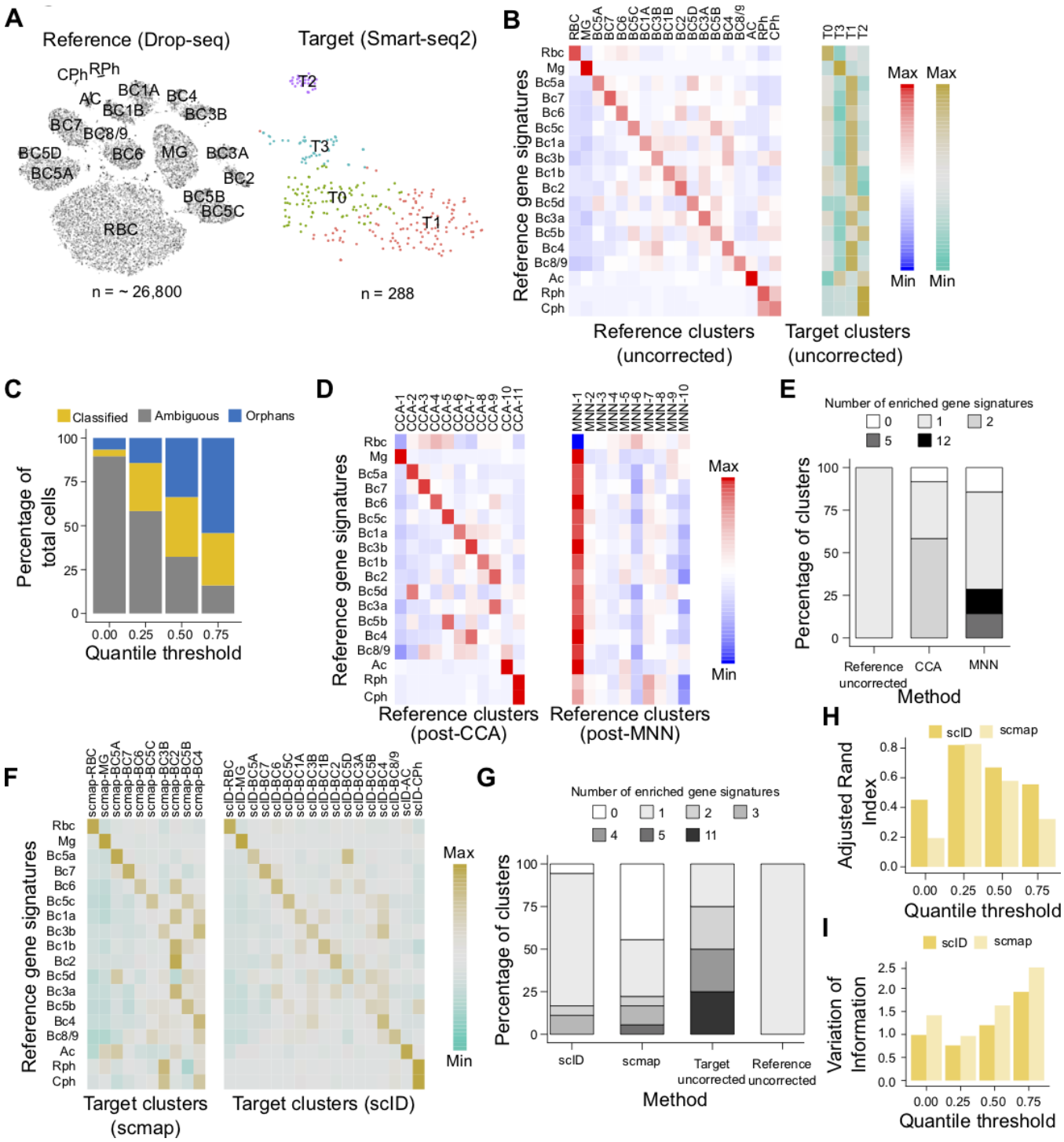

Figure 2. Reference based identification of equivalent cells across the mouse retinal bipolar neurons scRNA-seq datasets. *(**A**) t-SNE plot showing clusters in Drop-seq (reference) and Smart-seq2 (target) data of mouse retinal bipolar cells from Shekhar et al(Shekhar et al., 2016). Cluster membership of Drop-seq data (~26,800 cells) were taken from the publication. Smart-seq2 data (288 cells) were clustered using Seurat(Butler et al., 2018) and clusters name were assigned arbitrarily. (**B**) Heatmap showing z-score normalized average expression of gene signatures in the reference Drop-seq data (left) and in the target Smart-seq2 data (right). Red (khakhi) indicates enrichment and blue (turquois) indicates depletion of the gene signature levels relative to average expression of that gene signature across all clusters of reference (target) data. Row labels, shown in title case to distinguish it from cluster identity, indicate names of the gene signatures. Column names indicate cluster identity. The names of the Drop-seq data are as defined in the original publication(Shekhar et al., 2016). The names for the Smart-seq2 data are arbitrary names given to the clusters identified by Seurat. Presence of multiple dark red (khaki) pixels in a single column reflects merging of separate reference clusters; presence of multiple dark red (khaki) pixels in a single*

*row reflects reference cluster over-splitting. (**C**) Identification of target (Smart-seq2) cells equivalent to reference (Drop-seq) clusters using marker based approach. The top two differentially enriched genes (or marker genes) in each reference (Drop-seq) cluster were used to identify equivalent cells in the target (Smart-seq2) data using a thresholding approach. Bars represent percentage of classified and unassigned cells using various thresholds for normalized gene expression of the marker genes as indicated on the x-axis. Gray represents the percentage of cells that express markers of multiple clusters; yellow represents the percentage of cells that can be unambiguously classified to a single cluster; and blue represents the percentage of cells that do not express markers of any of the clusters. These cells are referred to as orphans. **(D)** Batch correction by CCA and MNN alters grouping of reference cells. Heatmap showing row-normalized average