Mapping transcriptionally equivalent populations across single cell RNA-seq datasets

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

Institute for Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK EH4 2XU

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

Abstract

Comparison of single cell RNA sequencing (scRNA-seq) data across individuals, tissues and perturbations is important but challenging due to technical differences in sample processing, number of cells, read depth and technical variation. To address these challenges we developed scID (single cell identification), which uses the linear discriminant analysis framework to identify transcriptionally related cells across scRNA-seq datasets. Through extensive characterization of published data we demonstrate that despite differences in the sequencing depth, cell coverage, tissue composition and technology used, scID outperforms existing approaches for identifying equivalent cell populations across datasets.

Introduction

Single cell RNA-sequencing (scRNA-seq)[1-6] is now being routinely used to characterize cell type specific changes in development, disease and perturbations[7] which require analysis of cross-data comparison[8-10]. However, there are technical and biological variations between the different datasets which may hamper the joint analysis of these data. Batch correction approaches such as CCA[10] and MNN[9] have been designed to integrate scRNA-seq datasets. The advantage of these methods is that batch correction allows merging of datasets leading to greater number of cells which provide more power to resolve cell types, particularly those that are highly similar or are rare. In both CCA and MNN, aligning cells is based on preserving the local correlation or neighbourhood structure of data. However, when the data are significantly imbalanced in the number of cells it is unclear how these methods perform in such situations. Depending on the extent of technical bias or differences in tissue composition, 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, it is worthwhile to reuse such high-quality information to identify known populations of cells in the target data. In addition to propagation of information such as cell labels and metadata across datasets, reference data can also be used to group cells in the target data in which closely related clusters or rare clusters are difficult to resolve via unsupervised clustering. A naive approach to identifying equivalent cell populations across dataset involves simply iterating through data from each cell and determining whether marker genes are present or absent. However, this naïve approach does not work effectively as many scRNA-seq data such as those from droplet-based methods usually have a very shallow sequencing coverage and suffer from an extreme dropout problem i.e. for a large proportion of genes, the expression values are not sampled[11-13]. To map cells from target scRNA-seq data to a reference, scmap[8] extracts features from each reference cluster and uses a distance or 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. Here we present a novel method called scID which uses Fisher's Linear Discriminant Analysis (LDA) for identifying transcriptionally related groups of cells across datasets. Through extensive comparison of published datasets with both technical and biological differences, we show that scID is more accurate than existing methods particularly for clusters that are transcriptionally close. Moreover, scID is relatively fast and scales linearly with the number of cells in the target data.

Results

Identification of transcriptionally similar cells across scRNA-seq datasets using Linear Discriminant Analysis

scID maps cells in the target data that are equivalent to clusters in the reference data in three key steps (**Fig. 1a**). First, genes that are differentially upregulated in a cluster are extracted from the reference. Second, the cells in the target cells that are most concordant with extracted features are identified using Fisher's LDA. Finally, conflicts in assignments, if any, are resolved. Conflicts in assignments can arise when the reference clusters are close in transcription space and the extracted features poorly discriminate between these groups.

The identification of transcriptionally related cells between the reference and the target data can be formulated as a classification problem. To identify cluster-of-interest (or "IN" class), scID projects the target data onto one dimension through a linear combination of normalized gene expression of the form

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

where *k* is the number of features, $\tilde{g}_i^j$ is the normalized gene expression of feature i in cell j and w_i 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 IN and the OUT population (i.e. the reference data without the IN cluster) are far while at the same time the sum of the within class scatter low. According to Fisher's LDA, the maximum separation between IN and OUT clusters are achieved with weights (w) of the form:

$$w_i = \frac{\mu_{iIN} - \mu_{iOUT}}{\sigma_{IN}^2 + \sigma_{OUT}^2} \qquad \textbf{(Eq. 2)}$$

where μ represents mean expression of the cluster and σ represents standard deviation of gene expression within the cluster.

It is likely that 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, discriminative power or weights for each of these genes is identified from the target data by first identifying a small set of positive and negative sets that can be identified confidently based on the proportion of cluster-specific features present in these cell. The scores learned by scID from the target data using the above approach in which the cell labels are unknown is highly concordant with the log fold change (logFC) in the expression of the selected discriminatory genes identified from the reference data (**Fig. 1b**).
Moreover, in many cases learning the weights from the target data improves accuracy of classification (**Fig. 1c**).

With the learned weights, each cell in the target data is projected using Eq. 1 and a mixture of Gaussians are fitted using the Expectation-Maximization algorithm to the projected scores to identify the valley or the threshold between the two clusters. 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. When closely related cells are present in the 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. scID resolves this by a relative normalized score strategy (see **Methods**). Through self-mapping experiments, we confirmed

that scID was properly able to assign cells to their respective cluster despite presence of very similar clusters (**Fig. 1d**).

scID can recover cluster identity of ultra-sparse data lacking expression of markers

To assess scID's accuracy relative to alternative approaches we aimed to identify subtypes of mouse monocytes from a highly sparse MARS-seq[6] based scRNA-seq data[14]. Mouse monocytes are well studied and are known to have two functionally dominant subtypes[14]. The first subtype, "subtype-H", is defined by high expression of *Ly6c* and low expression of *Nr4a1*; the second subtype, "subtype-L", is defined by low expression of *Ly6c* and high expression of *Nr4a1*. A third subtype, "subtype-I", is defined by intermediate expression of both *Ly6c* and *Nr4a1* genes. Mildner et al[14] performed MARS-seq on mouse circulating blood monocytes (~1,100 cells, ~400 genes per cell). Clustering peripheral blood mouse monocyte smart-seq2 data using SIMLR[15], SC3[16], CIDR[17] and Seurat[10] showed that the known subtypes can be uncovered via unsupervised clustering (**Fig. 2a**). However, none of these four methods were able to recapitulate the known clustering from the MARS-seq based mouse peripheral blood monocytes (**Fig. 2b**) despite 5 times more cells in this data than in the Smart-seq2[2] data (**Fig. 2a**). As nearly ~53% of the cells did not have expression of both Ly6c and *Nr4a1* ("orphan data") (**Fig. 2d**), the marker-based classification of subtype-L, subtype-H and subtype-I cells was not possible. scID was able to classify significantly more cells to clusters compared to scmap (**Fig. 2e**) and able to rescue significantly more orphan cells (~50%) than scmap (~5%) (**Fig. 2f**). As the markers of the two subtypes are well known and highly specific, we determine how well the features extracted from such cells (which made up 42%) in the MARS-seq can be used to classify orphan cells. We found scID was able to perform similarly as when the external smart-seq2 data was used as reference (**Fig. 2g**). The orphan cells assignment displayed highly concordant gene expression pattern when compared to the expression of genes in the non-orphan cells from the MARS-seq data (**Fig. 2h**). Accordingly, we conclude that for extremely sparse scRNA-seq data, scID identifies cell types of interest with low false negatives rate relative to existing methods.

