Check for updates

RESEARCH ARTICLE

Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data [version 1; peer review: 3 approved with reservations]

J. Javier Diaz-Mejia ¹⁻³, Elaine C. Meng³, Alexander R. Pico ⁴,
Sonya A. MacParland ⁵⁻⁷, Troy Ketela¹, Trevor J. Pugh^{1,8,9}, Gary D. Bader ^{2,10},
John H. Morris ³

¹Princess Margaret Cancer Centre, University Health Network, Toronto, ON, M5G 2M9, Canada

²The Donnelly Centre, University of Toronto, Toronto, ON, M5S 3E1, Canada

³Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA, 94143, USA

⁴Gladstone Institutes, San Francisco, CA, 95158, USA

⁵Multi-Organ Transplant Program, Toronto General Hospital Research Institute, Toronto, ON, M5G 2C4, Canada

⁶Department of Immunology, University of Toronto, Toronto, ON, M5S 1A8, Canada

⁷Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, M5G 1L7, Canada

⁸Department of Medical Biophysics, University of Toronto, Toronto, ON, M5G 1L7, Canada

⁹Ontario Institute for Cancer Research, Toronto, ON, M5G 0A3, Canada

¹⁰Department of Molecular Genetics, University of Toronto, Toronto, ON, M5G 1A8, Canada

V1 First published: 15 Mar 2019, 8(ISCB Comm J):296
<https://doi.org/10.12688/f1000research.18490.1>

Second version: 27 Aug 2019, 8(ISCB Comm J):296
<https://doi.org/10.12688/f1000research.18490.2>

Latest published: 14 Oct 2019, 8(ISCB Comm J):296
<https://doi.org/10.12688/f1000research.18490.3>

Abstract

Background: Identification of cell type subpopulations from complex cell mixtures using single-cell RNA-sequencing (scRNA-seq) data includes automated computational steps like data normalization, dimensionality reduction and cell clustering. However, assigning cell type labels to cell clusters is still conducted manually by most researchers, resulting in limited documentation, low reproducibility and uncontrolled vocabularies. Two bottlenecks to automating this task are the scarcity of reference cell type gene expression signatures and the fact that some dedicated methods are available only as web servers with limited cell type gene expression signatures.

Methods: In this study, we benchmarked four methods (CIBERSORT, GSEA, GSVA, and ORA) for the task of assigning cell type labels to cell clusters from scRNA-seq data. We used scRNA-seq datasets from liver, peripheral blood mononuclear cells and retinal neurons for which reference cell type gene expression signatures were available.

Results: Our results show that, in general, all four methods show a high performance in the task as evaluated by receiver operating characteristic curve analysis (average area under the curve (AUC) =

Open Peer Review

Approval Status

	1	2	3
version 3			
(revision)			
14 Oct 2019	view		
version 2			
(revision)			
27 Aug 2019	view	view	
version 1			
15 Mar 2019	view	view	view

1. **Saskia Freytag** , Harry Perkins Institute of Medical Research, Nedlands, Australia

2. **Jimmy Tsz Hang Lee**, Wellcome Sanger Institute, Hixton, UK

Tallulah Andrews , Wellcome Sanger

0.94, $sd = 0.036$), whereas precision-recall curve analyses show a wide variation depending on the method and dataset (average AUC = 0.53, $sd = 0.24$).

Conclusions: CIBERSORT and GSVA were the top two performers. Additionally, GSVA was the fastest of the four methods and was more robust in cell type gene expression signature subsampling simulations. We provide an extensible framework to evaluate other methods and datasets at https://github.com/jdime/scRNAseq_cell_cluster_labeling.

Keywords

single cell, RNA-seq, scRNA-seq, bioinformatics, benchmark, evaluation, labeling, cell type

This article is included in the **Bioinformatics** gateway.

Institute, Hinxton, UK

3. **Lindsay Cowell**, University of Texas

Southwestern Medical Center, Dallas, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: John H. Morris (scooter@cgl.ucsf.edu)

Author roles: **Diaz-Mejia JJ:** Conceptualization, Data Curation, Formal Analysis, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Meng EC:** Data Curation, Writing – Original Draft Preparation, Writing – Review & Editing; **Pico AR:** Conceptualization; **MacParland SA:** Data Curation, Writing – Original Draft Preparation; **Ketela T:** Writing – Original Draft Preparation; **Pugh TJ:** Writing – Original Draft Preparation; **Bader GD:** Conceptualization, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Morris JH:** Conceptualization, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: JJDM, ECM, ARP, and JHM are funded by grant number 2018-183120 from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. ARP, GDB and JHM are supported by the National Resource for Network Biology, P41GM103504 (NIGMS).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 Diaz-Mejia JJ *et al.* This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Diaz-Mejia JJ, Meng EC, Pico AR *et al.* **Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data [version 1; peer review: 3 approved with reservations]** F1000Research 2019, 8(ISCB Comm J):296 <https://doi.org/10.12688/f1000research.18490.1>

First published: 15 Mar 2019, 8(ISCB Comm J):296 <https://doi.org/10.12688/f1000research.18490.1>

Introduction

During the last five years a number of single-cell sequencing technologies have been developed to identify cell subpopulations from complex cell mixtures (Bakken *et al.*, 2017). For instance, recent advances in single-cell RNA-sequencing (scRNA-seq) enable the simultaneous measurement of expression levels of hundreds to thousands of genes across hundreds to thousands of individual cells. The resulting expression matrices of genes by cells are used (see below) to identify cell subpopulations with characteristic gene expression profiles and other biological properties (i.e. cell types).

A typical computational pipeline to process scRNA-seq data involves the following steps: i) quality control of sequencing reads, ii) mapping reads against a reference transcriptome, iii) normalization of mapped reads to correct batch effects and remove contaminants, iv) data dimensionality reduction with principal component analysis or alternative approaches, v) clustering of cells using principal component values, vi) detection of genes differentially expressed between clusters, vii) visualization of cell clusters in t-SNE or alternative plots, and viii) assignment of cell type labels to cell clusters. A number of computational tools, including Cell Ranger (Zheng *et al.*, 2017a) and Seurat (Butler *et al.*, 2018), allow automation of steps i to vii (Duò *et al.*, 2018;

Freytag *et al.*, 2018; Innes & Bader, 2018). However, assignment of cell type labels to cell clusters is still conducted manually by most researchers. The typical procedure involves manual inspection of the genes expressed in a cluster, combined with a detailed literature search to identify if any of those genes are known gene expression markers for cell types of interest. This manual approach has several caveats, including limited documentation and low reproducibility of cell type gene marker selection, use of uncontrolled and non-ontological vocabularies for cell type labels, and it can be time-consuming. For these reasons, computational tools that allow researchers to systematically, reproducibly and quickly assign cell type labels to cell clusters derived from scRNA-seq experiments are needed.

In this study we used three scRNA-seq datasets from liver cells (MacParland *et al.*, 2018), peripheral blood mononuclear cells (PBMCs) (Zheng *et al.*, 2017a) and retinal neurons (Shekhar *et al.*, 2016b) (Table 1) to compare four methods that can be used for assigning cell type labels to cell clusters: CIBERSORT (Newman *et al.*, 2015b), GSEA (Subramanian *et al.*, 2005), GSVA (Hänzelmann *et al.*, 2013) and ORA (Fisher, 1935; Goeman & Bühlmann, 2007) (Table 2). We chose these four methods to represent different categories of algorithms, ranging from first-generation enrichment analysis (ORA) to second-generation

Table 1. scRNA-seq datasets used in this study.

