A Strategy to Compare Single-Cell RNA Sequencing Data Sets Provides Phenotypic Insight into Cellular Heterogeneity Underlying Biological Similarities and Differences Between Samples

Bioinformatics and Biology Insights
Volume 18: 1–12

DOI: 10.1177/11779322241280866

Dan C Wilkinson[1], Elizabeth Tallman[1], Mishal Ashraf[1], Tatiana Gelaf Romer[1], Jeehoon Lee[2], Benjamin Burnett[1] and Pierre R Bushel[1]

[1]Bioinformatics, BlueRock Therapeutics, New York, NY, USA. [2]Cardiac Cell Biology, BlueRock Therapeutics, Toronto, ON, Canada.

ABSTRACT: Single-cell RNA sequencing (scRNA-seq) allows for an unbiased assessment of cellular phenotypes by enabling the extraction of transcriptomic data. An important question in downstream analysis is how to evaluate biological similarities and differences between samples in high dimensional space. This becomes especially complex when there is cellular heterogeneity within the samples. Here, we present scCompare, a computational pipeline for comparison of scRNA-seq data sets. Phenotypic identities from a known data set are transferred onto another data set using correlation-based mapping to average transcriptomic signatures from each cluster of cells' annotated phenotype. Statistically derived lower cutoffs for phenotype inclusivity allow for cells to be unmapped if they are distinct from the known phenotypes, facilitating potential novel cell type detection. In a comparison of our tool using scRNA-seq data sets from human peripheral blood mononuclear cells (PBMCs), we show that scCompare outperforms single-cell variational inference (scVI) in higher precision and sensitivity for most of the cell types. scCompare was used on a cardiomyocyte data set where it confirmed the discovery of a distinct cluster of cells that differed between the 2 protocols for differentiation. Further use of scCompare on cell atlas data sets revealed insights into the cellular heterogeneity underpinning biological diversity between samples. In addition, we used a cell atlas to better understand the effect of key parameters used in the scCompare pipeline. We envision that scCompare will be of value to the research community when comparing large scRNA-seq data sets.

KEYWORDS: scRNA-seq, transcriptomics, heterogeneity, biological variation, compare data sets

RECEIVED: December 15, 2023. **ACCEPTED:** August 15, 2024.

TYPE: Method and Protocol

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The co-authors are, or were at some point during the investigation reported in this manuscript, employees of BlueRock Therapeutics which is a for-profit company developing cellular therapies.

CORRESPONDING AUTHORS: Dan C Wilkinson, Bioinformatics, BlueRock Therapeutics, 450 E 29th Street, New York, NY 10016, USA. Email: dwilkinson@bluerocktx.com

Pierre R Bushel, Bioinformatics, BlueRock Therapeutics, 450 E 29th Street, New York, NY 10016, USA. Email: pbushel@yahoo.com

Introduction

Single-cell RNA sequencing (scRNA-seq) has provided unprecedented power to interrogate gene expression transcriptome-wide in single cells.[1] Coupled with advances in computational biology to analyze scRNA-seq data,[2-4] biological samples can be resolved to identify heterogeneity among the single cells in a pooled sample.[5] Analyzing a scRNA-seq data set or integrating several data sets can be challenging because the data tend to be extremely high dimensional and very sparse. However, several methods have been recently developed to aid in the computation.[6] For example, batch correction and harmonization of scRNA-seq data sets help to minimize systematic differences so that the data can be compared on the gene expression level.[7] Unfortunately, there is a lack of computational approaches that serve to compare scRNA-seq data sets according to similarity and differences in phenotypic heterogeneity.

We developed the scCompare computational pipeline to facilitate the mapping of phenotypic labels from 1 scRNA-seq data set to another. The aims of the pipeline are to establish comparability and to potentially discover unique cell types. In scCompare, a mapping scRNA-seq data set with known cell type identities is processed with the standard pipeline steps including the identification of Leiden clusters and projection of the single cells into uniform manifold approximation and projection (UMAP) space. Given phenotypic annotations of the single cells in the clusters, cell type–specific prototype signatures are generated based on the average gene expression of each cluster. Using those cells assigned to each cluster, distributions of the correlations of gene signatures between the cells and the corresponding prototype signature are determined. Statistical thresholds for inclusion or exclusion are derived from the resulting distribution and used to evaluate each single cell from a test scRNA-seq data set to assign a phenotypic label. Single cells that fall outside of the distributions are labeled as unmapped. We evaluated scCompare on benchmark scRNA-seq data sets from peripheral blood mononuclear cells (PBMCs) and compared the performance to single-cell variational inference (scVI), a state-of-the-art computational tool for general scRNA-seq analyses including label transfer.[8] In addition, we used publicly available scRNA-seq data sets from atlases and experimental protocols that differentiate human