expression of gene signatures in the reference Drop-seq data after batch correction of reference and target data with CCA (left) or MNN (right). Red indicates enrichment and blue indicates depletion of the gene signature levels relative to average expression of that gene signature across all clusters. Row labels indicate names of the reference gene signatures. Column names are arbitrary names given to the clusters in the batch corrected and merged reference and target data. Post-MNN corrected data was clustered with Seurat. (**E**) Assessment of the extent of cluster merging in post-CCA (**Figure 2D**, left) and post-MNN (**Figure 2D**, right) reference clusters compared to the uncorrected reference (**Figure 2A** and **2B**, left). For clusters generated by each method, the number of significantly expressed gene signatures per cluster were counted (as described in **Methods**). Each segment represents the percent of clusters with the indicated number of significantly expressed gene signatures. Ideally, the corrected reference should express similar pattern of gene signature enrichment compared to the uncorrected reference (i.e. 1 signature per cluster indicated by green colour). (**F**) Heatmap showing row normalized average gene signature expression in the target (Smart-seq2) retinal bipolar data grouped by scmap (left) and scID (right). Khaki represents enrichment and turquois represents depletion. Row labels indicate names of gene signatures. Column labels are those assigned by scmap or scID. **(G)** Assessment of the extent of cluster merging in scmap (**Figure 2F**, left) and scID (**Figure 2F**, right) mapped equivalent cells in the target. For each cluster, the number significantly expressed gene signatures were counted (as in **Figure 2E**). Each segment represents the proportion of clusters with the indicated number of significantly expressed gene signatures. For comparison, the number of enriched signatures per cluster are shown for uncorrected reference and for uncorrected target. **(H, I)** Assessment of scID and scmap. Target data that can be unambiguously labeled using marker based approach as shown in **Figure 2C** (yellow) was used as ground truth. To be unbiased, several thresholds of normalized gene expression are used as indicated in the x-axis. Higher Adjusted Rand Index and lower Variation of Information indicate better accuracy.*