Dataset Name	Description of scRNA-seq dataset	Number of genes in $\tilde{E}_{xy}$	Number of cells	Number of cell clusters	Number of cell type signatures	Reference
Liver	10X Chromium sample from liver cells from five human donors	20,007	8,444	20	10	(MacParland <i>et al.</i> , 2018)
PBMCs	10X Chromium sample from peripheral blood mononuclear cells from a human donor	17,786	68,579	12	22	(Zheng <i>et al.</i> , 2017a)
Retinal neurons	Drop-seq sample from retinal bipolar neurons from healthy mice	13,166	27,499	18	15	(Shekhar <i>et al.</i> , 2016b)

Table 2. Cell cluster labeling methods compared in this study. Details of the four methods compared in this study, and their computing times to classify cell clusters of indicated datasets. (b) refers to CIBERSORT 'binary' analysis mode, (c) refers to CIBERSORT 'continuous' analysis mode.

Acronym	Version	Name	Language	Computing time (s)			Reference
				Liver	PBMCs	Retinal neurons	
CIBERSORT	1.01	Cell type Identification by Estimating Relative Subsets of RNA Transcripts	R and Java	(b) = 44	(b) = 169 (c) = 700	(b) = 36	(Newman <i>et al.</i> , 2015b)
GSEA	3.0	Gene Set Enrichment Analysis	Java	93	78	98	(Subramanian <i>et al.</i> , 2005)
GSVA	1.30	Gene Set Variation Analysis	R	1.2	0.9	0.73	(Hänzelmann <i>et al.</i> , 2013)
ORA	R(3.5.1)	Over-representation Analysis	R	4	3	4	(Fisher, 1935; Goeman & Bühlmann, 2007)

approaches (GSEA and GSVA) and machine learning tools (CIBERSORT). Although ORA and GSEA were not originally developed to process RNA-seq data, they have been extensively used in transcriptomic studies for gene set enrichment analyses. GSVA was developed to analyze microarray and bulk RNA-seq data, and CIBERSORT was developed to estimate abundances of cell types in mixed cell populations from bulk RNA-seq data. We adapted all four methods to assign cell type labels to cell clusters from scRNA-seq data based on known sets of cell type marker genes. We evaluated these methods using two types of inputs: a matrix with the average expression of each gene x from all the cells in each cell cluster y ($\bar{E}_{xy}$) from scRNA-seq measurements, which we assume corresponds to the profile of a cell type or state, and known cell type gene expression signatures, represented as gene sets or continuous gene expression profiles (Figures 1A-C).

CIBERSORT uses gene expression profiles as training data for a machine learning algorithm to estimate abundances of known cell types in a mixed cell population and was originally developed to identify composition of known immune cell types in bulk RNA-seq sample measurements. In our evaluation, we used $\bar{E}_{xy}$ matrices instead of bulk RNA-seq data. GSEA uses a Kolmogorov–Smirnov (KS) like statistic to determine whether a gene set shows statistically significant, concordant differences between biological states. It was originally developed to analyze microarray gene expression data and has been applied to multiple genomic data types. GSVA transforms a gene by sample matrix to a gene set by sample matrix, and evaluates gene set enrichment for each sample. Like GSEA, GSVA uses a KS like statistic but GSVA bypasses explicitly modeling phenotypes within the enrichment scoring step. GSVA was originally developed to process microarray and bulk RNA-seq measurements. ORA uses the Fisher’s exact test to detect an overrepresentation of members of a gene set in a subsample of highly expressed genes, compared against both the total number of gene set members and the total number of genes measured in the sample.

Methods explicitly developed to assign cell type labels to cell clusters of scRNA-seq data have been reported (Alavi *et al.*, 2018; Alquicira-Hernandez *et al.*, 2018; Crow *et al.*, 2018). However, to our knowledge they are in beta, or implemented as web-servers to process cell types for which we could not find reference cell type annotations (Figure 1F) that we would require to include in our evaluation. For this reason, we included only the four methods described above, and we provide execution and benchmark scripts that will be useful to extend our comparisons to other methods in the future.

Methods

Generation of cell cluster average gene expression matrices ($\bar{E}_{xy}$)

For the liver dataset (MacParland *et al.*, 2018) (NCBI GEO: GSE115469) we followed the authors’ reported procedure to obtain cell clusters, and obtained the $\bar{E}_{xy}$ matrix for each cluster using the function `AverageExpression(use.raw = T)` from *Seurat* v2 (Butler *et al.*, 2018). For the PBMCs dataset (Zheng *et al.*, 2017a),

Fresh 68k PBMCs DonorA gene expression matrix files were obtained from 10X (Zheng *et al.*, 2017b) (NCBI Sequence Read Archive: SRX1723926). Normalization, data dimensionality reduction, cell clustering and $\bar{E}_{xy}$ matrix calculations were conducted with *Seurat* with the following functions: `FilterCells(low.thresholds = 200, -Inf, high.thresholds = 0.05, 10000)`; `FindClusters(reduction.type = “pca”, dims.use = 1:10, resolution = 0.4)`; `AverageExpression(use.raw = T)`. For the retinal neurons dataset (Shekhar *et al.*, 2016b) (NCBI GEO: GSE81905) the gene expression matrix and cell cluster assignments were obtained from (Shekhar *et al.*, 2016a) and the $\bar{E}_{xy}$ matrix calculation was conducted with `AverageExpression(use.raw = T)` from *Seurat*.

Generation of cell type gene expression signatures

A gene expression signature is defined simply as a set of genes characteristically and detectably expressed in a cell type. These are typically inferred from small-scale experiments that need to be manually identified in the literature, or by comparing the transcriptome of a given cell type against all other available cell type gene expression profiles, usually from the same experiment. The liver cell type gene set signatures were manually curated by us (author S.A.M.) and were originally used to manually annotate cell types in the liver dataset (MacParland *et al.*, 2018). We provide these gene sets on Zenodo (Diaz-Mejia, 2019a). For the PBMC dataset, we used a blood cell type gene expression profile signature compiled by the CIBERSORT developers called LM22, containing 547 genes and 22 cell types (Newman *et al.*, 2015a). Reference cell type assignments to the PBMCs by fluorescence-activated cell sorting (FACS) were obtained from (Zheng *et al.*, 2017c). The PBMC cell clusters we obtained with *Seurat* were mapped using cell barcode identifiers against the FACS assignments, and cell type names were manually matched to the LM22 signature. For the retinal neuron dataset (Shekhar *et al.*, 2016b), known cell type markers reported by the authors were used as cell type gene set signatures.

CIBERSORT requires as input a cell type gene expression signature in the form of gene expression profiles (i.e. a matrix of genes in rows and cell types in columns). For the PBMC dataset, we used two versions of the LM22 signature for CIBERSORT. First, we used the original LM22 signature (Newman *et al.*, 2015b) with continuous valued gene expression measurements, which we called CIBERSORT ‘continuous’. Second, for each cell type of the LM22 signature, a value of ‘1’ was assigned to 5% of genes with highest expression values in their column or a value of ‘0’ otherwise, and we called this approach CIBERSORT ‘binary’. The same 5% of genes was used to create cell type gene set signatures as inputs for GSEA, GSVA and ORA. For the liver dataset, we transformed the cell type gene set signature into a binary matrix of genes in rows and cell types in columns for CIBERSORT ‘binary’ analysis mode. To do this, each gene included in each cell type gene set m was assigned a value of ‘1’ in the column corresponding to m in the matrix, whereas other genes absent in m but present in other cell type gene sets were assigned a value of ‘0’. Similarly, for the retinal neuron dataset the ‘previously known markers’ for bipolar cell types provided in Table S2 of Shekhar *et al.* (2016b) were transformed into a binary matrix of genes by cell types for CIBERSORT ‘binary’ analysis.

