BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

scConsensus: combining supervised and unsupervised clustering for cell type identification in single-cell RNA sequencing data

Bobby Ranjan[1†], Florian Schmidt[1†], Wenjie Sun[1], Jinyu Park[1], Mohammad Amin Honardoost[1,2], Joanna Tan[1], Nirmala Arul Rayan[1] and Shyam Prabhakar[1*]

*Correspondence: prabhakars@gis.a-star.edu.sg
†Bobby Ranjan and Florian Schmidt have contributed equally to this work
[1] Laboratory of Systems Biology and Data Analytics, Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore
Full list of author information is available at the end of the article

Abstract

Background: Clustering is a crucial step in the analysis of single-cell data. Clusters identified in an unsupervised manner are typically annotated to cell types based on differentially expressed genes. In contrast, supervised methods use a reference panel of labelled transcriptomes to guide both clustering and cell type identification. Supervised and unsupervised clustering approaches have their distinct advantages and limitations. Therefore, they can lead to different but often complementary clustering results. Hence, a consensus approach leveraging the merits of both clustering paradigms could result in a more accurate clustering and a more precise cell type annotation.

Results: We present scConsensus, an **R** framework for generating a consensus clustering by (1) integrating results from both unsupervised and supervised approaches and (2) refining the consensus clusters using differentially expressed genes. The value of our approach is demonstrated on several existing single-cell RNA sequencing datasets, including data from sorted PBMC sub-populations.

Conclusions: scConsensus combines the merits of unsupervised and supervised approaches to partition cells with better cluster separation and homogeneity, thereby increasing our confidence in detecting distinct cell types. scConsensus is implemented in **R** and is freely available on GitHub at https://github.com/prabhakarlab/scConsensus.

Keywords: ScRNA-seq, Clustering, Cell type annotation, Consensus method

Background

Since the first single cell experiment was published in 2009 [1], single cell RNA sequencing (scRNA-seq) has become the quasi-standard for transcriptomic profiling of heterogeneous data sets. In contrast to bulk RNA-sequencing, scRNA-seq is able to elucidate transcriptomic heterogeneity at an unmatched resolution and thus allows downstream analyses to be performed in a cell-type-specific manner, easily. This has been proven to

be especially important for instance in case-control studies or in studying tumor heterogeneity [2]. Nowadays, due to advances in experimental technologies, more than 1 million single cell transcriptomes can be profiled with high-throughput microfluidic systems. Scalable and robust computational frameworks are required to analyse such highly complex single cell data sets.

The clustering of single cells for annotation of cell types is a major step in this analysis. There are two methodologies that are commonly applied to cluster and annotate cell types: (1) unsupervised clustering followed by cluster annotation using marker genes [3] and (2) supervised approaches that use reference data sets to either cluster cells [4] or to classify cells into cell types [5].

A wide variety of methods exist to conduct unsupervised clustering, with each method using different distance metrics, feature sets and model assumptions. The graph-based clustering method Seurat [6] and its Python counterpart Scanpy [7] are the most prevalent ones. In addition, numerous methods based on hierarchical [8], density-based [9] and k-means clustering [10] are commonly used in the field. Kiselev et al. [3] provide an extensive overview on unsupervised clustering approaches and discuss different methodologies in detail. Importantly, they conclude that there is currently no method available that can robustly be applied to any kind of scRNA-seq data set, as method performance can be influenced by the size of data sets, the number and the nature of sequenced cell types as well as by technical aspects, such as dropouts, sample quality and batch effects.

Unsupervised clustering methods have been especially useful for the discovery of novel cell types. However, the marker-based annotation is a burden for researchers as it is a time-consuming and labour-intensive task. Also, manual, marker-based annotation can be prone to noise and dropout effects. Furthermore, different research groups tend to use different sets of marker genes to annotate clusters, rendering results to be less comparable across different laboratories.

To overcome these limitations, supervised cell type assignment and clustering approaches were proposed. The major advantages of supervised clustering over unsupervised clustering are its robustness to batch effects and its reproducibility. This has been shown to be beneficial for the integrative analysis of different data sets [4]. A comprehensive review and benchmarking of 22 methods for supervised cell type classification is provided by [5]. While they found that several methods achieve high accuracy in cell type identification, they also point out certain caveats: several sub-populations of CD4+ and CD8+ T cells could not be accurately identified in their experiments. [5] traced this back to inappropriate and/or missing marker genes for these cell types in the reference data sets used by some of the methods tested. This exposes a vulnerability of supervised clustering and classification methods—the reference data sets impose a constraint on the cell types that can be detected by the method. Aside from this strong dependence on reference data, another general observation made was that the accuracy of cell type assignments decreases with an increasing number of cells and an increased pairwise similarity between them. Furthermore, clustering methods that do not allow for cells to be annotated as *Unkown*, in case they do not match any of the reference cell types, are more prone to making erroneous predictions.

In summary, despite the obvious importance of cell type identification in scRNA-seq data analysis, the single-cell community has yet to converge on one cell typing methodology [3]. Due to the diverse merits and demerits of the numerous clustering approaches, this is unlikely to happen in the near future. However, as both unsupervised and supervised approaches have their distinct advantages, it is desirable to leverage the best of both to improve the clustering of single-cell data. As exemplified in Additional file 1: Figure S1 using FACS-sorted Peripheral Blood Mononuclear Cells (PBMC) scRNA-seq data from [11], both supervised and unsupervised approaches deliver unique insights into the cell type composition of the data set. Specifically, the supervised RCA [4] is able to detect different progenitor sub-types, whereas Seurat is better able to determine T-cell sub-types. Therefore, a more informative annotation could be achieved by combining the two clustering results.