While unsupervised clustering of the full Drop-seq data with Seurat resulted in ~18 discrete clusters, unsupervised clustering of the Smart-seq2 data with Seurat and several other clustering methods (data not shown) identified only 4 clusters (**Figure 2A**) possibly due to low cell numbers (Heimberg et al., 2016). To determine how the transcriptional signatures of the reference clusters are distributed in the target data, we computed the average gene signature per cluster (**Figure 2B;** see **Methods**). The dominant diagonal pattern in the gene signature matrix for the reference data indicates the specificity of the extracted gene signatures. All the subtypes of bipolar cells (BCs) that were well separated in the reference Drop-seq data co-clustered in the target Smart-seq data (**Figure 2B**). Surprisingly, despite over 7 times greater number of genes per cell in the target Smart-seq2 data compared to the reference Drop-seq data, the biomarker approach of using top markers of clusters from the reference Drop-seq data resulted in unambiguous assignment of only a small proportion of cells (**Figure 2C**). Moreover, as apparent from the gene expression matrix, the alignment of the reference and the target data with either MNN (Haghverdi et al., 2018) or Seurat (Butler et al., 2018) altered the grouping of the cells in the post-aligned reference (**Figure 2D**; **Figure 2E**) as shown by the inconsistent gene signatures expression relative to the unaligned reference clusters (**Figure 2A,** left).

Thus, we assessed the ability of scID and scmap (Kiselev et al., 2018) to assign cluster identity to cells in the target Smart-seq2 data using Drop-seq data as reference. scID assigned 86.5% (n = 249) of the Smart-seq2 cells into 17 Drop-seq clusters while scmap was able to assign 63.2% (n = 182) of the Smart-seq2 cells into 10 Drop-seq clusters (**Figure 2F**). For 6 cell types, both scID and scmap were able to recover equivalent cells from the target correctly as revealed by the enrichment of only a single gene signature in these clusters (**Figure 2F**). However, for the other four clusters that were identified by both scmap and scID, the scID identified cell groups had lower expression of gene signatures of other cell types (**Figure 2F**). Quantification of the number significantly expressed gene signatures in each cluster (**Figure 2G**) suggest that scmap had higher proportion of merged clusters than scID clusters. Using target cells that specifically expressed cell type specific markers (**Figure 2C**) as ground truth,

two additional and independent metrics suggested that scID identified cells were classified more accurately than those by scmap (**Figure 2H,I**). Finally, scID was able to classify between 25 to 80 percent of orphan cells (i.e. the cells without expression of top markers of the reference due to dropout) depending on the gene expression threshold used to define orphan cells as shown in **Figure 2C**.

Reference based identification of equivalent cells in scRNA-seq and ultra-sparse single nuclei RNA-seq from mouse brain.

In many clinical settings where only post-mortem tissues are available, single nuclei-RNA seq (snRNA-seq) (Habib et al., 2017; Lake et al., 2016) may be possible. However, the transcript abundance in the nuclei is significantly lower than in the cytoplasm and as a consequence the complexity of scRNA-seq libraries from nuclei-scRNA-seq from the same tissue is significantly lower than in whole cell scRNA-seq (Habib et al., 2017) thus making it challenging to cluster despite large number of cells. To assess how well scID performs compared to alternatives in this setting, we obtained publicly available Chromium 10X (Zheng et al., 2017) based scRNA-seq data on mouse brain cells and snRNA-seq on mouse brain nuclei (see **Methods**).

The brain whole cells scRNA-seq data had ~9K cells and the brain nuclei snRNA-seq data had ~1K cells. Unsupervised clustering of the data separately identified different number of clusters which did not have an obvious one-to-one correspondence (**Figure 3A**). As the whole cell scRNA-seq data has more cells than the nuclei-seq we used the former as reference and obtained gene sets that are differentially expressed in each cluster compared to the rest of the cells (**Figure 3B**). Efforts to identify cluster membership of nuclei-seq data based on top markers in each of the clusters from the whole cell scRNA-seq data allowed unambiguous classification of only a small number of cells possibly due to shallow coverage which leads to low dynamic range in gene expression. Not surprisingly, there was significant number of orphan cells (i.e. those in which transcripts of cell markers were not captured) (**Figure 3C**). Attempts to identify equivalent cells via joint clustering after aligning the data with either CCA or MNN seem to have merged clusters of the original reference data (**Figure 3D,E**). scmap identified cells equivalent to all the reference clusters while scID failed to identify equivalent cells for one of the clusters (R13) (**Figure 3F**); however, the gene signature expression matrix for scID groupings of the target data had fewer merged clusters than scmap (**Figure 3G**). Using target cells that specifically expressed cell type specific markers (**Figure 3C**) as ground truth, scID was deemed more accurate than scmap in mapping target cells to reference clusters (**Figure 3H,I**). Depending on the gene expression threshold used to classify orphan cells (**Figure 3C**), scID was able to assign labels to between 50% to 90% of the orphan cells.