Figure 1. Schematic of a process to benchmark automated cell type detection methods. Two inputs are needed by automated cell type detection methods (**A–C**): (**A**) a matrix with the average expression of each gene x for each cell cluster y ($\bar{E}_{xy}$). (**B, C**) cell type gene marker signatures can be provided as either gene sets (lists of gene identifiers, **B**) or numeric gene expression profiles (**C**). Gene sets can be manually compiled from literature and are used for methods like GSEA, GSVA or ORA. Whereas gene-expression profiles are measurements from microarrays, bulk- or single-cell RNA-sequencing (scRNA-seq) experiments, and are used by methods like CIBERSORT. (**D, E**) Automated cell type detection methods produce a matrix of cell type likelihoods for each cell cluster. (**F**) Some authors of scRNA-seq studies assign cell type labels manually to cell clusters using empirical expertise or orthogonal experiments such as fluorescence activated cell sorting. These assignments can be used as references to benchmark automated cell type detections. (**G**) Top cell type predictions (red rectangles in **E**) are contrasted against annotation references (**F**) to assess the performance of cell type detection methods by receiver operating characteristic (ROC) curve and precision-recall (PR) curve analyses. (**H**) Robustness of cell type detection methods can be analyzed by gradually subsampling gene markers from cell type gene expression signatures (**B** or **C**) and repeating procedures of (**D–G**) to obtain distributions of the area under the curve (AUC) for ROC (ROC AUC) and PR (PR AUC) curves, which are shown as violin plots. We hypothesized that some detection methods may be more robust than others to the proportion of gene markers subsampled from cell type gene expression signatures.

Generation of subsampled cell type gene expression signatures and area under the curve (AUC) violin plots

Cell type gene set signatures (Figure 1B) were subsampled by randomly removing between 10 and ~99% of genes from each signature in increments of 10%, keeping a minimum of one gene. Each subsampling of gene sets was transformed into a binary matrix of genes by cell types for CIBERSORT ‘binary’ as indicated above. Cell type gene expression profile signatures (Figure 1C) were subsampled in two stages: first we selected the top 5% highest expressed genes for each cell type, then we randomly replaced the gene expression value of 10 to 100% of those genes from each cell type, in increments of 10%, by the minimum value of the cell type column. This resulted in subsampled gene expression profile signatures with identical size to the original profile signatures, but with values of the top highly expressed genes randomly replaced by the minimum score of each cell type. For percentage values between 10 to 100%, 1,000 subsampling replicates were generated for each cell type gene expression signature, and each replicate was processed as indicated by Figures 1D-G. Violin plots were used to show the resulting ROC and PR AUC distributions.

Transformation of tested methods’ enrichment metrics for ROC and PR analyses

The enrichment scores (ES) from CIBERSORT and GSVA were directly used as ranks for the benchmark comparisons against gold standard references, whereas the P-values from GSEA and ORA were first $-\log_{10}$ transformed and the resulting values were used as ranks for the benchmark analyses. For ORA, the universe of genes used was the intersection of genes present in the cell type gene expression signature and the $\bar{E}_{xy}$ matrix of each dataset. All methods were implemented locally using Java, R and Perl (Table 2), using the following libraries and programs: for CIBERSORT we used CIBERSORT.jar v1.01 and R(Rserve) 1.8.6, for GSEA we used gsea-3.0.jar, for GSVA we used R(GSVA) v1.30 and R(GSA) v1.3, and for ORA we used R(fisher.test) R v3.5.1.

Method computing time benchmark

We implemented wrapper scripts to execute each of the four methods tested, including a stopwatch to time the cell type prediction task. Other tasks, such as input and output preparation, were excluded from computing time values reported in Table 2. All computing time measurements were made using a 3.1-GHz Intel Core i5 CPU with 2 cores and 16 GB RAM.

The scripts used to run and benchmark cell type labeling methods described in this study are available on [GitHub](#) and archived at [Zenodo](#) (Diaz-Mejia, 2019b). An earlier version of this article can be found on [bioRxiv](https://doi.org/10.1101/562082) (<https://doi.org/10.1101/562082>).

Results

Benchmark of cell cluster labeling methods

We benchmarked the performance and computing time of four cell type labeling methods, namely: CIBERSORT, GSVA, GSEA and ORA (Table 2), using average gene expression profiles of scRNA-seq cell clusters and known cell type gene expression signatures. We used three scRNA-seq datasets: liver cells

(MacParland *et al.*, 2018), PBMCs (Zheng *et al.*, 2017a) and retinal neurons (Shekhar *et al.*, 2016b) (Table 1). Each method used two inputs: an $\bar{E}_{xy}$ matrix with the average gene expression for each cell cluster (Figure 1A) and a cell type gene expression signature, represented as either a gene set or a gene expression profile. Three of the four methods tested (GSVA, GSEA and ORA) used cell type gene set signatures (Figure 1B), whereas CIBERSORT used cell type gene expression profiles either with continuous or binarized values (Figure 1C). Each method produced a matrix of cell type predictions (Figure 1D, E) which was compared to manually annotated cell type references (Figure 1F) to conduct receiver operating characteristic (ROC) and precision-recall (PR) curve analyses (Figure 1G). The robustness of each method was assessed by randomly subsampling 10% to 100% of the genes from the cell type gene expression signatures and repeating the cell type detection and ROC and PR curve analyses for each subsample (Figure 1H).

ROC curve analysis

In general, we observed that all four methods showed high ROC AUC values for all three analyzed scRNA-seq datasets. An average ROC AUC = 0.97 was found for the liver dataset (Figure 2A), average ROC AUC = 0.92 for the PBMC dataset (Figure 2B) and average ROC AUC = 0.94 for the retinal neuron dataset (Figure 2C). Since CIBERSORT takes as input a cell type gene expression signature in the form of gene expression profiles (Figure 1C), and the only available signatures for the liver and retinal neuron datasets were in the form of gene sets, we transformed the gene sets into binary matrices and used them as inputs for CIBERSORT (Methods). Notably, the binary matrix approach, which we called CIBERSORT ‘binary’, produced the highest performance among all tested methods for the liver (ROC AUC = 1, Figure 2A) and retinal neurons datasets (ROC AUC = 0.95, Figure 2C). The CIBERSORT ‘binary’ approach performance was almost identical to that of the original LM22 cell type gene expression signature with continuous values, which we called CIBERSORT ‘continuous’, for the PBMC dataset (ROC AUC = 0.91 and 0.92, Figure 2B). GSVA was the top performer using the PBMC dataset (ROC AUC = 0.95, Figure 2B), closely followed by GSEA (ROC AUC = 0.94) and the two versions of CIBERSORT (‘binary’ ROC AUC = 0.92 and ‘continuous’ ROC AUC = 0.91), while ORA’s performance was slightly lower (ROC AUC = 0.86) (Figure 2B).