scID can identify equivalent cells across whole cell scRNA-seq and sparse nuclei-seq scRNA-seq

In many clinical settings where only post-mortem tissues are available, single nuclei-RNA seq (snRNA-seq)[18,19] 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[18]. To assess how well scID performs compared to alternatives in this setting, we obtained publically available Chromium 10X[5] based scRNA-seq data on mouse brain whole cells and nuclei (see **Methods**). The brain whole cells scRNA-seq data had ~9K cells and the brain nuclei scRNA-seq data had ~1K cells. Unsupervised clustering of the data separately identified different number of clusters (**Fig. 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 (**Fig. 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; not surprisingly, there was significant number of orphan cells (**Fig. 3c**). Attempts to identify equivalent cells via joint clustering after aligning the data with either CCA or MNN were unsuccessful judging from the significant alterations in cluster membership of the reference data after applying these methods (**Fig. 3d**). While scmap and scID were able to rescue similar number of orphan cells (**Fig. 3h**), multiple metrics for assessment of results of

mapping the target (snRNA-seq) data to reference (scRNA-seq) clusters suggested that scID was more accurate than scmap (**Fig. 3f,g,i**).

scID multiclass assignment resolution accurately identifies equivalent cells in data with highly similar cells

Given a large number of cells despite shallow coverage per cell, they provide more power to resolve highly similar cells compared to using a low number of cells with deep coverage per cell[20]. This is well illustrated by the mouse retinal bipolar neuronal Drop-seq[1]-based scRNA-seq data[21] which has highly similar cell types. In that study[21], scRNA-seq data were generated from mouse retinal bipolar neurons using two different technologies: Drop-seq[1] based had ~27500 cells (with ~700 genes per cell) and the plate-based Smart-seq2 data had 288 cells (~5000 genes per cell). 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 identified only 4 clusters (**Fig. 4a**). All the subtypes of bipolar cells (BCs) that were well separated in the Drop-seq data co-clustered in the Smart-seq data as seen from the similarity of gene expression patterns per cluster to the reference Drop-seq data (**Fig. 4b**). Thus, when cell numbers are low, cell types with high transcriptional similarity can co-cluster. Despite significantly greater gene coverage of Smart-seq2 data (target), the naive approach using markers of clusters from the Drop-seq data (reference) resulted in unambiguous assignment of only a small proportion of cells (**Fig. 4c**). Moreover, alignment of data with either MNN[9] or Seurat[10] followed by clustering altered the grouping of the reference in the merged data and showed inconsistent gene signatures expression relative to the clusters formed from the published and validated Drop-seq data (**Fig. 4d**).

Thus, we assessed the ability of scID and scmap[8] to assign cluster identity to cells in the Smart-seq2 (target) data using Drop-seq data as reference. scID assigned 86.5% (n = 249) of the Smart-seq2 cells into 18 Drop-seq clusters while scmap was able to assign 63.2% (n = 182) of the Smart-seq2 cells into 10 Drop-seq clusters (**Fig. 4e**). For the top six highly abundant cell types that were identified by both scID and scmap, expressed gene signatures were qualitatively similar to those expressed in the equivalent reference DropSeq clusters. 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 (**Fig 4e**), suggesting that scID-based mapping approach has better sensitivity and specificity. The specificity largely arose from the multiclass assignment approach of scID rather than due to extraction of more discriminating feature selection. Both methods were able to identify cell-types in the Smart-seq2 data with prevalence concordant to that in the Drop-Seq data but three of the four clusters identified only by scID had lower prevalence than the ones identified by scmap (**Fig. 4f**). Larger proportion of orphan cells were classified by scID than scmap (**Fig. 4j**) and several metrics suggested that scID identified cells were classified more accurately than those by scmap (**Fig. 4g,h,j**).

Discussion

Comparison of multiple scRNA-seq datasets across different tissues, individuals and perturbations are routinely done in order to determine similarities and differences between the data. However, comparison of scRNA-seq data 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 which may be lost during forced alignment of the data. Moreover, if the scRNA-seq data to be compared do not have a sufficiently large number of cells appropriate for the heterogeneity of the sample or are roughly

similar in quality (i.e. contain similar number of cells, dynamic range of gene expression and sparseness), separate clustering of the data may not lead to clusters formed from gene expression signals. In such cases of strong asymmetry in cell number or coverage of the cells, mapping target cells at the individual cell level rather than at the cluster level becomes a more appropriate route for identifying known transcriptionally related cells. High quality and validated tissue references can help in identification of cluster specific markers. However, many clusters do not have a unique set of markers but instead require a large set of genes (sometimes referred to as gene signatures) which collectively are highly enriched in that cluster to discriminate them from other clusters in the data. Even in the rare cases when highly expressed cluster specific markers may exist, detection of such markers in the highly sparse target data may not work effectively due to dropout of the marker genes.

We have developed scID which uses the framework of Fisher's LDA to distinguish each cluster from the remaining clusters within a reference scRNA-seq dataset and to project the cells onto a one dimension space that leads to good separation between clusters (**Fig. 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 which leads to better results than when weights are learned from the reference (**Fig. 1d**). scID down-weighs the genes within the list that are not discriminatory and is consequently tolerant to errors in the gene list. Moreover, through normalisation, scID is able to resolve clashes (which can occur between cells belonging to highly related clusters) in cell assignments relatively accurately (**Fig. 1e**).