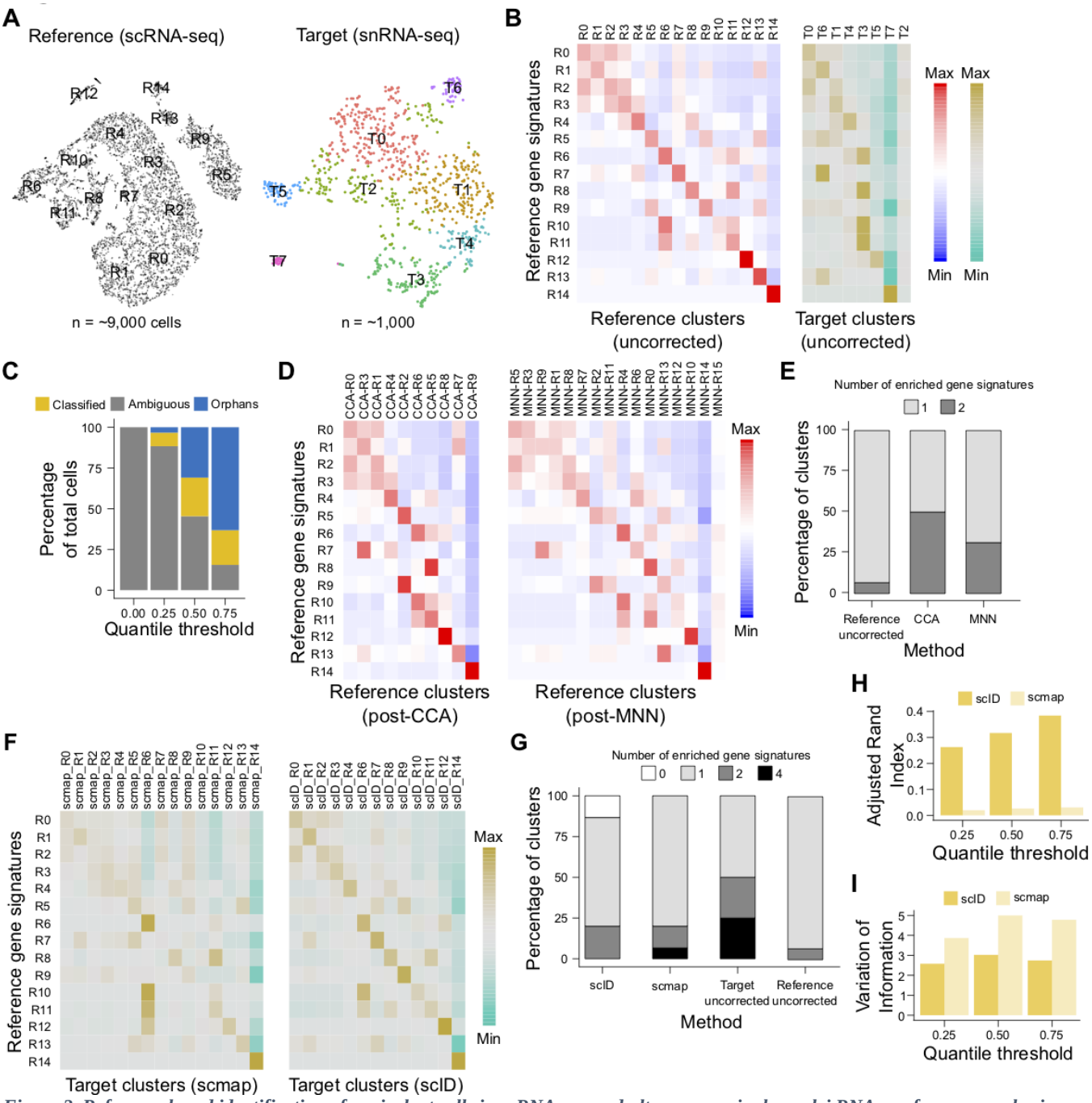

Figure 3. Reference based identification of equivalent cells in scRNA-seq and ultra-sparse single nuclei RNA-seq from mouse brain. *(**A**) t-SNE plot showing clusters in the mouse brain scRNA-seq data with ~9,000 cells (left) and tSNE of the mouse brain single nuclei RNA-seq (snRNA-seq) data with ~1,000 cells (right)**.** Data were clustered with Seurat and clusters were assigned names arbitrarily **(B)** Heatmap showing z-score normalized average expression of gene signatures in the reference scRNA-seq data (left) and in the target snRNA-seq data (right). Red (khakhi) indicates enrichment and blue (turquois) indicates depletion of the gene signature levels relative to average expression of that gene signature across all clusters of reference (target) data. Row labels indicate names of the gene signatures. Column names indicate cluster identity. The cluster names for the scRNA-sq and snRNA-seq data are arbitrary. Presence of multiple dark red (khaki) pixels in a single column reflects merging of separate reference clusters; presence of multiple dark red (khaki) pixels in a single row reflects reference cluster over-splitting. **(C)** Marker-based identification of cell types in target data using the two most differentially expressed markers from each reference cluster using a thresholding approach using various thresholds of quantile normalized gene expression as indicated on the x axis. Gray represents the percentage of cells that express markers of multiple clusters; yellow represents the percentage of cells that can be unambiguously classified to a single cluster; and blue represents the percentage of cells that do not express markers of any of the clusters. **(D)** Batch correction by CCA and MNN alters grouping of reference cells. Heatmap showing row-normalized average expression of gene signatures in the reference scRNA-seq data after batch correction of reference and target data with CCA (left) or MNN (right). Red indicates enrichment and blue indicates depletion of the gene signature levels relative to average*

*expression of that gene signature across all clusters. Row labels indicate names of the reference gene signatures. Column names are arbitrary names given to the clusters in the batch corrected and merged reference and target data. Post-MNN corrected data was clustered with Seurat. (**E**) Assessment of the extent of cluster merging in post-CCA (**Figure 3D**, left) and post-MNN (**Figure 3D**, right) reference clusters compared to the uncorrected reference (**Figure 3A** and **3B**, left). For clusters