Inspired by the consensus approach used in the unsupervised clustering method SC3, which resulted in improved clustering results for small data sets compared to graph-based approaches [3, 10], we propose scConsensus, a computational framework in **R** to obtain a consensus set of clusters based on at least two different clustering results.

Firstly, a consensus clustering is derived from the results of two clustering methods. This consensus clustering represents cell groupings derived from both clustering results, thus incorporating information from both inputs. Details on how this consensus clustering is generated are provided in "Workflow of scConsensus" section.

Secondly, the resulting consensus clusters are refined by re-clustering the cells using the union of consensus-cluster-specific differentially expressed genes (DEG) (Fig. 1) as features. Each initial consensus cluster is compared in a pair-wise manner with every other cluster to maximise inter-cluster distance with respect to strong marker genes.

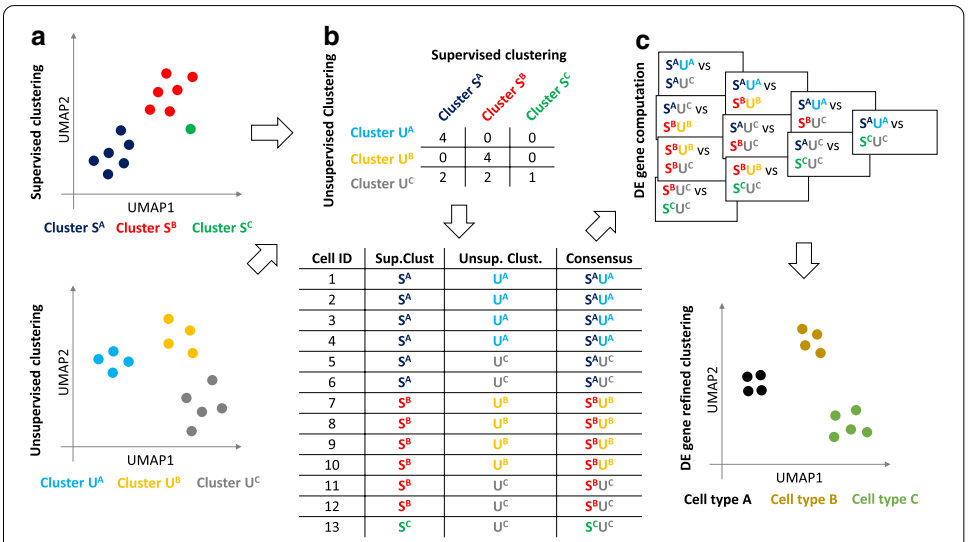

Fig. 1 The scConsensus workflow. **a** The scConsensus workflow considers two independent cell cluster annotations obtained from any pair of supervised and unsupervised clustering methods. **b** A contingency table is generated to elucidate the overlap of the annotations on the single cell level. A consensus labeling is generated using either an automated method or manual curation by the user. **c** DE genes are computed between all pairs of consensus clusters. Those DE genes are used to re-cluster the data. The refined clusters thus obtained can be annotated with cell type labels

Thereby, the separation of distinct cell types will improve, whereas clusters representing identical cell types not exhibiting distinct markers, will be merged together. This process can be seamlessly applied in an iterative fashion to combine more than two clustering results.

Here, we illustrate the applicability of the scConsensus workflow by integrating cluster results from the widely used Seurat package [6] and Scran [12], with those from the supervised methods RCA [4] and SingleR [13].

Methods

Data sets used

In total, we used five 10X CITE-Seq scRNA-seq data sets. Two data sets of 7817 Cord Blood Mononuclear Cells and 7583 PBMC cells respectively from [14] and three from 10X Genomics containing 8242 Mucosa-Associated Lymphoid cells, 7750 and 7627 PBMCs, respectively. Additionally, we downloaded FACS-sorted PBMC scRNA-seq data generated by [11] for CD14+ Monocytes, CD19+ B Cells, CD34+ Cells, CD4+ Helper T Cells, CD4+/CD25+ Regulatory T Cells, CD4+/CD45RA+/CD25- Naive T cells, CD4+/CD45RO+ Memory T Cells CD56+ Natural Killer Cells, CD8+ Cytotoxic T cells and CD8+/CD45RA+ Naive T Cells from the 10X website. Further details and download links are provided in Additional file 1: Table S1. Table 1 provides acronyms used in the remainder of the paper. Details on processing of the FACS sorted PBMC data are provided in Additional file 1: Note 3.

Data pre-processing and initial clustering

We used RCA (version 1.0) for supervised and Seurat (version 3.1.0) for unsupervised clustering (Fig. 1a). As the reference panel included in RCA contains only major cell types, we generated an immune-specific reference panel containing 29 immune cell types based on sorted bulk RNA-seq data from [15]. Details on the generation of this reference panel are provided in Additional file 1: Note 1.

All data pre-processing was conducted using the Seurat **R**-package. After filtering cells using a lower and upper bound for the *Number of Detected Genes (NODG)* and an upper bound for *mitochondrial rate*, we filtered out genes that are not expressed in at least 100 cells. Data set specific QC metrics are provided in Additional file 1:

Table 1 Overview on the number of cells contained in each considered scRNA-seq data set as well as on the acronyms used throughout this article

Dataset	Acronym	# cells
Cord Blood 10X	CBMC	7817
Peripheral Blood Drop-Seq	PBMC Drop-Seq	7583
Mucosa-Associated Lymphoid Tissue 10X	MALT	8242
Peripheral Blood 10X	PBMC	7750
Peripheral Blood 10X-VDJ	PBMC-VDJ	7627
PBMCs FACS	PBMC-FACS	25,389

Table S2. Note that we did not apply a threshold on the *Number of Unique Molecular Identifiers*. **R**-code is available in Additional file 1: Note 2.