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 differ significantly in the number of cells, sequencing coverage per cell, tissue type and complexity, and data procured from different sample processing methodologies and technologies. We have shown that scID outperforms existing approaches for recovering transcriptionally related cell populations from the target data which does not cluster concordantly with the reference (**Fig. 2-4**). More notably, scID is able to assign memberships to data that lack expression of top cluster-enriched markers. scID also works when the user provides a gene list which contain the majority of the genes that are highly and specifically enriched in the cell population of interest (**Fig. 5**).

Using reference information, scID can partition the target data into the known and the unknown, the latter can then be scrutinized further to identify features of interest that may reflect aspects of the experimental system or perturbation. 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, scID can be used for arbitrary user specified gene list without the need for providing information on expression level which will aid in validation of clusters.

An attractive feature of scID is that it scales linearly with number of cells in the target dataset (**Table 1**). The combination of accuracy and speed makes scID suitable for exploiting massive reference data which will have millions of cells across an arbitrary number of scRNA-seq datasets. Similar to indexing of genome assemblies, scID can index very large reference scRNA-seq data to create a database of discriminatory features which obviate the need for keeping local copies of reference data.

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. and N.N.B. interpreted the results. N.N.B. wrote the manuscript with input from K.B. 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 discriminative of a reference cell cluster c, scID ranks n cells expressing m genes ($k < m$) in a target population $\boldsymbol{G} \in \boldsymbol{\mathcal{R}}^{\boldsymbol{m}\times\boldsymbol{n}}$ by computing a score for cell j in the target population as a weighted linear sum:

$$s^j = \frac{\sum_{i=1}^{k} w_i \tilde{g}_i^j}{\sum_{i=1}^{k} w_i}$$

where $\tilde{g}_i^j \in [0,1]$ is the normalized gene expression values of the i-th gene in the j-th cell, w_i is a weighting factor that represents the discriminative power of gene i to identify the cells of interest in the target population, $i = 1, \dots, k$, and $j = 1, \dots, n$.

To reduce sensitivity to outliers, the normalized gene expression values are normalized to $[0,1]$ by 99th percentile as

$$\tilde{g}_i^j = \min\left(\frac{g_i^j}{P_{99}}, 1\right)$$

where g_i^j is the normalized gene expression of gene i in cell j and P_{99} is the 99th percentile of the expression of gene i across all cells of $\boldsymbol{G}$.

The weights are computed as

$$w_i = \frac{\mu_{iIN} - \mu_{iOUT}}{\sigma_{iIN}^2 + \sigma_{iOUT}^2}$$

where μ_i, σ_i represent the mean and standard deviation of gene i respectively, in the IN and OUT sets of cells. IN is a set of cells that express precisely and specifically the k genes and OUT is a set of the rest of the cells of the target data. This definition of the weights is a solution to Fisher's LDA with diagonal covariance matrices and acts as a regularizer, penalizing high variability and low mean expression in order to account for the following cases:

a. When a gene is not expressed uniquely in the IN population, μ_{OUT} 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 in a subpopulation of IN cells, σ_{iIN}^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, a false gene that causes contamination in the list of genes will be down-weighted, as μ_{iIN} will be low.

We have assumed that the covariance is diagonal for 2 reasons: 1) Estimation of full covariance from small clusters (i.e. IN) can be noisy and 2) Estimating the full covariance with often hundreds of features is computationally intensive.

Each term of the weight of gene i is in turn calculated as follows:

$$\mu_{iIN} = \frac{1}{l}\sum_{j\in IN}\widetilde{g_i^j}$$

$$\mu_{iOUT} = \frac{1}{n-l}\sum_{j\in OUT}\widetilde{g_i^j}$$

$$\sigma_{iIN}^2 = \frac{1}{l}\sum_{j\in IN}(\widetilde{g_i^j} - \mu_{iIN})^2$$

$$\sigma_{iOUT}^2 = \frac{1}{n-l}\sum_{j\in OUT}(\widetilde{g_i^j} - \mu_{iOUT})^2$$

To calculate the gene weights, scID identifies putative IN and OUT cells in the target data based on the precision and recall, which are defined as

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

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

by clustering the data in the precision-recall domain with k-means and $\mathrm{k} = 2$. In the ideal scenario, putative IN cells are close to $(1,1)$ and putative OUT cells are close to $(0,0)$ in the precision-recall domain. The weights learned from the target data reflects the cellular composition of the target data and quality (i.e. to the distribution of dropout and dynamic range of gene expression) of the target data (which influences the variation of the feature) both of which influence the discriminatory power of the features.

To partition the cells into IN and OUT population, scID fits a mixture of two Gaussian distributions to the scores with mean … and standard deviation. Accordingly, the probability of a target cell being a cell of interest is given by:

$$P(j \in IN \mid s^j) = \frac{\lambda\, \mathcal{N}\left(s^j \middle| \mu_1, \sigma_1\right)}{\lambda \mathcal{N}(s^j \mid \mu_1, \sigma_1) + (1-\lambda)\, \mathcal{N}(s^j \mid \mu_2, \sigma_2)}$$

where $\mu_1 > \mu_2$ and λ represents the mixing ratio of the two distributions. The parameters of the mixture model are estimated using an Expectation-Maximization algorithm.

The final step of scID involves resolution of multiclass assignment. When the reference data contains clusters that are highly similar transcriptionally, the features from a cluster (i.e. the differentially expressed genes in each clusters) can be correlated with the features from another cluster which leads to the assignment of more than one class to a target cell. To resolve this, the scores of matching cells are scaled with z-score normalization and cell is assigned to the reference cluster *u* if its normalized score for reference cluster *u* is higher than its normalized score for all other reference cluster.

Data source

Mouse Monocytes from Mildner et al. 2017[14]: This MARS-seq[6] data consists of 1098 Lin^- $CD11b^+$ $CD115^+$ monocytes isolated from the peripheral blood of adult C57BL/6 mice. On average there were 474 genes per cell. Data was obtained from GEO using accession number GSE95701.