induced pluripotent stems cells (hiPSCs) into cardiomyocytes (CMs) to test the utility of scCompare to detect similarities and differences in single-cell populations.

Materials and Methods

Data sets

Human Protein Atlas (proteinatlas.org). The Human Protein Atlas (HPA) scRNA-seq data were collected as previously described.[9] Briefly, scRNA-seq read count data from 81 cell types from 31 different data sets (Supplemental file 1) were downloaded. The scRNA-seq data are based largely on the Chromium single-cell gene expression platform from 10× Genomics (version 2 or 3), constituting a single-cell suspension from tissues without pre-enrichment of cell types. This includes only studies with >4000 cells and 20 million read counts, and only data sets whose pseudo-bulk transcriptomic expression profile is highly correlated with the transcriptomic expression profile of the corresponding HPA tissue bulk samples. An exception was made for the eye (~12.6 million reads allowed), the rectum (2638 cells allowed), and the heart muscle (plate-based scRNA-seq platform) to include these additional cell types in the analysis.

Tabula Sapiens. The Tabula Sapiens (TS) scRNA-seq data were collected as previously described.[10] Briefly, 24 tissues in total were collected from 2 cohorts of donors (Supplemental file 2). This allowed biological replicates for almost all tissues. More details of the samples are available from the metadata posted on Figshare (https://figshare.com/articles/dataset/Tabula_Sapiens_release_1_0/14267219). The scRNA-seq raw read count data from the 10× Genomics pipeline were downloaded (TS cell atlas) for analysis.

Peripheral blood mononuclear cells. scRNA-seq data from approximately 3000 single PBMCs (3k data set) from a donor were contributed by 10× Genomics as a public resource. The raw read count data were downloaded for analysis.[11] In addition, scRNA-seq data from approximately 68 000 single PBMCs (68k data set) from a donor were generated using the 10× Genomics platform.[12] The raw read count data were used for analysis.

Cardiomyocytes. hiPSCs were differentiated to CMs in 90 days using 2 different protocols as previously described.[13] Briefly, in protocol 1, CMs were differentiated using small molecules with CHIR99021 and IWP2. In protocol 2, CMs were differentiated using cytokines Activin A, BMP4, and XAV939 plus small molecules with CHIR99021. Samples were collected on days 0 (D0), 12 (D12) and 24 (D24) for protocol 1 and days 0 (D0), 14 (D14), and 26 (D26) for protocol 2. The difference in the days of collection is due to the lag in the differentiation initiation day between the 2 protocols. scRNA-seq raw read count data were generated using the SPLiT-seq[14] (split-pool barcoding) library preparation methodology and from sequencing on an Illumina NextSeq.

Data filtering, preprocessing, and clustering

Single-cell analysis in Python (scanpy)[4] v1.9.2 was used exclusively for filtering, preprocessing, clustering the data, and marker gene identification. For data processed in the R environment, the SeuratDisk converts Seurat objects to AnnData objects via the h5Seurat file format specification. The unique molecular identifiers (UMIs) in each data set were annotated using the human Ensembl gene model.[15] The data were filtered such that transcripts not present in at least 3 cells were excluded from further analysis. Cell barcode identifiers were filtered to only include those with a minimum of 2000 nonzero transcripts, those having ⩾5% mitochondrial transcripts, and those having ⩽10% ribosomal transcripts. The count data were normalized to counts per million (CPM) for each cell by dividing the expression of each gene by the total count of its respective cell, multiplying the result by a million, and then applying a log base 2 transformation with an offset of 1. The normalized data were scaled across single cells to a mean expression = 0 and variance = 1. Highly variable genes were selected using the variance-stabilizing transformation option in scanpy followed by principal component analysis (PCA) to reduce the dimension of the data. Using the Kneedle heuristic[16] to determine the number of principal components (PCs) according to the point of maximum curvature of the explained variance and $k = 100$ nearest neighbors, single cells were embedded in a graph with the edges represented as distances drawn between cells. Using a resolution of 0.8, the Leiden algorithm was applied to group the single cells into clusters. Finally, the clusters of cells were visualized in UMAP space. Note, these hyperparameters (k in the nearest neighbor graph construction and Leiden resolution) are only suggested starting points and if possible should be tuned to reflect known ground truth biology (eg, canonical gene expression).