Workflow of scConsensus

scConsensus takes the supervised and unsupervised clustering results as input and performs the following two major steps:

1. Generation of consensus annotation using a contingency table consolidating the results from both clustering inputs,
2. Refinement of the consensus cluster labels by re-clustering cells using DE genes.

The entire pipeline is visualized in Fig. 1.

Generating a consensus clustering

First, we use the *table* function in **R** to construct a contingency table (Fig. 1b). Each value in the contingency table refers to the extent of overlap between the clusters, measured in terms of number of cells.

scConsensus provides an automated method to obtain a consensus set of cluster labels $\mathcal{C}$. Starting with the clustering that has a larger number of clusters, referred to as $\mathcal{L}$, scConsensus determines whether there are any possible sub-clusters that are missed by $\mathcal{L}$. To do so, we determine for each cluster $l \in \mathcal{L}$ the percentage of overlap for the clustering with fewer clusters ($\mathcal{F}$) in terms of cell numbers: $|l \cap f|$. By default, we consider any cluster f that has an overlap $\geq 10\%$ with cluster l as a sub-cluster of cluster l, and then assign a new label to the overlapping cells as a combination of l and f. For cells in a cluster $l \in \mathcal{L}$ with an overlap $< 10\%$ to any cluster $f \in \mathcal{F}$, the original label will be retained. We note that the overlap threshold can be changed by the user. For instance by setting it to 0, each cell will obtain a label based on both considered clustering results $\mathcal{F}$ and $\mathcal{L}$. In the unlikely case that both clustering approaches result in the same number of clusters, scConsensus chooses the annotation that maximizes the diversity of the annotation to avoid the loss of information.

In addition to the automated consensus generation and for refinement of the latter, scConsensus provides the user with means to perform a manual cluster consolidation. This approach is especially well-suited for expert users who have a good understanding of cell types that are expected to occur in the analysed data sets.

Refinement by re-clustering cells on DE genes

Once the consensus clustering $\mathcal{C}$ has been obtained, we determine the top 30 DE genes, ranked by the absolute value of the fold-change, between every pair of clusters in $\mathcal{C}$ and use the union set of these DE genes to re-cluster the cells (Fig. 1c). Note that the number of DE genes is a user parameter and can be changed. Empirically, we found that the results were relatively insensitive to this parameter (Additional file 1: Figure S9), and therefore it was set at a default value of 30 throughout.Typically, for UMI data, we use the Wilcoxon test to determine the statistical significance (q-value ≤ 0.1) of differential expression and couple that with a fold-change threshold (absolute log fold-change ≥ 0.5) to select differentially

expressed genes. Upon DE gene selection, Principal Component Analysis (PCA) [16] is performed to reduce the dimensionality of the data using the DE genes as features. The number of principal components (PCs) to be used can be selected using an elbow plot. For the datasets used here, we found 15 PCs to be a conservative estimate that consistently explains majority of the variance in the data (Additional file 1: Figure S10). We then construct a cell-cell distance matrix in PC space to cluster cells using Ward's agglomerative hierarchical clustering approach [17].

Sequential merging of multiple clustering methods

scConsensus can be generalized to merge three or more methods sequentially. The merging of clustering results is conducted sequentially, with the consensus of 2 clustering results used as the input to merge with the third, and the output of this pairwise merge then merged with the fourth clustering, and so on. This process is repeated for all the clusterings provided by the user. By default, the input clusterings are arranged in decreasing order of the number of clusters.

Clustering of antibody tags to derive a ground truth for CITE-Seq data

We used antibody-derived tags (ADTs) in the CITE-Seq data for cell type identification by clustering cells using Seurat. The raw antibody data was normalized using the Centered Log Ratio (CLR) [18] transformation method, and the normalized data was centered and scaled to mean zero and unit variance. Dimension reduction was performed using PCA. The cell clusters were determined using Seurat's default graph-based clustering. More details, along with the source code used to cluster the data, are available in Additional file 1: Note 2.

Since these cluster labels were derived solely using ADTs, they provide an unbiased ground truth to benchmark the performance of scConsensus on scRNA-seq data. For each antibody-derived cluster, we identified the top 30 DE genes (in scRNA-seq data) that are positively up-regulated in each ADT cluster when compared to all other cells using the Seurat FindAllMarkers function. The union set of these DE genes was used for dimensionality reduction using PCA to 15 PCs for each data set and a cell-cell distance matrix was constructed using the Euclidean distance between cells in this PC space. This distance matrix was used for Silhouette Index computation to measure cluster separation.

Metrics for assessment of clustering quality

Normalized Mutual Information (NMI) to compare cluster labels

The Normalized Mutual Information (NMI) determines the agreement between any two sets of cluster labels $\mathcal{C}$ and $\mathcal{C}'$. We compute $NMI(\mathcal{C}, \mathcal{C}')$ between $\mathcal{C}$ and $\mathcal{C}'$ as

$$NMI(\mathcal{C}, \mathcal{C}') = \frac{[H(\mathcal{C}) + H(\mathcal{C}') - H(\mathcal{C}\mathcal{C}')]}{\max(H(\mathcal{C}), H(\mathcal{C}'))}, \tag{1}$$

where $H(\mathcal{C})$ is the entropy of the clustering $\mathcal{C}$ (see Chapter 5 of [19] for more information on entropy as a measure of clustering quality). The closer the NMI is to 1.0, the better is the agreement between the two clustering results.