Human Metastatic Melanoma immune cells from Tirosh et al. 2016[22]: This Smart-seq2[23] data consists of malignant, immune and stromal cells from metastatic melanoma tumours from 19 patients, a total of 4645 cells. We have used the 3254 immune ($CD45^+$) cells for our analysis which on average had 3925 genes per cell. Data was downloaded from the Broad Institute Single Cell Portal.

Mouse Retinal Bipolar Neurons from Shekhar et al. 2016[21]*:* 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 X genes per cell. These data was downloaded from the Broad Institute Single Cell Portal.

Human Breast Tumour microenvironment immune cells from Azizi et al. 2018[24]: This 10X data consists of 45000 immune cells from 8 patients and four tissues; breast carcinomas, normal breast tissue, blood and lymph node. For our analysis we have used the 19790 cells from the breast carcinomas, with an average of ~735 genes per cell. Data was obtained from GEO (GSE114725).

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 have 9128 cells with an average of ~2500 genes per cell and the single nuclei RNA-seq data have X cells with an average of Y genes per cell. Both the datasets were downloaded from 10X Genomics.

Software tools and parameter settings

We used the following R packages: SC3[16] (version 1.7.7), Seurat[10] (version 2.3.1), CIDR[17] (version 0.1.5), SMILR[15] (version 1.4.1), scran (version 1.6.9) and scmap[8] (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, SIMLR and CIDR we adjusted the number of principal components according to the number of expected populations. Seurat's CCA[10] was run with default parameters using canonical correlation components based on the elbow plot. MNN[9] 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 assessment of scID weights as a metric to sort the features by discriminating power we compared the resulted gene weights using all reference datasets to the calculated log fold-change values of the genes from MAST using Spearman rank correlation. High correlation shows the ability of scID to recover the importance of marker genes.

For the comparison of scID and scmap classification of cells to their known labels (in reference data) 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.

Consensus matrix

The consensus matrix ($\boldsymbol{C}$) for combining clustering results of multiple clustering algorithms is an $n \times n$ matrix, where n is the total number of cells in the dataset, that shows the aggregate result of several clustering algorithms. Each value $C_{ij}, i = 1, \ldots, n \; and \; j = 1, \ldots n,$ represents how often the two cells i and j are located in the same cluster across all clustering results. Values range between 0 and 1, with 0 given to a pair of cells that no clustering algorithm groups together and 1 to a pair of cells that all clustering algorithms group together. The diagonal of the consensus matrix is equal to 1 for all cells. Intermediate values show disagreement of the different clustering algorithms and ambiguous grouping of cells.

Comparing prevalence of identified cells to the expected values (as shown in Fig. 4f)

To estimate the expected prevalence of each cell type in the Smart-seq2 data, we randomly sampled 288 cells from the reference Drop-Seq data to obtain equal number of cells with the Smart-seq2 and calculated the percentage of each cell type. To estimate confidence intervals of the cell proportions we repeated the procedure 10000 times.

Availability

scID is implemented as an R package and is available on Github (https://github.com/BatadaLab/scID).

Figure Legends

Figure 1. Mapping cells across scRNA-seq dataset using scID. (**a**) The three main steps involved in mapping cells across scRNA-seq data are i) feature extraction from reference data; ii) the identification of cells in the target described by the feature vector obtained from the reference (i.e. scID); iii) resolving multiclass assignments. (**b**) scID can recover relative importance of cluster specific genes from the target data within hidden labels. Each dot represents spearman rank correlation between scID weights learned from data with hidden labels and log fold change (logFC) computed from labeled data. (**c**) scID annotation of target cells has higher True Positive Rate (TPR) when weights are estimated from the target cells (green) compared to weights estimated from the reference cells (gray). False Positive Rate (FPR) is equally low in both cases. (**d**) Assessment of accuracy of multiclass resolution by scID and scmap via self-projection of published annotated data.

Figure 2. Reference based identification of cell populations when marker genes are not detected (**a**) Consensus clusters of Smart-seq2 mouse peripheral blood monocytes (196 cells) using four unsupervised clustering methods (CIDR[17], SIMLR[15], SC3[16] and Seurat[10]) reveals three clusters that agree with the three subtypes, defined in the literature which we refer to as subtype-H (*Ly6c2*^{high}*Nr4a1*^{low}) (upper left) , subtype-L (*Ly6c2*^{low}*Nr4a1*^{high}) (middle) and subtype-I (*Ly6c2*^{int}*Nr4a1*^{int}) (lower right). **(b)** Consensus clustering of mouse peripheral blood monocytes MARS-seq based scRNA-seq data (~1000 cells)[14] does not recover the three stable groups seen in smart-seq2 based data (shown in 1a) but instead shows high discordance between methods (See **Methods**). **(c)** Use of marker-based (i.e. based on Ly6c2 and Nr4a1 levels) approach fails to classify majority of the cells in the MARS-seq data (these cells referred to as orphans) due to high level of dropouts **(d)** Accuracy of scID and scmap in identifying non-orphan subtype-H and subtype-L cells from the MARS-seq dataset, using the smart-seq2 dataset as reference. **(e, f)** Percentage of MARS-seq classified orphan cells with scID and scmap using the Smart-seq2 (external) data or the non-orphans (internal) MARS-seq cells as reference. **(g)** Assessment of classification of orphan cells by scID. Heatmap shows expression of marker genes (rows) of subtype-H (green) and subtype-L (grey) in cells identified by scID. Cells of the same cell type are further grouped into those that could be identified using *Ly6c2* and *Nr4a1* (yellow) and the orphan cells that lack the two markers (blue).