The analysis of ROC AUC robustness showed that, in general, performance of all methods decayed as a function of removing genes from cell type gene expression signatures. However, GSVA tolerated removal of up to 90% of the genes from the PBMC signature to maintain ROC AUCs ≥ 0.8 . ORA tolerated removal of up to 60% of genes at the same ROC AUC cutoff (Figure 3B), whereas GSEA and the two versions of CIBERSORT gave ROC AUCs < 0.8 when $\geq 30\%$ of the genes were removed from the PBMC cell type signatures. For the liver dataset, GSVA and GSEA tolerated removal of up to 60% of genes from the liver signature to maintain ROC AUCs ≥ 0.8 , whereas CIBERSORT ‘binary’ and ORA tolerated removal of up to 50% of the genes at the same ROC AUC cutoff (Figure 3A). For the retinal neuron dataset,

Figure 2. Performance analysis of automated cell type detection methods using single-cell RNA-seq (scRNA-seq) data. Receiver operating characteristic (ROC) and precision-recall (PR) curve analyses of four automated cell type detection methods (CIBERSORT, GSEA, GSVA and ORA) (Table 2) using three scRNA-seq datasets (Table 1). ROC curve analyses for datasets from: (A) human liver cells, (B) human PBMCs, and (C) mouse retinal neurons. PR curve analyses for the same datasets: (D) human liver cells, (E) human peripheral blood mononuclear cells (PBMCs), and (F) mouse retinal neurons. The ROC area under the curve (AUC) and PR AUC are shown for each method using each dataset. For the PBMCs dataset, two analyses were conducted with CIBERSORT, one using the original LM22 cell type gene expression signature with continuous gene expression values, that we called CIBERSORT 'continuous' (CIBER(c)), and another where the LM22 profiles were thresholded and binarized, which we called CIBERSORT 'binary' (CIBER(b), see Methods). The same thresholded signature was used to create cell type gene sets for GSEA, GSVA and ORA (Methods). For the liver and retinal neuron datasets, only gene set signatures were available and they were transformed into binary matrices for CIBERSORT 'binary' (CIBER(b)).

GSVA and ORA tolerated removal of up to 50% of the genes from the signature to maintain ROC AUCs ≥ 0.8 , whereas GSEA and CIBERSORT 'binary' tolerated removal of 30% and 20%, respectively, for the same ROC AUC cutoff (Figure 3C).

Precision-Recall curve analysis

When benchmarking the four methods compared in this study, we classified each cell cluster positively into a single cell type and negatively into the remaining cell types of their corresponding dataset signature. This produced a skewed distribution with few positive predictions and several negative predictions. To ameliorate this imbalance, we used PR curve analyses in addition to ROC curve analyses. In general, the PR AUCs were smaller than the ROC AUCs (Figure 2, top vs. bottom panels). Some methods clearly separated from the rest using PR curve analyses.

For instance, GSEA showed the lowest PR AUC values for both the liver and retinal neurons datasets (PR AUCs = 0.51 and 0.28), compared with CIBERSORT (PR AUCs = 0.98 and 0.5), ORA (PR AUCs = 0.90 and 0.53), and GSVA (PR AUC = 0.89 and 0.56) (Figure 2D, F). GSEA also displayed the lowest AUC in the ROC curve analyses for the liver and retinal neurons datasets, and the performance differences between GSEA and the other methods were more pronounced using PR curve analyses. In contrast, the two versions of CIBERSORT for the PBMC dataset ranked very close to the other three methods using ROC curve analyses (all ROC AUCs were > 0.9 , Figure 2B), but they were relatively low using PR curve analyses (CIBERSORT 'continuous' PR AUC = 0.22 and CIBERSORT 'binary' PR AUC = 0.24), compared with GSVA (PR AUC = 0.56), ORA (PR AUC = 0.42) and GSEA (PR AUC = 0.34) (Figure 2E).

Figure 3. Receiver operating characteristic (ROC) area under the curve (AUC) robustness analysis of automated cell type detection methods. The cell type gene expression signatures used for ROC curve analyses in Figure 2 were randomly subsampled 1,000 times, keeping 10 to 100% of genes from the original signatures each time. Automated cell type detection was repeated for each subsample and violin plots representing the distribution of resulting ROC AUCs are shown for datasets from: (A) human liver cells, (B) human peripheral blood mononuclear cells (PBMCs), and (C) mouse retinal neurons. For the PBMC dataset, two analyses were conducted with either the original LM22 cell type gene expression signature with continuous gene expression values (CIBER(c)) or with a thresholded and binarized version (CIBER(b)). For the liver and retinal neuron datasets only binary matrices for CIBER(b) were used.

The PR AUC robustness analysis showed that all methods' performance decayed as a function of removing genes from cell type gene expression signatures. Interestingly, using the liver dataset all four methods showed higher PR AUCs than for the PBMC and retinal neuron datasets (Figure 4A-C). GSVA and ORA tolerated removal of up to 60% of genes from the liver dataset signatures to maintain PR AUCs ≥ 0.5 . CIBERSORT 'binary' tolerated removal of 50% of genes for the same PR AUC cutoff (Figure 4A), whereas GSEA PR AUCs were < 0.5 using either the full liver cell type signature or any subsampling of it. For the retinal neuron dataset, CIBERSORT 'binary', GSVA and ORA tolerated removal of up to 20% of the genes from the signature to maintain average PR AUC ≥ 0.5 , whereas for GSEA the average was < 0.5 at any fraction of genes in the signature. For the PBMC dataset, GSVA was the only method showing PR AUC > 0.5 with the full signature (Figure 2E) and it tolerated removal of up to 20% of genes from the signature to maintain average PR AUC > 0.5 (Figure 4B).

Computing time benchmark

As shown in Table 2, the computing times of method implementations varied from 0.73 s for GSVA processing the retinal

neurons dataset, up to 700 s for CIBERSORT 'continuous' processing the PBMC dataset. For all three datasets, GSVA was the fastest method to process cell type classification tasks. ORA ranked second with computing times between 3 and 5 times longer than GSVA. GSEA showed computing times between 77 and 134 times longer than GSVA, and CIBERSORT showed computing times between 37 and 777 times longer than GSVA. The size of the cell type gene expression signatures used for CIBERSORT influenced the speed of the classification task. For CIBERSORT 'continuous' we used the original LM22 signature, which contained 547 genes for the PBMC dataset, whereas the thresholded binary matrix used for CIBERSORT 'binary' had 248 genes, and it took 169 s, or 24% of the time that took CIBERSORT 'continuous' for the same task. For comparison, we created a second 'continuous' signature by restricting the original LM22 signature to the 248 genes present in the thresholded binary matrix. This 'reduced continuous' signature approach showed a performance (ROC AUC = 0.92, PR AUC = 0.32) which was similar to the full CIBERSORT 'continuous' (ROC AUC = 0.92, PR AUC = 0.24) and 'binary' modes (ROC AUC = 0.91, PR AUC = 0.22), and the computing time was reduced substantially to 189 s, or 27% of the time that took CIBERSORT 'continuous' for the same task.

Figure 4. Precision-recall (PR) area under the curve (AUC) robustness analysis of automated cell type detection methods. The same procedure described in Figure 3 for ROC AUCs was used here for PR AUCs. Please see Figure 3 legend for details.