generated by each method, the number of significantly expressed gene signatures per cluster were counted (as described in **Methods**). Each segment represents the percent of clusters with the indicated number of significantly expressed gene signatures. Ideally, each cluster should express only one signatures (shown in green), if the features extracted are unique per cluster which is not always possible depending on the similarity of the clusters in the reference. **(F)** Heatmap showing row normalized average gene signature expression in the target (snRNA-seq) data grouped by scmap (left) and scID (right). Khaki represents enrichment and turquois represents depletion. Row labels indicate names of gene signatures. Column labels are those assigned by scmap or scID. **(G)** Assessment of the extent of cluster merging in scmap (**Figure 3F**, left) and scID (**Figure 3F**, right) mapped equivalent cells in the target. For each cluster, the number significantly expressed gene signatures were counted (as in **Figure 3E**). Each segment represents the proportion of clusters with the indicated number of significantly expressed gene signatures. For comparison, the number of enriched signatures per cluster are shown for uncorrected reference and for uncorrected target. **(H, I)** Assessment of scID and scmap. Target data that can be unambiguously labeled using marker based approach as shown in **Figure 3C** (yellow) was used as ground truth. To be unbiased, several thresholds of normalized gene expression are used as indicated in the x-axis. Higher Adjusted Rand Index and lower Variation of Information indicate better accuracy.*

Discussion

Comparison of multiple scRNA-seq datasets across different tissues, individuals and perturbations are routinely done in order to reveal potential biological mechanisms underlying phenotypic diversity. However, comparison of scRNA-seq datasets is challenging because even scRNA-seq data from the same tissue but with different technology or donor can have significant technical or batch effect confounding the technical with the biological variability. scRNA-seq datasets generated from different scRNA-seq methods and quality (i.e. similar number of cells and library depth per cell) are in general much easier to correct than data generated from different methods and quality. Particularly, in cases of strong asymmetry in quality between datasets to be compared, mapping target cells across data at the individual cell level rather than performing batch correction can be a more appropriate route for identifying known transcriptionally related cells.