Figure 3. Identification of equivalent cells in scRNA-seq and single nuclei RNA-seq . (**a**) t-SNE plot showing clusters in the mouse brain scRNA-seq data with ~9K cells (left) and tSNE of the mouse brain single nuclei RNA-seq data with 1K cells (right)**. (b)** Heatmap showing expression of gene signatures in the 9K brain scRNA-seq data (reference) and the 1K brain nuclei scRNA-seq data (target). Rows represent gene signatures of each reference cluster and columns represent clusters. (**c**) Marker-based identification of cell types in nuclei scRNA-seq dataset (target) using two most differentially expressed markers from each reference cluster. Each bar shows percentage of classified and unassigned cells using different thresholds of quantile normalized gene expression (shown in x axis). Stacked barplot shows relative levels of ambiguously classified cells i.e. that are expressing markers of multiple cell types (gray), cells that could be successfully classified (yellow) and cells that do not express any of the markers i.e. orphans (blue). **(d)** Heatmap showing expression of gene signatures in the reference cells after batch effect correction with CCA (left) and MNN (right). Rows represent gene signatures and columns clusters. **(e)** Heatmap showing expression of reference gene signatures in the identified groups in the nuclei dataset by scMap (left) or scID (right). Empty gray columns, reflecting groups for which equivalent cells were not identified, were inserted to align data of equivalent cell types identified by the two methods. **(f)** Comparison of the

percentage of orphan cells classified by scID (turquois) and scMap (gray). **(g, h)** Assessment of scID (turquois) and scMap (gray) assigned labels to target cells compared to marker-based identification shown in **3c** for the different thresholds of normalized gene expression. ARI, Adjusted Rand Index; VI, Variation of Information.

Figure 4. Reference based Identification of equivalent cells in scRNA-seq target datasets with low number of cells. (**a**) t-SNE plot showing clusters in Drop-seq (reference) and Smart-seq2 (target) mouse retinal bipolar cells from Shekhar et al[21]. Cluster membership of Drop-seq data (~27500 cells) were taken from the publication and visualized with Seurat. Smart-seq2 data (288 cells) were clustered using Seurat[10]. (**b**) Heatmap showing expression of Drop-seq cluster gene signatures in the original Drop-seq clusters (left) and in Smart-seq2 clusters (right). Gene signatures of clusters in Drop-seq ("gs_") are shown in rows. Column names represent cell types in the Drop-seq data as defined in the original publication[21]. (**c**) Marker-based identification of cell types in Smart-seq2 dataset using two most differentially expressed markers from each Drop-seq cluster. Each bar shows percentage of classified and unassigned cells using different thresholds of normalized gene expression (*x*-axis). Gray represents the percentage of cells that are expressing markers of multiple cell types, yellow represents the percentage of cells that could be successfully classified and blue represents the percentage of cells that did not express any of the markers over the selected threshold (orphans). **(d)** Heatmap showing expression of gene signatures of the various retinal cell types in the reference Drop-seq cells clustered after batch correction with CCA (left) or MNN (right). Red represents enrichment and blue represents depletion. Empty gray rows were inserted to align data of equivalent cell types identified by the two methods. (**e**) Heatmap showing expression of gene signatures of the retinal bipolar neuron subtypes in the Smart-seq2 groups identified by scMap (left) and scID (right). Empty gray rows were inserted to align data of equivalent cell types identified by the two methods. **(f)** Comparison of the proportion of cell types in Smart-seq2 identified by scMap (gray) and scID (turquois). The expected proportion of cell types by random sampling from the reference data is shown in black. Error bars represent 2 standard deviations from the mean. **(g, h)** Comparison of scID and scMap labels to marker-based identification shown in **Fig. 4c** for the different thresholds of normalized gene expression using Adjusted Rand Index (ARI) and Variation of Information (VI). **(i)** Percentage of classified orphan cells with scID and scMap.