Assessment of cluster quality using bootstrapping

We used both (1) Cosine Similarity $cs_{x,y}$ [20] and (2) Pearson correlation $r_{x,y}$ to compute pairwise cell-cell similarities for any pair of single cells (x, y) within a cluster c according to:

$$cs_{x,y} = \frac{\sum\limits_{g\in\mathcal{G}} x_g y_g}{\sqrt{\sum\limits_{g\in\mathcal{G}} x_g^2}\sqrt{\sum\limits_{g\in\mathcal{G}} y_g^2}}, \tag{2}$$

$$r_{x,y} = \frac{\sum\limits_{g\in\mathcal{G}} (x_g - \hat{x})(y_g - \hat{y})}{\sqrt{\sum\limits_{g\in\mathcal{G}} (x_g - \hat{x})^2}\sqrt{\sum\limits_{g\in\mathcal{G}} (y_g - \hat{y})^2}}. \tag{3}$$

To avoid biases introduced by the feature spaces of the different clustering approaches, both metrics are calculated in the original gene-expression space $\mathcal{G}$ where x_g represents the expression of gene g in cell x and y_g represents the expression of gene g in cell y, respectively. We apply two cut-offs on $\mathcal{G}$ with respect to the variance of gene-expression (0.5 and 1), thereby neglecting genes that are not likely able to distinguish different clusters from each other. Using bootstrapping, we select 100 genes 100 times from the considered gene-expression space $\mathcal{G}$ and compute the mean cosine similarity cs_c^i as well as the the mean Pearson correlation r_c^i for each cluster $c \in \mathcal{C}$ in each iteration i:

$$cs_c^i = \frac{1}{|c|}\sum_{(x,y)\in c} cs_{x,y}, \tag{4}$$

$$r_c^i = \frac{1}{|c|}\sum_{(x,y)\in c} r_{x,y}. \tag{5}$$

The scores cs_c and r_c are computed for all considered data sets and all three clustering approaches, scConsensus, Seurat and RCA. The closer cs_c and r_c are to 1.0, the more similar are the cells within their respective clusters. Statistical significance is assessed using a one-sided Wilcoxon–Mann–Whitney test.

Testing accuracy of cell type assignment on FACS-sorted data

Using the FACS labels as our ground truth cell type assignment, we computed the F1-score of cell type identification to demonstrate the improvement scConsensus achieves over its input clustering results by Seurat and RCA. The F1-score for each cell type t is defined as the harmonic mean of precision ($Pre(t)$) and recall ($Rec(t)$) computed for cell type t. In other words,

$$F1(t) = 2\frac{Pre(t)Rec(t)}{Pre(t) + Rec(t)}, \tag{6}$$

$$Pre(t) = \frac{TP(t)}{TP(t) + FP(t)}, \tag{7}$$

$$Rec(t) = \frac{TP(t)}{TP(t) + FN(t)}. \tag{8}$$

Here, a *TP* is defined as correct cell type assignment, a *FP* refers to a mislabelling of a cell as being cell type t and a *FN* is a cell whose true identity is t according to the FACS data but the cell was labelled differently.

Visualizing scRNA-seq data using UMAP

To visually inspect the scConsensus results, we compute DE genes between every pair of ground-truth clusters and use the union set of those DE genes as the features for PCA. Next, we use the Uniform Manifold Approximation and Projection (UMAP) dimension reduction technique [21] to visualize the embedding of the cells in PCA space in two dimensions.

Results

scConsensus: a hybrid approach for clustering single cell data

scConsensus is a general **R** framework offering a workflow to combine results of two different clustering approaches. Briefly, scConsensus is a two-step approach. First, scConsensus creates a consensus clustering using the Cartesian product of two input clustering results. Next, scConsensus computes the DE genes between all pairs of consensus clusters. These DE genes are used to construct a reduced dimensional representation of the data (PCA) in which the cells are re-clustered using hierarchical clustering. The scConsensus pipeline is depicted in Fig. 1.

scConsensus produces clusters that are more consistent with antibody-derived clusters

We used the Antibody-derived Tag (ADT) signal of the five considered CITE-seq data sets to generate a ground truth clustering for all considered samples (Fig. 2a). Next, we compute all differentially expressed (DE) genes between the antibody based clusters using the scRNA-seq component of the data. As shown in Fig. 2b (Additional file 1: Fig. S2), the expression of DE genes is cluster-specific, thereby showing that the

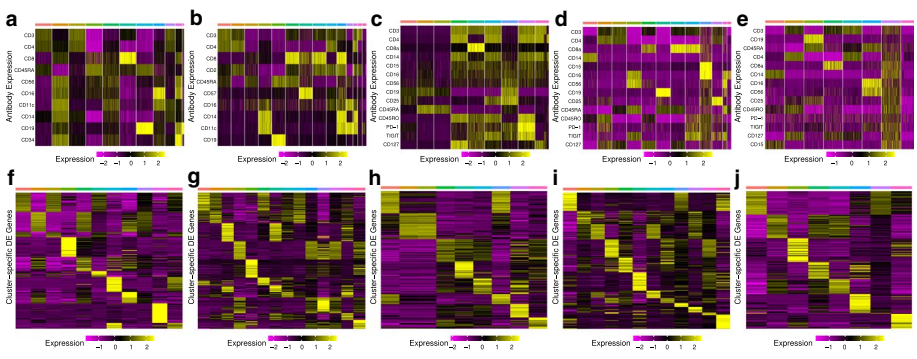

Fig. 2 Antibody-derived ground truth for CITE-Seq data. **a**–**e** Cluster-specific antibody signal per cell across five CITE-Seq data sets. **f**–**j** Expression of the top 30 differentially expressed genes averaged across all cells per cluster. (a, f) CBMC, (b, g) PBMC Drop-Seq, (c, h) MALT, (d, i) PBMC, (e, j) PBMC-VDJ

antibody-derived clusters are separable in gene expression space. Therefore, these DE genes are used as a feature set to evaluate the different clustering strategies.