Discussion

The size and volume of scRNA-seq datasets are continually increasing, and several methods are available to normalize scRNA-seq measurements and cluster cells. In contrast, cell type labeling of cell clusters is still conducted manually by most researchers. This is in part due to a scarcity of reference cell type gene expression signatures and also because most methods to address this challenge are only available via web servers with limited number of cell types (Alavi *et al.*, 2018; Alquicira-Hernandez *et al.*, 2018; Crow *et al.*, 2018), making it difficult for users to adapt them for their needs. In this study we used three scRNA-seq datasets to benchmark four methods that can address these challenges. Although three of the four tested methods (GSEA, GSVA and ORA) were not explicitly developed to identify cell types, their extensive use in gene set enrichment tasks and their wide portability motivated us to test them as cell type classifiers. CIBERSORT is implemented both as a webserver and a local distribution that can be licensed by developers, allowing users to benchmark it with relatively low programmatic effort.

In general, our results show that for the three scRNA-seq datasets tested (liver, PBMCs and retinal neurons) all four tested methods achieved good performance by ROC curve analyses. However, ROC curves tend to overestimate methods' performance

when the ratio of positive to negative predictions is highly skewed. For this reason, we decided to also conduct PR curve analyses. GSVA was consistently one of the top performers by both ROC and PR curve analyses for the three datasets, and its performance was more robust in analyses where we subsampled genes from cell type gene expression signatures. This is particularly important at this stage of the scRNA-seq field, as only limited information on cell type gene expression signatures is available. Notably, despite its relative simplicity, ORA showed a performance comparable to GSVA. CIBERSORT's performance was good, particularly for the liver dataset by both ROC and PR analyses, albeit lower than that of GSVA or ORA in the PBMC dataset, and it was comparable using the retinal neuron dataset. CIBERSORT's computing times were orders of magnitude higher than those of GSVA and ORA. Our results showed that CIBERSORT 'binary' performed as well as CIBERSORT 'continuous' by both ROC and PR curve analyses and used only one quarter of the computing time. In the present implementation, GSEA performed worse than the other three methods, particularly in the PR curve analyses.

The size of current publicly available scRNA-seq datasets is currently typically on the order of thousands of cells clustered into dozens of cell clusters. In our tests, each of the four tested

methods completed the cell type prediction tasks in seconds or minutes. However, bigger datasets from the Human Cell Atlas (Rozenblatt-Rosen *et al.*, 2017) and other sources are expected to have millions of cells (e.g. 1.3 million brain cell from E18 mice, NCBI GEO: GSE93421) grouped into thousands of clusters, for which the fastest method implementations will be preferred. In this sense, we found that GSVA is the best option since its computing time for the tested datasets was fastest (one to two orders of magnitude faster than GSEA and CIBERSORT). ORA also offers a good option for cell cluster labeling as its ROC and PR curve benchmarks were comparable to GSVA and its computing times were only 3 to 5 times longer than those of GSVA. One extra requirement for ORA compared with the other three methods is that the $\hat{E}_{xy}$ matrix profiles need to be thresholded. In this study we used an arbitrary cutoff, based on the overall distribution of gene expression values, but future analyses could evaluate iterative thresholding.

One of the limitations of this study is that we included only three scRNA-seq datasets (liver, PBMCs and retinal neurons). This was due to the lack of reference cell type annotations needed for the ROC and PR curve analyses. As more scRNA-seq datasets become available and authors provide gold standard annotations of their cell types, those annotations could be used as references to benchmark methods with other scRNA-seq datasets. This is exemplified by the LM22 signature, which was constructed by Newman *et al.* (2015b) from microarray gene expression measurements to predict cell types from bulk RNA-seq data, and we have shown here that LM22 could also be used to detect cell types from scRNA-seq data. Thus, in the future, we envision that methods to detect differentially expressed genes can be used as part of pipelines to produce cell type gene expression signatures. As with any classification task, researchers would need to control for circularity between training, test and validation cell-annotation data and also will need to evaluate generalizability.

One of the challenges that we faced while adapting the LM22 signature to detect cell types in the scRNA-seq cell clusters generated by Zheng *et al.* (2017a) was that, even though both datasets correspond to PBMCs, the granularity of their cell type labels was different. For instance, the LM22 signature contains six T-cell types, including three CD4+ (naïve, memory resting, and memory activated), follicular helper, regulatory and gamma delta, whereas the dataset of Zheng *et al.* (2017a) contained labels for four T-cell related cell types: CD4+/CD25 T Regulatory, CD4+/CD45RO+ Memory, CD4+/CD45RA+/CD25- Naïve T and CD4+ T Helper2. Thus, even though these two datasets both classify PBMCs, they cannot be easily related one-to-one. This could be addressed with an ontology analogous to the Gene Ontology (Ashburner *et al.*, 2000) but dedicated to cell type annotations (Bakken *et al.*, 2017; Bard *et al.*, 2005). Fortunately, the Cell Ontology is being developed for this purpose. This is particularly important as an increasing number of signatures are expected to arise from initiatives like the Human Cell Atlas (Rozenblatt-Rosen *et al.*, 2017).

Data availability

Underlying data

The datasets used in this study were processed from the below underlying source data:

Single-cell RNA-sequencing data from human liver cells. Accession number, GSE115469. <https://identifiers.org/geo/GSE115469>.

Single-cell RNA-sequencing data from human peripheral blood mononuclear cells. Accession number, SRX1723926. <https://identifiers.org/insdc.sra/SRX1723926>.

Single cell RNA-sequencing of retinal bipolar cells. Accession number, GSE81905. <https://identifiers.org/geo/GSE81905>.

Extended data

Zenodo: Supplementary data for “Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data”. <http://doi.org/10.5281/zenodo.2575050> (Diaz-Mejia, 2019a).

This project contains the three processed scRNA-seq datasets—from liver cells (MacParland *et al.*, 2018), peripheral blood mononuclear cells (Zheng *et al.*, 2017a) and retinal neurons (Shekhar *et al.*, 2016b)—examined in this study.

Software availability

R and Perl scripts used to run and benchmark cell type labeling methods available from: https://github.com/jdime/scRNAseq_cell_cluster_labeling.

Archived code at time of publication: <http://doi.org/10.5281/zenodo.2583161> (Diaz-Mejia, 2019b).

License: MIT license.

Grant information

JJDM, ECM, ARP, and JHM are funded by grant number 2018-183120 from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. ARP, GDB and JHM are supported by the National Resource for Network Biology, P41GM103504 (NIGMS).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We are thankful to Jeff Liu and Brendan Innes from the Bader lab for advice processing the liver dataset and implementing GSVA; to Danielle Croucher and Laura Richards from the Pugh lab for feedback collecting benchmark datasets; and to Rene Quevedo from the Pugh lab for help implementing R scripts.