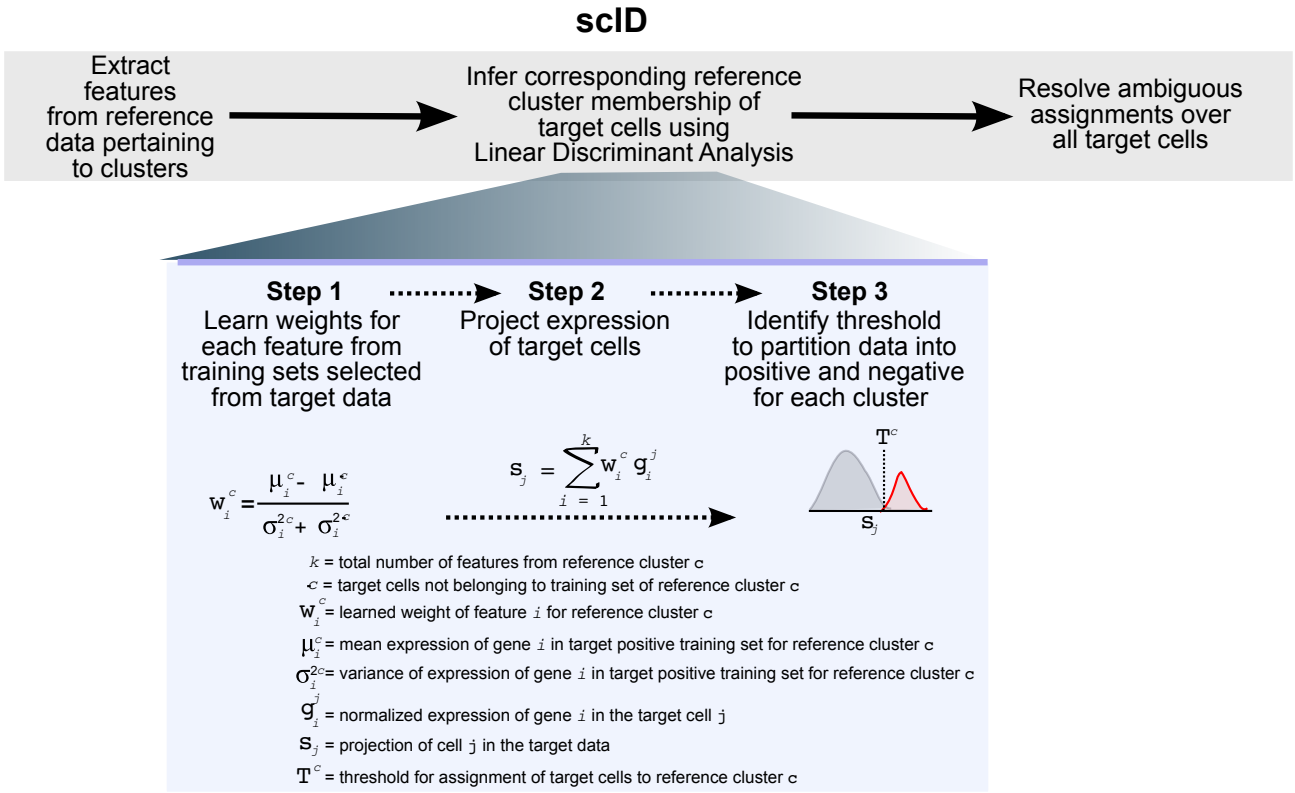
References

1. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
2. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171-81 (2014).
3. Ramskold, D. et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**, 777-82 (2012).
4. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* **2**, 666-73 (2012).
5. Zheng, G.X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).
6. Jaitin, D.A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776-9 (2014).
7. Regev, A. et al. The Human Cell Atlas. *Elife* **6**(2017).
8. Kiselev, V.Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods* **15**, 359-362 (2018).
9. Haghverdi, L., Lun, A.T.L., Morgan, M.D. & Marioni, J.C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421-427 (2018).
10. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
11. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133-45 (2015).
12. Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol* **17**, 63 (2016).
13. Vallejos, C.A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J.C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods* **14**, 565-571 (2017).
14. Mildner, A. et al. Genomic Characterization of Murine Monocytes Reveals C/EBPbeta Transcription Factor Dependence of Ly6C(-) Cells. *Immunity* **46**, 849-862 e7 (2017).
15. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods* **14**, 414-416 (2017).
16. Kiselev, V.Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **14**, 483-486 (2017).
17. Lin, P., Troup, M. & Ho, J.W. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol* **18**, 59 (2017).
18. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* **14**, 955-958 (2017).
19. Lake, B.B. et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586-90 (2016).
20. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst* **2**, 239-250 (2016).
21. Shekhar, K. et al. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166**, 1308-1323 e30 (2016).
22. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-96 (2016).

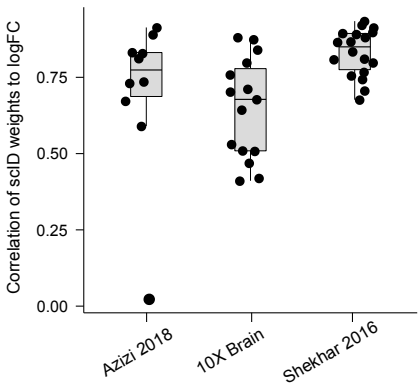
23. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096-8 (2013).
24. Azizi, E. et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174**, 1293-1308 e36 (2018).