scCompare core functionalities: generate bulk signatures, statistical thresholding, and mapping phenotypic labels

After phenotypic labels were derived, either provided from external ground truth information sources or arrived at via unsupervised clustering, the scCompare pipeline was initiated. Normalized transcript count measurements and phenotypic labels from the training (mapping) data were used to build phenotypic label-specific prototype signatures that were based on the average expression of each phenotypic label using only highly variable genes. For each phenotypic label, distributions of the correlations of each cell's highly variable genes to the phenotypic label's prototype were generated. The correlation of each cell's gene signature to each prototype signature was

calculated, and the cell was initially assigned the phenotypic label to which it has the highest correlation.

The median absolute deviation (MAD), a measure of dispersion in data, uses the median as a statistic of central tendency and is robust against outliers.[17] If $x_1, x_2, \ldots, x_n$ represent a set of n Pearson correlation coefficients between the scRNA-seq expression data of signature genes for each cell in a specific phenotypic annotation from the mapping data set, and if $\tilde{X}$ is the median, then

$$MAD = \text{median}_i \left\{ \left| x_i - \widetilde{X} \right| \right\}.$$

We primarily used 5*MAD below the median as the statistical cutoff for the distribution of the Pearson correlation coefficients for a given phenotypic annotation to exclude phenotypic label assignment in the single cells in the testing data set. As an alternative approach to determining the statistical cutoff for distributions that might be highly skewed, we employed the Fisher transformation to convert the Pearson correlation coefficients to z-scores as follows:

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

where ln is the natural logarithm, and r is the Pearson correlation coefficient. The z-scores follow an approximately normal distribution with the standard deviation equal to

$$\frac{1}{\sqrt{N-3}}$$

where N is the sample size. Two or more standard deviations below the mean as the statistical cutoff demarcates ⩾97.5% of the correlations. Subsequently, the training data set was mapped back to the bulk signatures. Single cells that fell below the statistical cutoffs for the phenotypic annotation of the cells that are most correlated with were labeled as "unmapped." If a higher stringency for "unmapped" label assignment is desired, the MAD cutoff parameter may be user-adjusted, where a lower MAD cutoff would result in a higher number of "unmapped" cells (notably those whose signatures are most dissimilar from the prototype) allowing for further analysis. The coefficient of determination (R^2), Spearman rank correlation coefficient (R_s), and scatter plots were used to assess similarity and compare the proportion of phenotypes between the training and test data sets.

Marker gene identification

We looked for marker genes for each CM differentiation protocol (small molecule and cytokine-driven) at matched timepoints. Specifically, D12 and D24 of the small molecules in protocol 1 were compared with D14 and D26 of the small molecules and a cytokine in protocol 2, respectively. Marker genes were identified using scanpy's rank_genes_groups method. A t test with overestimated variance was used to detect genes that are statistically different between clusters of cells. The resulting genes were filtered requiring a minimum in-group fraction of 0.5, maximum out-group fraction of 0.5, and a minimum fold change of 2.

Performance metrics for the phenotypic mapping

Accuracy is the proportion of the mapping of the correct phenotypic labels to the test data set over all levels.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

Specificity is the probability of not falsely mapping the correct phenotypic label to the test data set.