Here, we assessed the agreement of the Scran, SingleR, Seurat and RCA, and their pairwise scConsensus results with the antibody-based single-cell clusters in terms of Normalized Mutual Information (NMI), a score quantifying similarity with respect to the cluster labels. In most cases, we observed that using scConsensus to combine a clustering result with one other method improved its NMI score. Further, in 4 out of 5 datasets, we observed a greater performance improvement when one supervised and one unsupervised method were combined, as compared to when two supervised or two unsupervised methods were combined (Fig. 3).

For a visual inspection of these clusters, we provide UMAPs visualizing the clustering results in the ground truth feature space based on DE genes computed between ADT clusters, with cells being colored according to the cluster labels provided by one of the tested clustering methods (Additional file 1: Figs. S5–S8). We compared the PBMC data set clustering results from Seurat, RCA, and scConsensus using the combination of Seurat and RCA (which was most frequently the best performing combination in Fig. 3). By visually comparing the UMAPs, we find for instance that Seurat cluster 3 (Fig. 4b), corresponds to the two antibody clusters 4 and 7 (Fig. 4a). In contrast to the unsupervised results, this separation can be seen in the supervised RCA clustering (Fig. 4c) and is correctly reflected in the unified clustering by scConsensus (Fig. 4d). Another illustration for the performance of scConsensus can be found in the supervised clusters 3, 4, 9, and 12 (Fig. 4c), which are largely overlapping. In the ADT cluster space, the corresponding cells should form only one cluster (Fig. 4a). Here scConsensus picks up the cluster information provided by Seurat (Fig. 4b), which reflects the ADT labels more accurately (Fig. 4d). These visual examples indicate the capability of

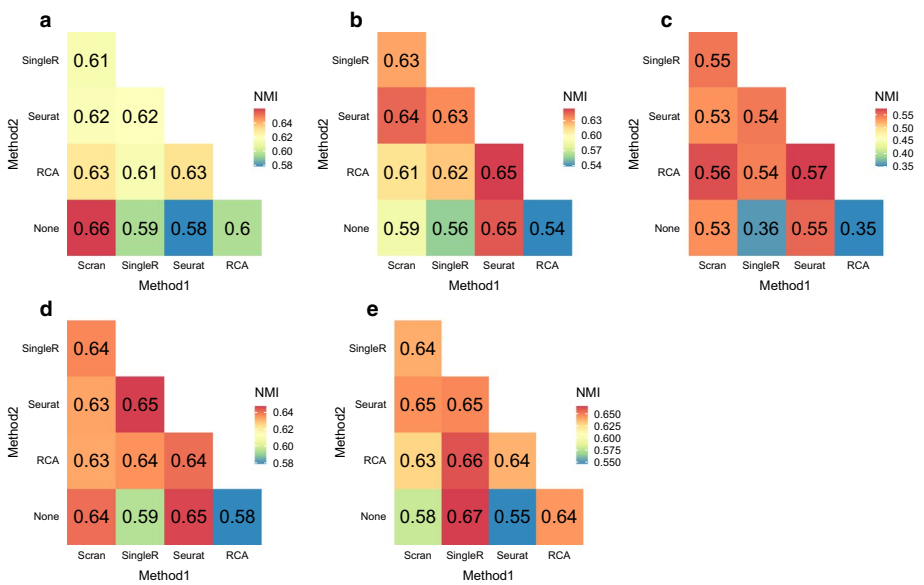

Fig. 3 Normalized Mutual Information (NMI) of antibody-derived ground truth with pairwise combinations of Scran, SingleR, Seurat and RCA clustering results. **a**–**e** Pair-wise combinations on the five CITE-Seq datasets: (a) CBMC, (b) PBMC Drop-Seq, (c) MALT, (d) PBMC, (e) PBMC-VDJ. "None" refers to no combination i.e. the clustering method's output was directly used to compute NMI