Figure 1

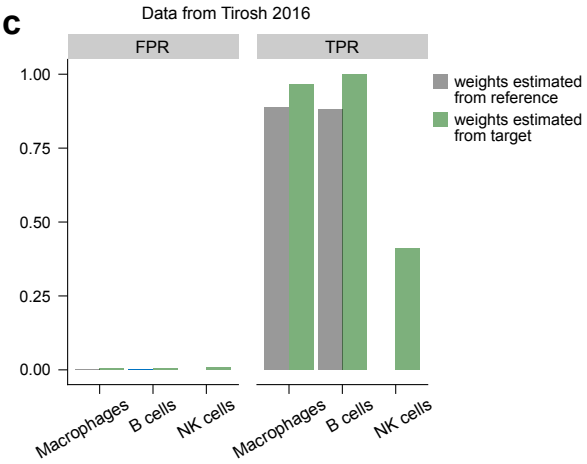
a



b



c



d

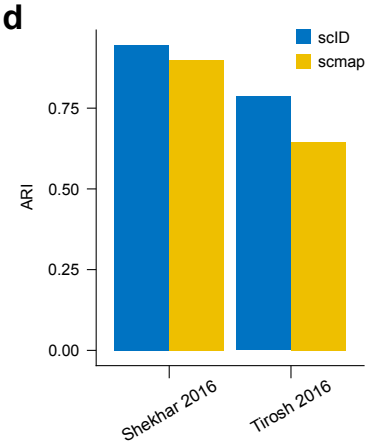


Figure 2

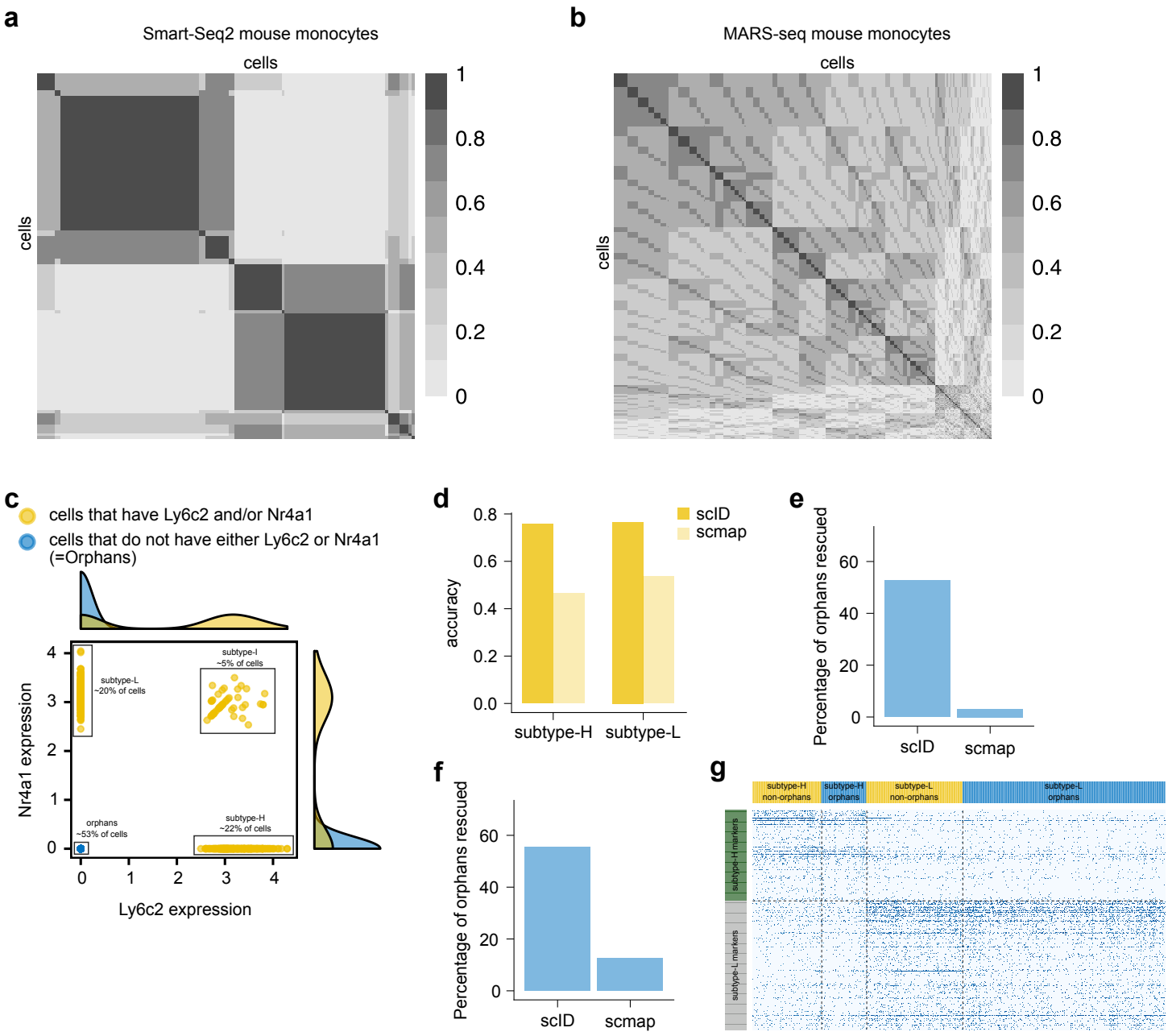

a
Smart-Seq2 mouse monocytes
cells
cells
1
0.8
0.6
0.4
0.2
0
b
MARS-seq mouse monocytes
cells
cells
1
0.8
0.6
0.4
0.2
0
c
cells that have Ly6c2 and/or Nr4a1
cells that do not have either Ly6c2 or Nr4a1 (=Orphans)
subtype-L ~20% of cells
subtype-I ~5% of cells
orphans ~53% of cells
subtype-H ~22% of cells
Nr4a1 expression
Ly6c2 expression
d
scID
scmap
accuracy
subtype-H
subtype-L
e
Percentage of orphans rescued
scID
scmap
f
Percentage of orphans rescued
scID
scmap
g
subtype-H non-orphans
subtype-H orphans
subtype-L non-orphans
subtype-L orphans
subtype-H markers
subtype-L markers