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{False Positives} + \text{True Negatives}}$$

Recall (sensitivity) is the probability of mapping the correct phenotypic label to the test data set.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Results

scCompare overview

scCompare is a pipeline computational process using a training (mapping) scRNA-seq data set with known phenotypic annotations of single cells to label cells from a test data set of scRNA-seq data. The primary goal is to compare the phenotypic characteristics of the single cells (Figure 1). The mapping data set is processed to cluster the single cells for projection into UMAP 2-dimensional space. Cell type-specific prototype signatures are created using phenotypic labels (either provided by external sources or arrived at by unsupervised clustering), by calculating the average gene expression of all cells within each label. Then, distributions are obtained by comparing each cell's gene signature to the prototype of its respective label. Next, the gene signature from each single cell in the test data set is evaluated against statistical thresholds for each label distribution to assign a phenotypic annotation. Single cells that fall outside of the distributions are labeled as "unmapped." Finally, a comparison of the proportion of assigned cluster between data sets is

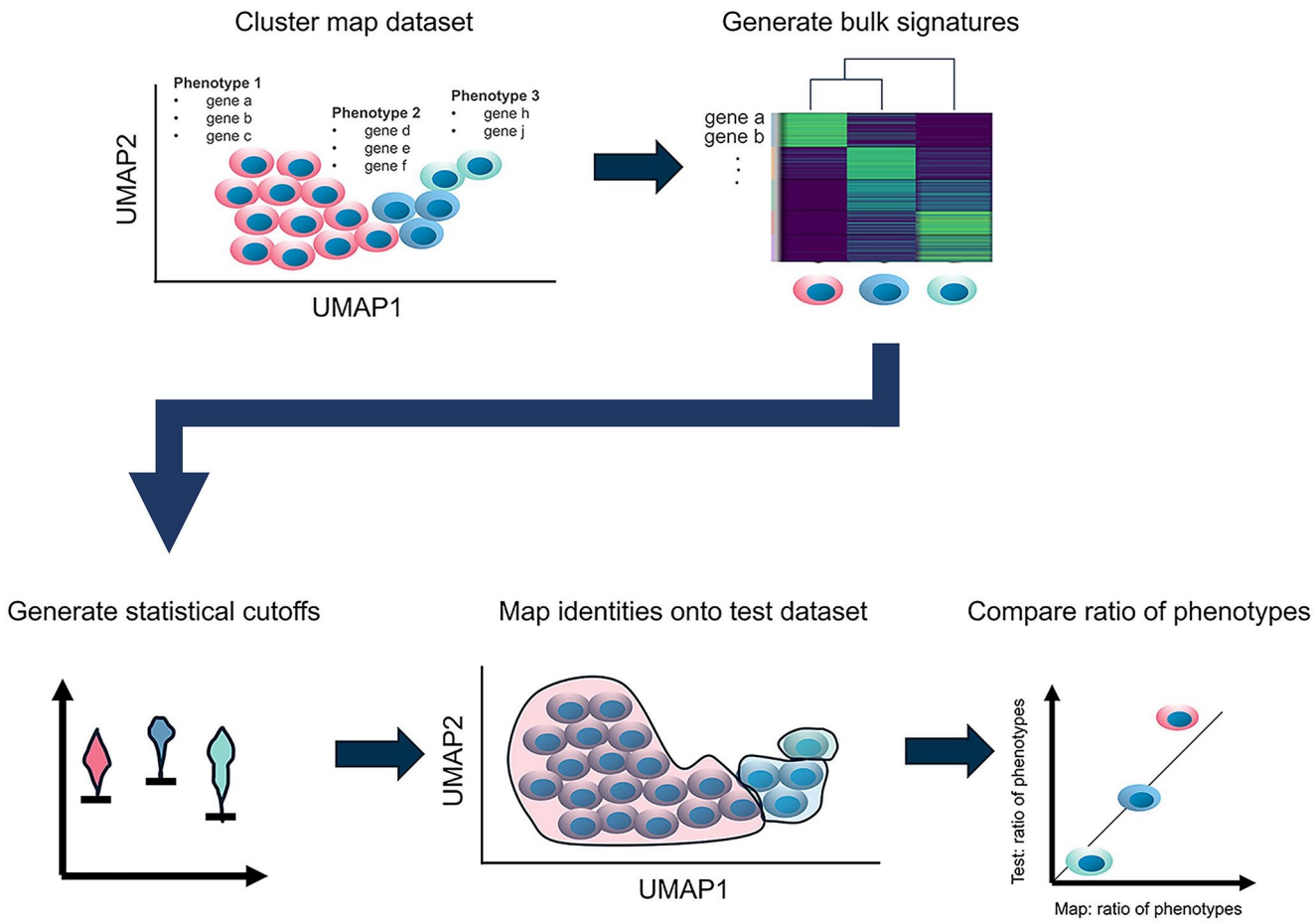


Figure 1. Compare workflow. Semi-supervised clustering is performed to group the mapping data set into phenotypically relevant subsets. Bulk signatures derived from the prototype gene expression signature for each cluster are generated, and distributions of within-cluster Pearson correlations of each member cell to the cluster's prototype signature are formed. These provide statistical cutoffs for cluster inclusivity. The cells from the test data set are correlated with each prototype signature, and the Pearson correlation is compared with the statistical cutoffs. Cells that pass the threshold are considered mapped and are labeled with the cluster's phenotype; otherwise, they are labeled as unmapped. Finally, fractions of mapped cells are compared between the mapping and testing data sets to facilitate comparability of phenotypic representations.

visualized in a scatter plot, and R^2 is calculated as a statistic measuring the similarity between the phenotypic composition of the 2 data sets.

Comparison of scCompare to scVI to map phenotypic labels

The scVI[8] is a scRNA-seq analysis tool that uses an optimization routine and deep neural networks to join information across cells and genes that are similar, and to estimate the distributions that represent the observed expression values. An advantage of scVI is that it uses a neural network-based approach to construct latent-space representations from the matrix of read counts and batch information to account for experimental and technical variation in the data, and then estimate biological differences between cells. A couple of disadvantages of scVI is that it uses a non-deterministic algorithm leading to alternative results with different initializations, and for genes with few cells, the prior and the inductive bias of the neural network may not fit the data optimally. The expected read count data (batch-corrected and normalized) are used in scVI for comparison with scCompare.