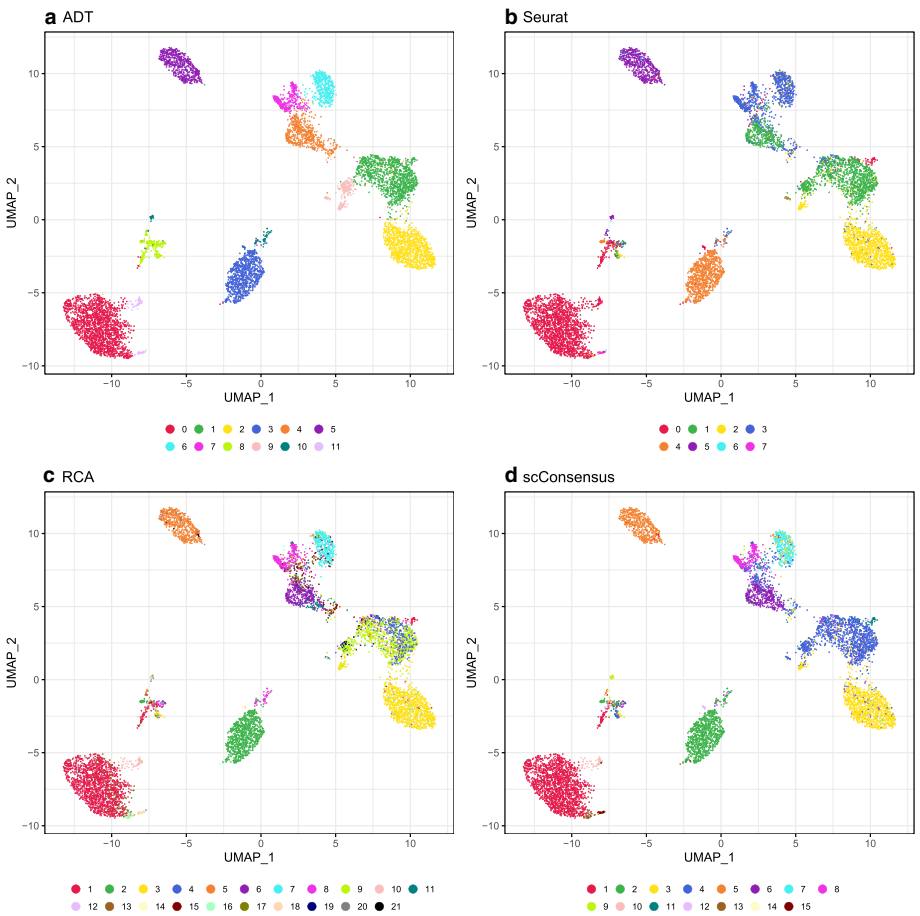

Fig. 4 ADT-based clustering of the PBMC data set. **a**–**d** UMAPs anchored in DE gene space colored by cluster IDs obtained from **a** ADT data, **b** Seurat clusters, **c** RCA and **d** scConsensus

scConsensus to adequately merge supervised and unsupervised clustering results leading to a more appropriate clustering. Similar examples can be found for the other data sets (CBMC, PBMC Drop-Seq, MALT and PBMC-VDJ) in Additional file 1: Figs. S5–S8.

In addition to the NMI, we assessed the performance of scConsensus in yet another complementary fashion. We quantified the quality of clusters in terms of within-cluster similarity in gene-expression space using both Cosine similarity and Pearson correlation. Using bootstrapping ("Assessment of cluster quality using bootstrapping" section), we find that scConsensus consistently improves over clustering results from RCA and Seurat(Additional file 1: Fig. S3 and Additional file 1: Fig S4) supporting the benchmarking using NMI. While the advantage of this comparisons is that it is free from biases introduced through antibodies and cluster method specific feature spaces, one can argue that using all genes as a basis for comparison is not ideal either. However, paired with bootstrapping, it is one of the fairest and most unbiased comparisons possible. A similar approach has been taken previously by [22] to compare the expression profiles of CD4+ T-cells using bulk RNA-seq data. Analogously to the NMI comparison, the number of resulting clusters also does not correlated to our performance estimates using Cosine similarity and Pearson correlation.

Merging more than two clustering methods is not beneficial

Using Scran, SingleR, Seurat and RCA, we demonstrated scConsensus' ability to sequentially merge up-to 3 clustering results. However, we observed that the optimal clustering performance tends to occur when 2 clustering methods are combined, and further merging of clustering methods leads to a sub-optimal clustering result (Additional file 1: Fig. S11).

scConsensus accurately reproduces FACS-sorted PBMC cell type labels

Using data from [11], we clustered cells using Seurat and RCA, as the combination of these methods performed well in the benchmarking presented above. After annotating the clusters, we provided scConsensus with the two clustering results as inputs and computed the F1-score ("Testing accuracy of cell type assignment on FACS-sorted data" section) of cell type assignment using the FACS labels as ground truth.

Figure 5a shows the mean F1-score for cell type assignment using scConsensus, Seurat and RCA, with scConsensus achieving the highest score. Fig. 5b depicts the F1 score in a cell type specific fashion. Figure 5 shows the visualization of the various clustering results using the FACS labels, Seurat, RCA and scConsensus. A striking observation is that CD4 T Helper cells could neither be captured by RCA nor by Seurat, and hence also not by scConsensus. Fig. 5b also illustrates that scConsensus does not hamper with and can even slightly further improve the already reliable detection of B cells, CD14+ Monocytes, CD34+ cells (Progenitors) and Natural Killer (NK) cells even compared to RCA and Seurat. Importantly, scConsensus is able to isolate a cluster of Regulatory T cells (T Regs) that was not detected by Seurat but was pinpointed through RCA (Fig. 5b). The scConsensus approach extended that cluster leading to an F1-score of 0.6 for T Regs. However, the cluster refinement using DE genes lead not only to an improved result for T Regs and CD4 T-Memory cells, but it also resulted in a slight drop in performance of scConsensus compared to the best performing method for CD4+ and CD8+ T-Naive as well as CD8+ T-Cytotoxic cells. As indicated by a UMAP representation colored by the FACS labels (Fig. 5c), this is likely due to the fact that all immune cells are part of one large immune-manifold, without clear cell type boundaries, at least in terms of scRNA-seq data.

Another example for the applicability of scConsensus is the accurate annotation of a small cluster to the left of the CD14 Monocytes cluster (Fig. 5c). Using Seurat, the majority of those cells are annotated as stem cells, while a minority are annotated as CD14 Monocytes (Fig. 5d). RCA annotates these cells exclusively as CD14+ Monocytes (Fig. 5e). However, according to FACS data (Fig. 5c) these cells are actually CD34+ (Progenitor) cells, which is well reflected by scConsensus (Fig. 5f).