Figure 3

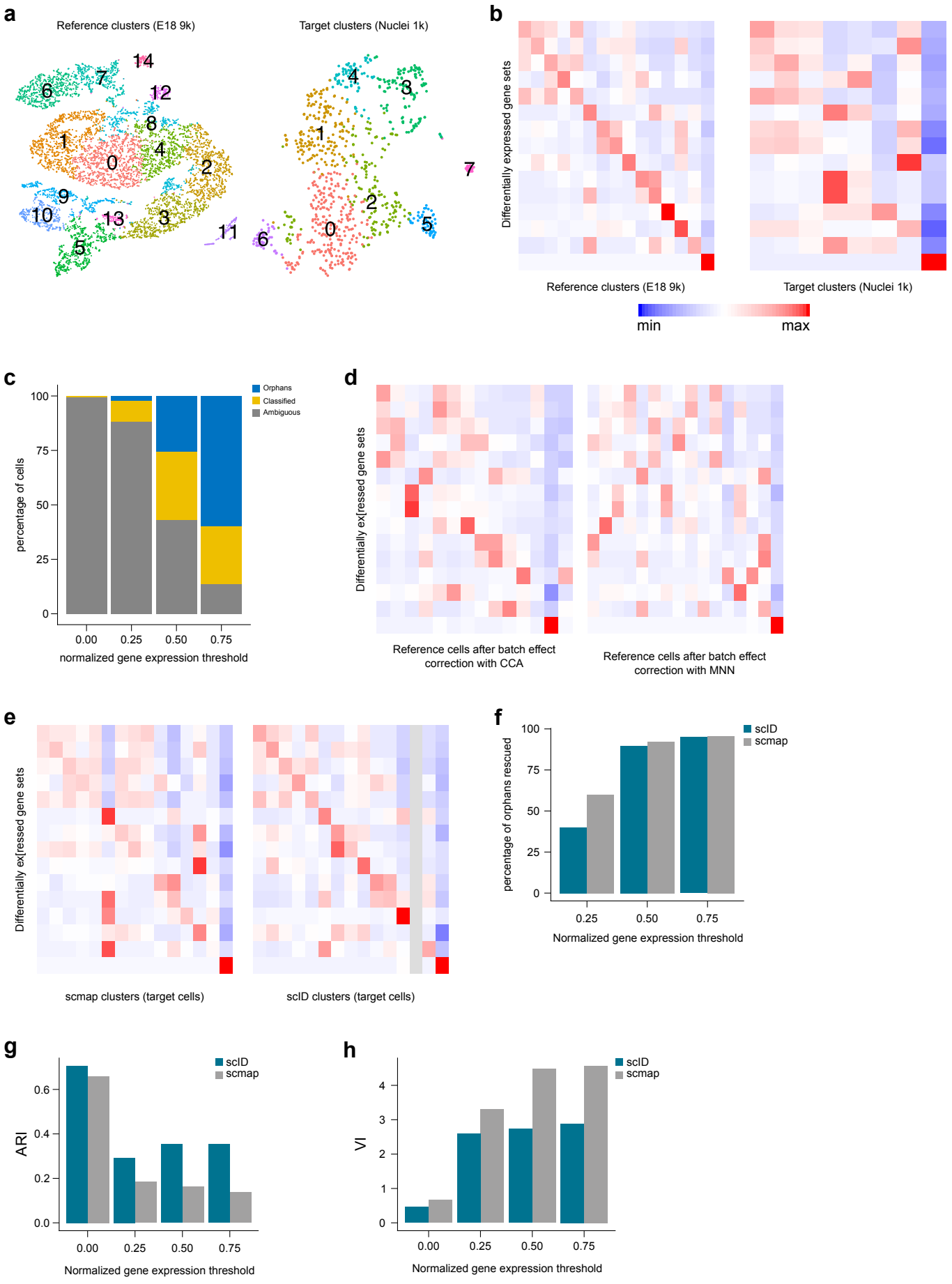
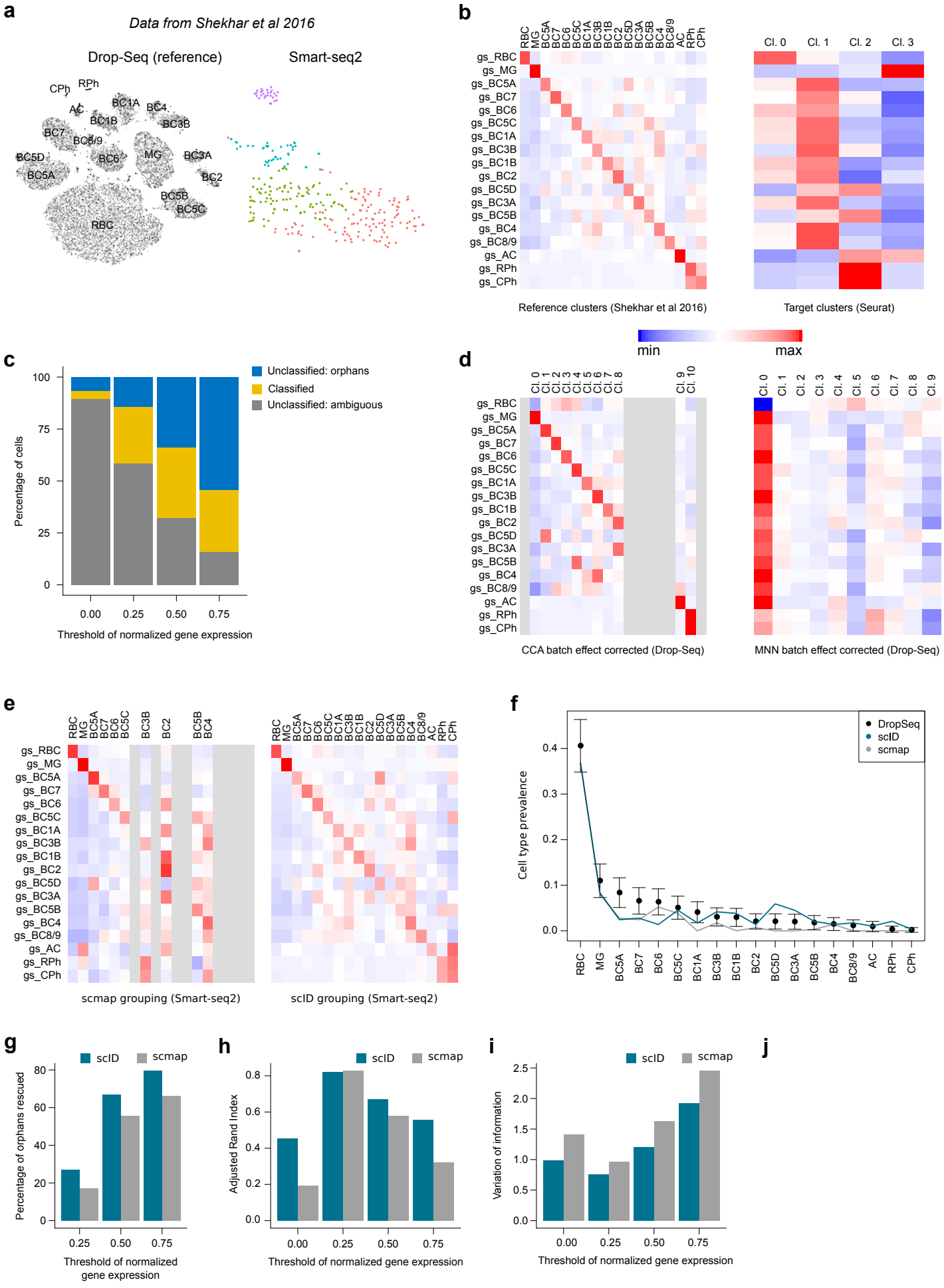

a
Reference clusters (E18 9k)
Target clusters (Nuclei 1k)
b
Differentially expressed gene sets
Reference clusters (E18 9k)
Target clusters (Nuclei 1k)
min
max
c
Orphans
Classified
Ambiguous
percentage of cells
normalized gene expression threshold
d
Differentially ex[ressed gene sets
Reference cells after batch effect correction with CCA
Reference cells after batch effect correction with MNN
e
Differentially ex[ressed gene sets
scmap clusters (target cells)
scID clusters (target cells)
f
scID
scmap
percentage of orphans rescued
Normalized gene expression threshold
g
scID
scmap
ARI
Normalized gene expression threshold
h
scID
scmap
VI
Normalized gene expression threshold

Figure 4


a
Data from Shekhar et al 2016
Drop-Seq (reference)
Smart-seq2
CPh
RPh
AC
BC1A
BC4
BC1B
BC3B
BC7
BC8/9
BC5D
BC6
MG
BC3A
BC5A
BC2
BC5B
BC5C
RBC
b
Reference clusters (Shekhar et al 2016)
Target clusters (Seurat)
Cl. 0
Cl. 1
Cl. 2
Cl. 3
c
Percentage of cells
Threshold of normalized gene expression
Unclassified: orphans
Classified
Unclassified: ambiguous
d
min
max
CCA batch effect corrected (Drop-Seq)
MNN batch effect corrected (Drop-Seq)
e
scmap grouping (Smart-seq2)
scID grouping (Smart-seq2)
f
Cell type prevalence
DropSeq
scID
scmap
g
Percentage of orphans rescued
Threshold of normalized gene expression
h
Adjusted Rand Index
Threshold of normalized gene expression
i
Variation of information
Threshold of normalized gene expression
j