To assess the ability of scCompare and scVI to align biological annotations between scRNA-seq data sets, we randomly sampled the 3k PBMCs scRNA-seq data 50 times in training and testing splits (80:20) and used the top 2000 highly variable genes for signature construction. Performance is based on the accuracy, precision, and sensitivity of scCompare and scVI mapping the phenotypic labels to the test data sets. As shown in Table 1, the accuracy of predictions was similar, but scCompare matches or outperforms scVI in precision and sensitivity for most cell types, most strikingly with the dendritic cells and megakaryocytes. However, scVI scored slightly higher for CD4 T cells, FCGR3A monocytes, and natural killer (NK) cells in precision for the former and sensitivity for the latter two.

Table 1. Performance comparison between scCompare and scVI using the 3k PBMCs scRNA-seq data set.

CELL TYPES	ACCURACY		PRECISION		RECALL	
	SCCOMPARE	SCVI	SCCOMPARE	SCVI	SCCOMPARE	SCVI
B	0.998	0.995	0.990	0.969	0.996	0.996
CD14 monocytes	0.982	0.976	0.948	0.943	0.943	0.905
CD4T	0.975	0.955	0.981	0.986	0.962	0.911
CD8T	0.967	0.949	0.829	0.721	0.877	0.861
Dendritic	0.998	0.992	**0.957**	**0.628**	**0.925**	**0.665**
FCGR3A monocytes	0.991	0.984	0.915	0.830	0.938	0.948
Megakaryocytes	1.000	0.995	**1.000**	**0.160**	**0.993**	**0.133**
NK	0.984	0.979	0.897	0.822	0.903	0.947

Performance parameters based on the average of 50 iterations of 80%:20% splits of the data training to test and using 2000 highly variable genes. Bold indicates that scCompare most strikingly outperforms scVI in precision and recall with the dendritic cells and megakaryocytes.

We also evaluated the ability of scCompare and scVI to map known transcriptomic signatures from the 3k PBMCs scRNA-seq data set to the 68k PBMCs scRNA-seq data set (Figure 2). As shown in Figure 2B, the cell type identities from the 3k data set were mapped equivalently to the 68k data set except for the megakaryocytes in the case of scVI where they were labeled as NK cells. The gene representations of the clusters of single cells for the phenotypic labels are comparably similar except for the megakaryocytes (Figure 2B). Furthermore, scCompare identified a previously unannotated cell type, plasmacytoid dendritic cell.[18] The application of the MAD-based statistical cutoff allowed for discovery of potentially novel cell types, although not all unmapped cells would be novel (not accounted for in the mapping data set). To determine whether a group of unmapped cells was novel, 2 parameters are assessed: (1) a differentially expressed gene-based metric and (2) a UMAP Euclidean distance-based metric (Figure 2C, Table S1, Figure S1, and Figure S2). The cell cluster identified as plasmacytoid dendritic cells was observed to have a gene signature divergent from that of dendritic cells (the pre-cutoff assigned phenotype) and was separated from the mapped dendritic cells in UMAP space. Furthermore, differential gene expression analysis revealed the expression of MZB1 in this cluster, a marker gene for plasmacytoid dendritic cells.[18] This not only demonstrates scCompare's usefulness in mapping phenotypes between data sets, but it also provides the added ability to discover novel cell types not accounted for in the mapping data.