Overall, these examples demonstrate the power of combining reference-based clustering with unsupervised clustering and showcase the applicability of scConsensus to identify and cluster even closely-related sub-types in scRNA-seq data.

Discussion

Many different approaches have been proposed to solve the single-cell clustering problem, in both unsupervised [3] and supervised [5] ways. However, all approaches have their own advantages and disadvantages and do not necessarily lead to similar results,

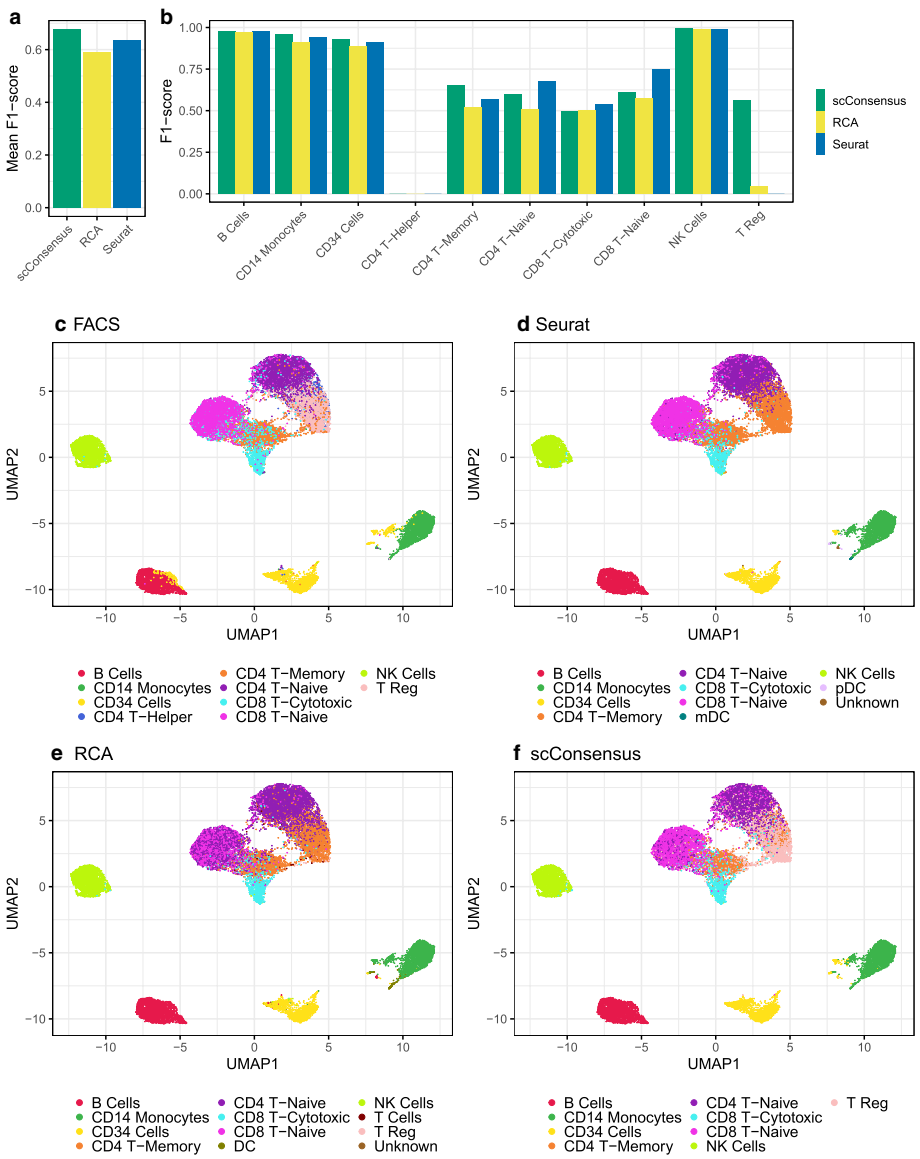

Fig. 5 Performance assessment of cell type assignment on FACS sorted PBMC data. **a** Mean F1-score across all cell types. **b** F1-score per cell type. **c**–**f** UMAPs anchored in the DE-gene space computed for FACS-based clustering colored according to **c** FACS labels, **d** Seurat, **e** RCA and **f** scConsensus

as exemplified in Additional file 1: Fig. 1. While benchmarking scConsensus we also found that there is no consistent ranking between the tested supervised and unsupervised approaches. On some data sets, e.g. the FACS sorted PBMC data shown in Fig. 5, the unsupervised Seurat performs better than the supervised RCA, while the latter achieves better performance than Seurat on the CITE-seq data sets (Fig. 3). In fact, this observation stresses that there is no ideal approach for clustering and therefore also motivates the development of a consensus clustering approach. With scConsensus we propose a computational strategy to find a consensus clustering that provides the best possible cell type separation for a single-cell data set.

scConsensus builds on known intuition about single-cell RNA sequencing data, i.e. homogeneous cell types will have consistent differentially expressed marker genes when compared with other cell types. scConsensus computes DE gene calls in a pairwise fashion, that is comparing a distinct cluster against all others. Together with a constant number of DE genes considered per cluster, scConsensus gives equal weight to rare sub-types, which may otherwise get absorbed into larger clusters in other clustering approaches. We have demonstrated this using a FACS sorted PBMC data set and the loss of a cluster containing regulatory T-cells in Seurat compared to scConsensus.