We have developed scID which uses the framework of Fisher's linear discriminant analysis to map cells from a target data to a reference dataset (**Figure 1A**). scID adapts to characteristics of the target data in order to learn the discriminative power of cluster-specific genes extracted from the reference in the target data. scID down-weighs the genes within the signature that are not discriminatory in the target data. Through extensive characterization of multiple published scRNA-seq datasets, we assessed scID's accuracy relative to alternative approaches in situations where the reference and target datasets have strong asymmetry in quality and have shown that scID outperforms existing methods for recovering transcriptionally related cell populations from the target data which does not cluster concordantly with the reference (**Figure 2-3**).

Due to reliance on gene signatures to discriminate between clusters, a limitation of scID is that it only works for reference cluster that have their own gene signatures. Although, this requirement may not always be satisfied for all datasets, this situation appears to be rare in the datasets that have been analyzed here. Moreover, scID only uses genes that are specifically upregulated in the cluster of interest for discriminating clusters. scID can easily be extended by including genes that are specifically downregulated in the cluster of interest as well by modifying the way the scores are calculated. Another limitation of scID is that if the target contains a cell type that is related to but not exactly the cell type in the reference, scID is susceptible to selecting these closely related cells. This problem can be mitigated by better approaches for identifying gene signatures and including a minimum score threshold to be classified to a cluster in the last stage of scID, in which the scores are fit to two normal distributions.

As the scale of single cell RNA-seq data increases and the numbers of clusters obtained are so large that manual annotation is cumbersome, scID can enable automatic propagation of annotations and metadata across these data. In addition, unlike scmap, CCA and MNN, scID can be used for arbitrary user specified gene list without the need for providing information on gene expression levels thus aiding in the validation of clusters from unsupervised methods.

Author contributions

K.B. and N.N.B. conceived the study. K.B, S.S. and N.N.B. developed the computational framework. K.B. implemented the method and performed the analysis. K.B., S.S. and N.N.B. interpreted the results. N.N.B. wrote the manuscript with input from K.B and S.S. N.N.B. supervised the study.

Acknowledgements

K.B. and N.N.B. were funded by the Chancellor's Fellowship from the University of Edinburgh to N.N.B.

Competing Interests

The authors declare no competing interests

Methods

scID workflow

Given a set of k features (genes) that are upregulated in a reference cluster c, scID computes the probability, $P(j \in c \mid s_j^c)$, of a target cell j belonging to that cluster given a score s_j^c, which is a weighted linear sum given by

$$s_j^c = \frac{\sum_{i=1}^{k} w_i^c \tilde{g}_i^j}{\sum_{i=1}^{k} w_i^c} \qquad \textbf{(Eq. 1)}$$

for $j = 1, \ldots, n$ and n is the number cells in the target and $\tilde{g}_i^j \in [0,1]$ is the normalized gene expression values of the i-th gene in the j-th cell, $w_i^c > 0$ is a weighting factor that represents the discriminative power of gene i to identify the cells from cluster c in the reference population. To reduce sensitivity to outliers, the gene expression values are normalized to $[0,1]$ by 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

$$w_i^c = \frac{\mu_i^c - \mu_i^{c^-}}{\sigma_i^{c2} + \sigma_i^{c-2}} \qquad \textbf{(Eq. 2)}$$

where μ_i^c, σ_i^c represent the mean and standard deviation, respectively, of gene i in the cluster c . c^- are the set of cells in the reference data not in c. Each term of the weight of gene i is in turn calculated as follows:

$$\mu_i^c = \frac{1}{l^c}\sum_{j\in c}\widetilde{g_i^j}$$

$$\mu_i^{c^-} = \frac{1}{N-l^c}\sum_{j\in c^-}\widetilde{g_i^j}$$

$$\sigma_i^{c2} = \frac{1}{l^c}\sum_{j\in c}(\widetilde{g_i^j} - \mu_i^c)^2$$

$$\sigma_i^{c^-2} = \frac{1}{N-l^c}\sum_{j\in c^-}(\widetilde{g_i^j} - \mu_i^{c^-})^2$$

where l^c is the number of cells in cluster c and N is the total number of cells in the data. Only the signature genes with positive w_i^c are retained.