Mapping between scRNA-seq cell atlases

To demonstrate a broader applicability of the scCompare pipeline, we performed cell atlas mapping using HPA as the training data set and TS as the test data set. These atlases contain samples obtained from primary human adult tissue from multiple individuals across many tissues. These samples have also been extensively annotated with various levels of descriptive labels including cell phenotype, organ-of-origin, sex, and other information, typifying the breadth and depth of heterogeneity present within human biology. To demonstrate scCompare's performance on atlas-to-atlas mapping, phenotypic label alignment between atlases was necessary for a ground truth comparison. First, for each data set, we combined phenotypic and organ-of-origin annotations and excluded groupings (phenotype; organ-of-origin, for example: T cells; salivary gland) that did not contain at least 300 cells or were not present in both HPA and TS. We then removed groupings that denoted phenotypes that were too broadly defined, likely containing more than 1 distinct phenotype. For example, stromal cells may be comprised of fibroblasts, adipocytes, pericytes, vascular cells, etc. We then removed all cells that were annotated as stem cells as it is unclear how to align phenotypes of unknown maturation state across atlases. Finally, after filtering, resulting labels were hand-aligned by ensuring overlap between the annotated phenotype and the organ-of-origin (Table S2) and each resulting grouping subset to 300 cells to streamline processing.

We obtained bulk signatures corresponding to each of the 22 annotation groups in HPA sampled at available tissue level. Bulk signatures were captured using the top 2000 highly variable genes. These bulk signatures were then applied to the TS data set comprising 50 corresponding sub-categories sampled from 21 different tissue locations throughout the body. No statistical cutoff was applied as we were testing direct 1:1 mapping fidelity (Table 2).

Figure 3 depicts a summary of the classification performance based on differential expression analysis, along with visualizations showing the alignment of classification mapping between HPA and TS cell atlases. Differential gene expression analysis was conducted on HPA, using the same parameters as in the cardiac differentiation comparison. The top differentially expressed genes from HPA are displayed for both atlases, allowing for a direct comparison (Figure 3A). In this comparison, cells from the TS atlas were not ground truth labels, but