A major feature of the scConsensus workflow is its flexibility - it can help leverage information from any two clustering results. Here, we focus on Seurat and RCA , two complementary methods for clustering and cell type identification in scRNA-seq data. However, the intuition behind scConsensus can be extended to any two clustering approaches. For example, even using the same data, unsupervised graph-based clustering and unsupervised hierarchical clustering can lead to very different cell groupings. Upon encountering this issue, users typically tend to pick the clustering result that agrees best with their domain knowledge, while completely ignoring the information provided by the other clustering. Thus, we propose scConsensus as a valuable, easy and robust solution to the problem of integrating different clustering results to achieve a more informative clustering.

Conclusions

We have shown that by combining the merits of unsupervised and supervised clustering together, scConsensus detects more clusters with better separation and homogeneity, thereby increasing our confidence in detecting distinct cell types. As scConsensus is a general strategy to combine clustering methods, it is apparent that scConsensus is not restricted to scRNA-seq data alone. Any multidimensional single-cell assay whose cell clusters can be separated by differential features can leverage the functionality of our approach. For instance, for single-cell ATAC sequencing data, there are various clustering approaches available that lead to different clustering results [23]. scConsensus could be used out of the box to consolidate these clustering results and provide a single, unified clustering result. Therefore, we believe that the clustering strategy proposed by scConsensus is a valuable contribution to the computational biologist's toolbox for the analysis of single-cell data.

Availability and requirements

Project name scConsensus

Project home page https://github.com/prabhakarlab/

Operating system(s) Windows, Linux, Mac-OS

Programming language **R** ($\geq$ 3.6)

Other requirements **R** packages: mclust, circlize, reshape2, flashClust, calibrate, WGCNA, edgeR, circlize, ComplexHeatmap, cluster, aricode

License MIT Any restrictions to use by non-academics: None

Abbreviations
DE: Differentially expressed; DEG: Differentially expressed genes; scRNA-seq: Single cell RNA sequencing; PBMC: Peripheral blood mononuclear cells; NODG: Number of detected genes; PCA: Principal component analysis; PC: Principal component; ADTs: Antibody-derived tags; CLR: Centered log ratio; NMI: Normalized mutual information; NK: Natural killer cells; T Reg: Regulatory T cells.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04028-4.

Additional file 1. Supplementary Figures and Tables.

Acknowledgements
The authors thank all members of the Prabhakar lab for feedback on the manuscript. This publication is part of the Human Cell Atlas—www.humancellatlas.org/publications.

Authors' contributions
BR, WS, JP, MAH and FS were involved in developing, testing and benchmarking scConsensus. NAR and JT developed the immune reference panel. BR and FS wrote the manuscript. BR, FS and SP edited and reviewed the manuscript. All authors have read and approved the manuscript.

Funding
Salaries for BR and FS have been paid by Grant# CDAP201703-172-76-00056 from the Agency for Science, Technology and Research (A*STAR), Singapore. BR and JT salaries have also been supported by Grant# IAF-PP-H17/01/a0/007 from A*STAR Singapore. Computational resources and NAR's salary were funded by Grant# IAF-PP-H18/01/a0/020 from A*STAR Singapore. The funding bodies did not influence the design of the study, did not impact collection, analysis, and interpretation of data and did not influence the writing of the manuscript.

Availability of data and materials
All data generated or analysed during this study are included in this published article and on Zenodo (https://doi.org/10.5281/zenodo.3637700).

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Laboratory of Systems Biology and Data Analytics, Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore. [2] Department of Medicine, School of Medicine, National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119077, Singapore.

Received: 12 May 2020 Accepted: 15 February 2021
Published online: 12 April 2021

References

1. Tang F, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6(5):377–82.
2. Lawson DA, et al. Tumour heterogeneity and metastasis at single-cell resolution. Nat Cell Biol. 2018;20(12):1349–60.
3. Kiselev VY, et al. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019;20(5):273–82.
4. Li H, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet. 2017;49(5):708–18.
5. Abdelaal T, et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol. 2019;20(1):194.
6. Butler A, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20.
7. Wolf FA, et al. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15.
8. Lin P, et al. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017;18(1):59.
9. Ester M, Kriegel H-P, Sander J, Xu X, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol. 96. p. 226–31; 1996.
10. Kiselev V, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14(5):483–6.
11. Zheng GX, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
12. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. F1000Research. 2016;5:2122.

13. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163–72.
14. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14(9):865.
15. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carre C, Burdin N, Visan L, Ceccarelli M, Poidinger M, et al. RNA-seq signatures normalized by MRNA abundance allow absolute deconvolution of human immune cell types. Cell Rep. 2019;26(6):1627–40.
16. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometr Intell Lab Syst. 1987;2(1–3):37–52.
17. Ward JH Jr. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58(301):236–44.
18. Aitchison J. The statistical analysis of compositional data. J R Stat Soc Ser B (Methodol). 1982;44(2):139–60.
19. Schütze H, Manning CD, Raghavan P. Introduction to Information Retrieval, vol. 39. Cambridge: Cambridge University Press; 2008.
20. Pesquita C, et al. Semantic similarity in biomedical ontologies. PLoS Comput Biol. 2009;5(7):1000443.
21. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
22. Durek P, Nordstrom K, et al. Epigenomic profiling of human CD4+ T cells supports a linear differentiation model and highlights molecular regulators of memory development. Immunity. 2016;45:1148–61.
23. Chen H, et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biol. 2019;20(1):241.

Publisher's Note