Although we can compute this weight from the reference, we found that computing this weight from the target data instead leads to improved accuracy (**Figure 1C**), possibly due to adjustments to the data quality in the target data. However, to do this we need to select target cells that likely belong to the cluster c. The target cells (denoted as $\tilde{c}$) that express the k signature genes precisely and specifically are selected as equivalent to the reference cluster c by clustering the data in the precision-recall domain with *k*-means and $\mathrm{k} = 2$. The precision and recall are defined as follows

$$precision = \frac{n_{me}}{n_e} = \frac{number\ of\ signature\ genes\ expressed}{total\ number\ of\ expressed\ genes}$$

$$recall = \frac{n_{me}}{n_m} = \frac{number\ of\ signature\ genes\ expressed}{total\ number\ of\ signature\ genes}$$

In the ideal scenario, $\tilde{c}$ cells are close to $(1,1)$ and $\tilde{c}^-$ cells are close to $(0,0)$ in the precision-recall domain. Analogous to computing the weights in the reference data, they can be computed from the target data.

This definition of the weights (**Eq. 2**) is a solution to Fisher's linear discriminant analysis which aims to find an optimal linear projection that maximizes the separation between two classes. The separation between the two classes is defined as the of the variance between the classes to the variance within the classes. Choosing the weights in this way penalizes high variability and low mean expression in order to account for the following cases:

a. When a gene is not expressed uniquely in the $\tilde{c}$ population, $\mu_i^{\tilde{c}}$ will increase, reducing the weight as it does not provide sufficient evidence of belonging to the cells of interest.
b. When a gene is expressed only within the subpopulation of $\tilde{c}$, $\sigma_i^{\tilde{c}2}$ 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 $\mu_i^{\tilde{c}} - \mu_i^{\tilde{c}^-}$ will be low.

The weights estimated from the target data reflects 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 partition the cells into c and c^- population, scID fits a mixture of two Gaussian distributions. Accordingly, the likelihood of a target cell being a cell of interest is given by:

$$P(j \in c \mid s_j^c) = \frac{\lambda\, N\left(s_j^c \middle| \mu^c, \sigma^c\right)}{\lambda N\left(s_j^c \mid \mu^c, \sigma^c\right) + (1-\lambda)\, N\left(s_j^c \mid \mu^{c^-}, \sigma^{c^-}\right)} \qquad \textbf{(Eq. 3)}$$

where $\mu^c > \mu^{c^-}$ and λ represents the mixing ratio of the two distributions. The parameters of the mixture model are estimated using the Expectation-Maximization algorithm. Cells are assigned to cluster c, if their score was such that the above probability was greater than 0.95 unless specified otherwise. 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 et al., 2016): This Smart-seq2(Picelli et al., 2013) 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 et al., 2016)*:* This study performed Drop-seq and Smart-seq2 experiments on Vsx2-GPF 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. These 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 subverticular 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).

Intestinal stem cells from Biton et al. 2018 (Biton et al., 2018): This combination of plate-based and droplet-based scRNA-seq data of mature epithelial cell types from mouse intestinal stem cells consists of 24,649 cells with an average of 1,956 genes per cell. Unsupervised clustering of the cells resulted in 12 clusters. Data was downloaded from the Broad Institute Single Cell Portal.

Murine tracheal epithelium cells from Montoro et al. 2018 (Montoro 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 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.

scID implementation

The R implementation and tutorial for scID is available on Github (https://github.com/BatadaLab/scID).

scID has the following user-specified options:

1. `logFC [Default: 0.5]`: Specify threshold for extracting markers to select a subset of the significantly differentially expressed markers. We used default except for the **Figure 1B**, Hu et al 2017 data where a value of `0.3` was used due to high similarity of the cell populations.
2. `use_reference_for_weights [Default: FALSE]`: Estimate weights from either reference or target data for calculation of gene signature weights. For all figures in this paper, we used the default option.
3. `likelihood_threshold [Default: 0.99]`: Minimum required likelihood of gene signature score for a cell to be assigned to a reference cluster (**Eq. 3**). We used default except for **Figure 1B**, where we used a value of 1.0 for Montoro et al 2018 and `0.95` for other datasets.