References

- Alavi A, Ruffalo M, Parvangada A, *et al.*: **A web server for comparative analysis of single-cell RNA-seq data.** *Nat Commun.* 2018; **9**(1): 4768.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Alquicira-Hernandez J, Nguyen Q, Powell JE: **scPred: scPred: Cell type prediction at single-cell resolution.** *bioRxiv.* 2018.
[Publisher Full Text](#)
- Ashburner M, Ball CA, Blake JA, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet.* 2000; **25**(1): 25–29.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bakken T, Cowell L, Aevermann BD, *et al.*: **Cell type discovery and representation in the era of high-content single cell phenotyping.** *BMC Bioinformatics.* 2017; **18**(Suppl 17): 559.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bard J, Rhee SY, Ashburner M: **An ontology for cell types.** *Genome Biol.* 2005; **6**(2): R21.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Butler A, Hoffman P, Smibert P, *et al.*: **Integrating single-cell transcriptomic data across different conditions, technologies, and species.** *Nat Biotechnol.* 2018; **36**(5): 411–420.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Crow M, Paul A, Ballouz S, *et al.*: **Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor.** *Nat Commun.* 2018; **9**(1): 884.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Diaz-Mejia JJ: **Supplementary data for 'Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data' (Diaz-Mejia JJ, *et al.*, 2019).** 2019a; [Accessed February 21, 2019].
<http://www.doi.org/10.5281/zenodo.2575050>
- Diaz-Mejia JJ: **Supplementary code for "Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data" (Diaz-Mejia JJ *et al.*, 2019) (Version v1.0).** *Zenodo.* 2019b.
<http://www.doi.org/10.5281/zenodo.2583161>
- Duò A, Robinson MD, Sonesson C: **A systematic performance evaluation of clustering methods for single-cell RNA-seq data [version 1; referees: 2 approved with reservations].** *F1000Res.* 2018; **7**: 1141.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Fisher RA: **The Logic of Inductive Inference.** *J R Stat Soc.* 1935; **98**(1): 39–82.
[Publisher Full Text](#)
- Freytag S, Tian L, Lönnstedt I, *et al.*: **Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data [version 1; referees: 1 approved, 2 approved with reservations].** *F1000Res.* 2018; **7**: 1297.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Goeman JJ, Bühlmann P: **Analyzing gene expression data in terms of gene sets: methodological issues.** *Bioinformatics.* 2007; **23**(8): 980–987.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Hänzelmann S, Castelo R, Guinney J: **GSVA: gene set variation analysis for microarray and RNA-seq data.** *BMC Bioinformatics.* 2013; **14**: 7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Innes BT, Bader GD: **scClustViz – Single-cell RNAseq cluster assessment and visualization [version 1; referees: 2 approved with reservations].** *F1000Res.* 2018; **7**: 1522.
[Publisher Full Text](#)
- MacParland SA, Liu JC, Ma XZ, *et al.*: **Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations.** *Nat Commun.* 2018; **9**(1): 4383.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Newman AM, Liu CL, Green MR, *et al.*: **Robust enumeration of cell subsets from tissue expression profiles.** *LM22 signature.* 2015a.
[Reference Source](#)
- Newman AM, Liu CL, Green MR, *et al.*: **Robust enumeration of cell subsets from tissue expression profiles.** *Nat Methods.* 2015b; **12**(5): 453–457.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rozenblatt-Rosen O, Stubbington MJT, Regev A, *et al.*: **The Human Cell Atlas: from vision to reality.** *Nature.* 2017; **550**(7677): 451–453.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Shekhar K, Lapan SW, Whitney IE, *et al.*: **Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics.** 2016a.
[Reference Source](#)
- Shekhar K, Lapan SW, Whitney IE, *et al.*: **Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics.** *Cell.* 2016b; **166**(5): 1308–1323.e30.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Subramanian A, Tamayo P, Mootha VK, *et al.*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A.* 2005; **102**(43): 15545–15550.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zheng GX, Terry JM, Belgrader P, *et al.*: **Massively parallel digital transcriptional profiling of single cells.** *Nat Commun.* 2017a; **8**: 14049.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zheng GXY, Terry JM, Belgrader P, *et al.*: **Fresh 68k PBMCs (Donor A).** 2017b.
[Reference Source](#)
- Zheng GXY, Terry JM, Belgrader P, *et al.*: **Single Cell RNA-seq Secondary Analysis of 68k PBMCs.** 2017c.
[Reference Source](#)

Open Peer Review

Current Peer Review Status: ? ? ?

Version 1

Reviewer Report 01 April 2019

<https://doi.org/10.5256/f1000research.20232.r45813>

© 2019 Cowell L. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Lindsay Cowell

Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA

The authors address an important problem, which is the need for systematic and reproducible approaches for assigning cell type labels based on single cell transcriptome data. They use three data sets with gold standard cell type annotations available and compare the performance of four computational tools on these data sets. The authors measure performance using ROC curves and plots of precision versus recall. They also assess performance over subsamples of the data used as reference gene expression patterns for cell types (either cell type-specific gene sets or cell type-specific expression profiles). In general, they found that all four methods perform reasonably well, although ORA and GSVA perform more consistently well across the three data sets. I do have some questions about the details of how the work was done. The answers to these questions are important for interpreting the results, reproducing the work, or extending it to include additional tools.

1. Presumably the approach to creating the cell clusters, and how dense versus diffuse the clusters are, can have an impact on performance and confidence in the output?
2. How exactly were clusters mapped to cell types? From Figure 1E, it appears that each of the four tools generates a numerical vector for each cell type that contains a score for each cluster, presumably corresponding to the likelihood that that cluster is of the corresponding cell type.
 - Is a cluster always assigned to the cell type corresponding to its highest score? (presumably yes).
 - In the example, each cell type and each cluster has only a single high score with all other scores being very small. What is the distribution of scores typically? Do clusters sometimes have multiple high scores? Were ties ever observed?
 - Can multiple clusters map to the same cell type?
 - Must a cluster be assigned to a cell type? Or could some remain unassigned?
3. How were the performance curves generated? What parameter was varied?

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: computational immunology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 17 Aug 2019

J. Javier Diaz-Mejia

R3-Q1) Presumably the approach to creating the cell clusters, and how dense versus diffuse the clusters are, can have an impact on performance and confidence in the output?

R3-A1) We agree that cluster density and other structure in the data will likely impact automatic cluster annotation performance. Investigating the relationship between a given structure in the data (e.g. density vs. sparseness) and performance would require simulations that may not be realistic. Thus, we limited our analysis to published data with available gold standards. We have now added this point to the discussion.

R3-Q2)

[a] How exactly were clusters mapped to cell types? From Figure 1E, it appears that each of the four tools generates a numerical vector for each cell type that contains a score for each cluster, presumably corresponding to the likelihood that that cluster is of the corresponding cell type.
[b] Is a cluster always assigned to the cell type corresponding to its highest score? (presumably yes).

[c] In the example, each cell type and each cluster has only a single high score with all other scores being very small. What is the distribution of scores typically? Do clusters sometimes have multiple high scores? Were ties ever observed?

[d] Can multiple clusters map to the same cell type?

[e] Must a cluster be assigned to a cell type? Or could some remain unassigned?

R3-A2)

[a] Correct, each tool generates a numerical vector as the reviewer describes.

[b] Yes, a cluster is always assigned to the cell type corresponding to its highest score.

[c] In the methods that we compared, each cell cluster vs. each cell type receives only one score. As can be observed in our new Figure 6E, most cell clusters which were incorrectly classified (i.e. that were not the top-1 ranked prediction) still had top-ranks (ticker distribution in the violin plots closer to the top-1 ranks), which indicates that some clusters can have multiple high scores. We found that 118 out of all 1,276 (9.2%) cell cluster labeling predictions we ran showed ties in the top-score: 65 of the 118 ties (68%) corresponded to METANEIGHBOR 'binary', 24 (20%) to ORA, 15 (13%) to METANEIGHBOR 'continuous', 10 (8%) to GSEA, and 4 (3%) to GSVA. None of the CIBERSORT analyses showed ties.

[d] Yes, multiple clusters can map to the same cell type and this is particularly the case for the newly incorporated Tabula Muris dataset, where 130 cell clusters map to 53 cell types. This doesn't affect our evaluation because a method is not penalized for predicting that multiple clusters have the same cell type annotation.