Software tools and parameter settings

We used the following R packages: SC3 (Kiselev et al., 2017) (version 1.7.7), Seurat (Butler et al., 2018) (version 2.3.1), scran (version 1.6.9) and scmap (Kiselev et al., 2018) (version 1.1.5). In general, we used the guidelines defined by the developers of each method. For Seurat we selected number of principal components based on the elbow plot. For SC3 we adjusted the number of principal components as recommended by the authors. Seurat's CCA (Butler et al., 2018) was run with default parameters using canonical correlation components based on the elbow plot. MNN (Haghverdi et al., 2018) was run with default parameters using 150 selected features to estimate the batch effect corrected expression values for the two datasets. Aligned and merged expression matrix were clustered using Seurat with default parameters with number of principal components determined by the elbow plot.

Statistical tests

For the comparison of scID and scmap classification of cells to their known 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.

Gene signature matrix

Each row of gene signature expression matrix (shown in **Figure 2B,D,F** and **Figure 3B,D,F**) shows the average expression of signature genes in cells grouped by cluster. For visualization as heatmaps, the values of the average values per cluster are z-score normalized by row.

Computation of number of enriched signatures per cluster

To quantitively compare the extent of cluster merging (shown in **Figure 2E,G** and **Figure 3E,G**) we count the number of gene signatures that are significantly enriched in each cluster. If the average expression of a signature in a cluster exceeds the 95th percentile of the values of this signature across all clusters, that cluster is considered to express that signature. We repeat this for all clusters and signatures and count the number of expressed signatures per cluster.

References

Aggarwal, C.C., Hinneburg, A., Keim, D.A (2001). On the surprising behaviour of distance metrics in high dimensional space. International Conference on Database Theory, 420-434.

Bacher, R., and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol *17*, 63.

Biton, M., Haber, A.L., Rogel, N., Burgin, G., Beyaz, S., Schnell, A., Ashenberg, O., Su, C.W., Smillie, C., Shekhar, K., *et al.* (2018). T Helper Cell Cytokines Modulate Intestinal Stem Cell Renewal and Differentiation. Cell *175*, 1307-1320 e1322.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol *36*, 411-420.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., *et al.* (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol *16*, 278.

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S.R., Aguet, F., Gelfand, E., Ardlie, K., *et al.* (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods *14*, 955-958.

Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol *36*, 421-427.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep *2*, 666-673.

Heimberg, G., Bhatnagar, R., El-Samad, H., and Thomson, M. (2016). Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. Cell Syst *2*, 239-250.

Hu, P., Fabyanic, E., Kwon, D.Y., Tang, S., Zhou, Z., and 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 e1007.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., *et al.* (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776-779.

Kiselev, V.Y., Andrews, T.S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet.

Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., *et al.* (2017). SC3: consensus clustering of single-cell RNA-seq data. Nat Methods *14*, 483-486.

Kiselev, V.Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. Nat Methods *15*, 359-362.

Lake, B.B., Ai, R., Kaeser, G.E., Salathia, N.S., Yung, Y.C., Liu, R., Wildberg, A., Gao, D., Fung, H.L., Chen, S., *et al.* (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science *352*, 1586-1590.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., *et al.* (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202-1214.

Montoro, D.T., Haber, A.L., Biton, M., Vinarsky, V., Lin, B., Birket, S.E., Yuan, F., Chen, S., Leung, H.M., Villoria, J., *et al.* (2018). A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. Nature *560*, 319-324.

Picelli, S., Bjorklund, A.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods *10*, 1096-1098.

Picelli, S., Faridani, O.R., Bjorklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc *9*, 171-181.
Ramskold, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., *et al.* (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol *30*, 777-782.
Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., *et al.* (2017). The Human Cell Atlas. Elife *6*.
Shekhar, K., Lapan, S.W., Whitney, I.E., Tran, N.M., Macosko, E.Z., Kowalczyk, M., Adiconis, X., Levin, J.Z., Nemesh, J., Goldman, M., *et al.* (2016). Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. Cell *166*, 1308-1323 e1330.
Stegle, O., Teichmann, S.A., and Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet *16*, 133-145.
Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., *et al.* (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189-196.
Vallejos, C.A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J.C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods *14*, 565-571.
Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., *et al.* (2017). Massively parallel digital transcriptional profiling of single cells. Nat Commun *8*, 14049.