[e] Yes, a cluster must be assigned a cell type in our case because all clusters have a cell type assignment in our gold standards. In the case of the newly incorporated PBMC-SeqWell data (Gierahn *et al.*, 2017), some of the cell clusters were labeled as 'Removed_' by the authors, and they didn't classify those clusters into cell types, thus we did not include these in our analysis.

As mentioned above in response to reviewers 1 and 2, we've updated the Methods section "Implementation of tested methods and transformation of enrichment metrics for ROC and PR analyses" to clarify all of these points.

R3-Q3) How were the performance curves generated? What parameter was varied?

R3-A3) As mentioned above in response to reviewers 1 and 2, we've updated the Methods section "Implementation of tested methods and transformation of enrichment metrics for ROC and PR analyses" to clarify this. For each dataset, we combine all cell type gene set prediction scores for a method across all clusters into one column and vary the prediction score threshold to compute the ROC and PR curves.

Competing Interests: No competing interests were disclosed.

Reviewer Report 22 March 2019

<https://doi.org/10.5256/f1000research.20232.r45811>

© 2019 Andrews T et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Jimmy Tsz Hang Lee

Wellcome Sanger Institute, Hinxton, UK

Tallulah Andrews

Wellcome Sanger Institute, Hinxton, UK

Diaz-Mejia et al. test the ability of four different algorithms to correctly annotate a set of clusters identified in single-cell RNA-seq data. They find that GSVA tends to be the most accurate and fastest method, interestingly they find ORA and GSVA are much more robust to small numbers of marker genes than GSEA or CIBERSORT. This is a very useful and timely study, as manual annotation of cell-types is currently the main bottleneck when analyzing single-cell RNA-seq data.

Comments:

1. It was unclear to me how the accuracy of the classification methods was evaluated. What was the gold standard truth used for each dataset? Were clusters assigned to (a) the single cell-type for which they had the greatest score or (b) all cell-types where their score exceeded some threshold, or (c) to the single cell-type for which they had the greatest score provided that score was above some threshold or another approach? This is crucial to interpreting the PR and ROC curves presented in the results.
Based on the first sentence of the "Precision-Recall curve analysis" section: I inferred you to be using method (c), but using such a method should not necessarily lead to recall values of 1 as clusters which are more similar to an incorrect cell-type than to the correct cell-type would never become true positives. Thus, I had inferred you to be using method (b) based on Figure 2. It would be very helpful to add a section to the Methods explaining precisely how the accuracy was evaluated.
2. In addition, I suggest adding figures/tables for the accuracy of each classification approach (% of clusters correctly assigned) when all clusters are simply assigned to the cell-type for which they have the highest score, since I expect this to be the most common approach users of these classifications would take.
3. The main weakness of the paper, as the authors admit, is the small number of datasets used to test the classification methods, particularly since the variability in performance between datasets was high. It would be useful to show reproducibility of the results in additional datasets.
We acknowledge identifying marker gene lists for many different tissues can be very time consuming, there are datasets similar to those the authors have already have markers for that they could use. E.g. mouse retina: Shekhar et al. 2016¹, PBMCs: Gierahn et al. 2017 (Seq-Well)². Alternatively, they could do cross-comparisons using the two mouse cell atlas (Tabula Muris³ and Mouse Cell Atlas⁴). Or use datasets such as Pollen et al., 2014⁵ where gold-standard cell-type identity is known by design.
4. The authors show that performance degrades when small numbers of marker genes are used by the classifiers. Is it the case that more marker genes is always better or does performance also degrade if too many genes are used?

References

1. Shekhar K, Lapan SW, Whitney IE, Tran NM, et al.: Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*. 2016; **166** (5): 1308-1323.e30 [PubMed Abstract](#) | [Publisher Full Text](#)

2. Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, et al.: Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods*. 2017; **14** (4): 395-398 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, et al.: Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*. **562** (7727): 367-372 [PubMed Abstract](#) | [Publisher Full Text](#)
4. Han X, Wang R, Zhou Y, Fei L, et al.: Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*. 2018; **172** (5): 1091-1107.e17 [PubMed Abstract](#) | [Publisher Full Text](#)
5. Pollen AA, Nowakowski TJ, Shuga J, Wang X, et al.: Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotechnol*. 2014; **32** (10): 1053-8 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, single-cell RNA-seq, clustering, network inference

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 17 Aug 2019

J. Javier Diaz-Mejia

R2-Q1) *It was unclear to me how the accuracy of the classification methods was evaluated. What was the gold standard truth used for each dataset? Were clusters assigned to (a) the single cell-type for which they had the greatest score or (b) all cell-types where their score exceeded some threshold, or (c) to the single cell-type for which they had the greatest score provided that score was above some threshold or another approach? This is crucial to interpreting the PR and ROC*

curves presented in the results.

Based on the first sentence of the "Precision-Recall curve analysis" section: I inferred you to be using method (c), but using such a method should not necessarily lead to recall values of 1 as clusters which are more similar to an incorrect cell-type than to the correct cell-type would never become true positives. Thus, I had inferred you to be using method (b) based on Figure 2. It would be very helpful to add a section to the Methods explaining precisely how the accuracy was evaluated.

R2-A1) Apologies for the confusion around this point. We have now clarified how the ROC and PR curves were computed in Figure 1 and the text, as described for reviewer 1, above. We combine all cell type gene set prediction scores for a method across all clusters into one column and vary the prediction score threshold to compute the ROC and PR curves. A cluster is only allowed to be correctly labeled using one cell type, as enforced by our gold standard cluster annotation data (the set of cell types an author used to label their given cell clusters). So this matches strategy (c).

R2-Q2) *In addition, I suggest adding figures/tables for the accuracy of each classification approach (% of clusters correctly assigned) when all clusters are simply assigned to the cell-type for which they have the highest score, since I expect this to be the most common approach users of these classifications would take.*

R2-A2) Percent of clusters correctly assigned is now included in Figure 6C and Supplementary Table 1. It is useful to have a range of performance indicators to capture different performance facets.

R2-Q3) *The main weakness of the paper, as the authors admit, is the small number of datasets used to test the classification methods, particularly since the variability in performance between datasets was high. It would be useful to show reproducibility of the results in additional datasets. We acknowledge identifying marker gene lists for many different tissues can be very time consuming, there are datasets similar to those the authors have already have markers for that they could use. E.g. mouse retina: Shekhar et al. 20161, PBMCs: Gierahn et al. 2017 (Seq-Well)2. Alternatively, they could do cross-comparisons using the two mouse cell atlas (Tabula Muris3 and Mouse Cell Atlas4). Or use datasets such as Pollen et al., 20145 where gold-standard cell-type identity is known by design.*

R2-A3) We thank the reviewer for suggesting these datasets. We already used Shekhar *et al.* 2016 in version 1 of our paper. In the current version, we added Gierahn *et al.* (2017) as the authors provide cell type labels for the cell clusters, and used the LM22 cell type signatures as input for the prediction methods. We also added the Tabula Muris dataset. We contacted one of the Tabula Muris authors (Angela Pisco), who kindly gave us access to a set of cell type signatures curated by experts on tissues of the Tabula Muris dataset. From the 20 tissues provided in the Tabula Muris data we could map 11 of them into the dataset of cell type signatures and that is what we used as 'Tabula Muris 11' in the current version of our paper. We also investigated the influence of cell type signatures using the PBMC datasets (10X and Seq-Well) using either the full LM22 signature database, that we call 'PBMC-22', or only the six cell types expected to occur in the data, that we call 'PBMC-6'. Altogether, we provide analysis using eight dataset variants, up from the three in our initial manuscript.

R2-Q4) *The authors show that performance degrades when small numbers of marker genes are used by the classifiers. Is it the case that more marker genes is always better or does performance also degrade if too many genes are used?*

R2-A4) In general, having more marker genes is better, but not always. We approached this question by examining the influence of the number of genes in each gene set (x-axis) and asking what rank does the corresponding gold standard positive receive (y-axis). As can be seen in Figure 6E, the most common scenario is that the fewer the number of genes in the signatures, the more chances that the prediction is incorrect (i.e. assigned a rank lower than the top-rank). However, there are a few exceptions, like ORA in the 11-20 genes bin, where we found more incorrect predictions than having 6-10 genes, or CIBERSORT, which had higher error rate in the 31-50 genes category, than in the 11-20 or 21-30 categories. Thus, it is possible to use too many genes, but it is not always clear how many genes this will be and the performance drop is not great for most of the cases we have data for. We have added this analysis to the paper.

Competing Interests: No competing interests were disclosed.

Reviewer Report 20 March 2019

<https://doi.org/10.5256/f1000research.20232.r45814>

© 2019 Freytag S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Saskia Freytag

Epigenetics and Genomics, Harry Perkins Institute of Medical Research, Nedlands, WA, Australia

Diaz-Mejia et al have produced a nice research article on assessing methods for assigning cluster labels to cell clusters from scRNA-seq. I think this work is of great importance, but I felt that some crucial cluster labelling methods were not compared. I hope the authors can update the article with some of the suggestions:

- The biggest suggestion for improvement is the choice of methods that the authors compare. The authors chose to adapt 4 methods originally developed for bulk RNA-seq in order to label clusters. While their approach is commendable, none of their adapted methods reflect the current standard practice in the field. Additionally, their claim that methods for cluster labelling in scRNA-seq are too immature or implemented as web-servers is not true. The website scRNA-tools.org lists 29 methods in this category. Many of these methods, such as scMCA, MetaNeighbour and scmap, are well-established and frequently used in the field. Furthermore, many of these methods recommend using annotated scRNA-seq datasets as references instead of bulk data. Hence, it would be great if the authors could include some of these tools in their analysis.

- I am confused as to which classifier parameter was varied in order to generate the ROCs. Were these comparable across the different methods?
- It would be interesting to see what the effect of varying the cluster resolution is to the ability of the methods to accurately label the populations. Do you obtain more diverse labelling when there are more clusters?
- LM22 is a great reference dataset, but recently a new dataset has become openly accessible. This dataset, generated by Monaco et al¹, characterizes 29 human immune cell types by RNA-seq and flow cytometry. It would be interesting to see if the use of this dataset leads to an improvement.
- I think it would be helpful for the reader if the authors could summarize their results. The sheer number of comparisons made, means that the reader can feel overwhelmed at the end. A figure summarizing the various results for each method in each dataset could help clarify the message.

Thank you for making your code publicly available.

References

1. Monaco G, Lee B, Xu W, Mustafah S, et al.: RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep.* 2019; **26** (6): 1627-1640.e7
[PubMed Abstract](#) | [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 17 Aug 2019

J. Javier Diaz-Mejia

R1-Q1) *The biggest suggestion for improvement is the choice of methods that the authors compare. The authors chose to adapt 4 methods originally developed for bulk RNA-seq in order to label clusters. While their approach is commendable, none of their adapted methods reflect the current standard practice in the field. Additionally, their claim that methods for cluster labelling in scRNA-seq are too immature or implemented as web-servers is not true. The website scRNA-tools.org lists 29 methods in this category. Many of these methods, such as scMCA, MetaNeighbour and scmap, are well-established and frequently used in the field. Furthermore, many of these methods recommend using annotated scRNA-seq datasets as references instead of bulk data. Hence, it would be great if the authors could include some of these tools in their analysis.*

R1-A1) We thank the reviewer for their comments. We have added MetaNeighbor to the methods compared. The implementation of MetaNeighbor required considerable communication with one of the method developers (M. Crow) who kindly guided us on which parts of the MetaNeighbor source code we needed to modify to make one of its variants (MetaNeighborUS) compatible with the type of task in our study. As we detail in the Methods section 'Implementation of tested methods and transformation of enrichment metrics for ROC and PR analyses', the original goal of MetaNeighbor is to quantify cell type replicability across scRNA-seq datasets (that the authors call 'studies'); whereas in our comparison, we are "comparing" all the clusters in one scRNA-seq dataset against known cell type specific gene sets or gene expression profiles. Similar to MetaNeighbor, Scmap projects cells from a scRNA-seq experiment on to the cell types or individual cells identified in a different experiment. We would need to apply similar workarounds and modify its code to use it in our study. Although we acknowledge that adding more methods to our comparison would make our results more complete, these other methods were not designed for the specific task we evaluate and would require code modifications to work on our input data. However, we provide an extensible framework, code and datasets that others can use for additional benchmarks. We now clarify this point in the paper.

R1-Q2) *I am confused as to which classifier parameter was varied in order to generate the ROCs. Were these comparable across the different methods?*

R1-A2) Sorry for the confusion. For a set of cell clusters, a given method was used to score each cluster against all cell type gene sets resulting in a matrix of cell type prediction scores per cluster. All scores in this matrix were combined into one column to capture all cell type prediction scores across all clusters and this set of prediction scores was varied to generate the ROC and PR curves. This is now clarified in Figure 1 and the text.

R1-Q3) *It would be interesting to see what the effect of varying the cluster resolution is to the ability of the methods to accurately label the populations. Do you obtain more diverse labelling when there are more clusters?*

R1-A3) Presumably yes. However, there is a methodological barrier that prevents us from investigating this aspect of the data using our current evaluation design. Authors of the

analyzed datasets provided gold standard annotations only at a single resolution per dataset, and we use these. Reclustering the original data to test other resolutions would require gold standards to be created for those resolutions (ideally by the original authors). However, we agree with the reviewer that studying the influence of cell cluster resolution is an interesting question. As the field moves towards increasing the number of scRNA-seq datasets annotated following standard ontology-based cell type annotations that consider a hierarchy of cell types at multiple granularities, this question could be addressed. We have added this to our discussion.

R1-Q4) *LM22 is a great reference dataset, but recently a new dataset has become openly accessible. This dataset, generated by Monaco et al1, characterizes 29 human immune cell types by RNA-seq and flow cytometry. It would be interesting to see if the use of this dataset leads to an improvement.*

R1-A4) Thanks for the pointer. We decided to keep the LM22 dataset because only six of the 22 cell types represented in it could be mapped into the PBMC data we analyzed. The Monaco dataset does not improve this number. Only five of the 17 cell types represented in the Monaco signature for RNA seq data are present in the PBMC data we analyzed. Furthermore, the ROC AUC and PR AUC values obtained using the LM22 and the Monaco signature are comparable to each other (Supplementary Table 2).

R1-Q5) *I think it would be helpful for the reader if the authors could summarize their results. The sheer number of comparisons made, means that the reader can feel overwhelmed at the end. A figure summarizing the various results for each method in each dataset could help clarify the message.*

R1-A5) Thanks for raising this point. We have now included summary Figure 6.

R1-Q6) *Thank you for making your code publicly available.*

R1-A6) Thanks. We have updated our GitHub repository with the MetaNeighbor implementation and modifications to our main wrapper to make it easier to incorporate new methods.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research