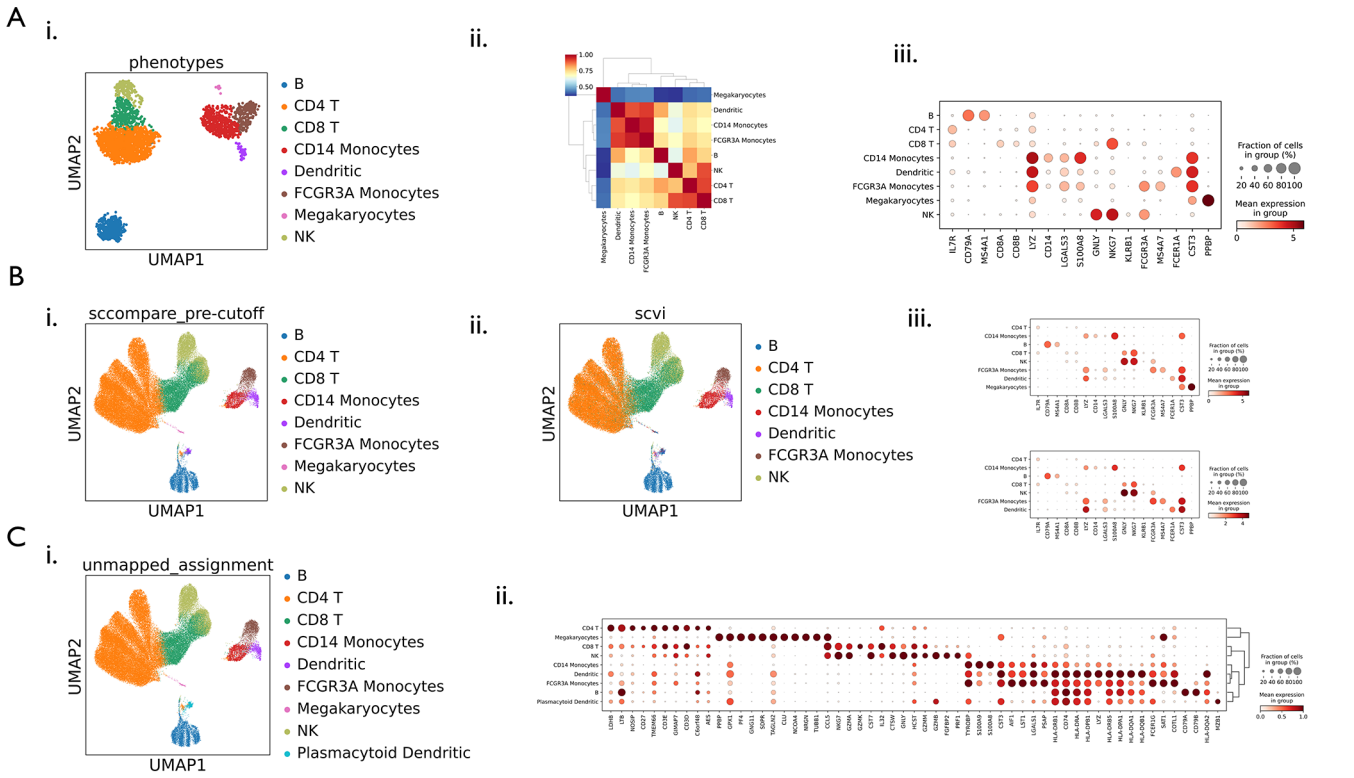


Figure 2. scCompare and scVI 3k PBMCs data set-derived models applied to the 68k PBMCs data set and the discovery of an unannotated cell type. (A) 3k PBMCs data set: (i) UMAP of 3k PBMCs data set, (ii) correlation dendrogram of 3k PBMCs phenotypes, and (iii) 3k PBMCs gene expression dotplot showing expression of key phenotypic markers. (B) Results of mapping 3k PBMCs identities onto the 68k PBMCs data set using scCompare and scVI. (i-ii) UMAPs showing phenotypes assigned by scCompare (i) and scVI (ii), (iii) gene expression dotplots showing consistency of key gene expression between the 3k PBMCs data set and the assigned phenotypes in the 68k data set (top—scCompare, bottom—scVI). scVI does not predict any megakaryocytes in the 68k PBMCs data set despite evidence for their presence (small PPBP-expressing subcluster (A iii)). The intensity scale represents the mean expression of a gene within a cell type. The size of the dots represents the fraction of the cells within a cell type. (C) Statistical cutoff-assisted discovery of plasmacytoid dendritic cells. Cells not meeting statistical cutoff were assessed for novel phenotype status using differential gene expression list analysis and UMAP Euclidean distance metrics (Figure S1). After assessment and remapping, a cluster of plasmacytoid dendritic cells was discovered. This phenotype is confirmed by the expression of MZB1. Further characterization can be found in Figure S1.

rather were categorized according to their phenotypic labels assigned by scCompare. The results show a broad agreement in the expression patterns of the differentially expressed genes derived from HPA. Key marker genes for each phenotype are highlighted in the legend to provide additional biological context and phenotypic validation. Circos plots display a visual representation of classification accuracy along a variety of axes including tissue of origin, cell type, and cell type sub-categorizations (Figure 3B). The vast majority of cells fall into the correct corresponding phenotypic grouping with more than 90% overall classification accuracy. The class weighted and unweighted prediction accuracy scores were 90.2% and 88.1%, respectively. Cell type confusion did occur between closely related subtypes of cells, such as different subtypes of immune cells and epithelial cells. This can likely be attributed to shared lineage and functional categories. For instance, confusion among T lymphocytes (T cells), B lymphocytes (B cells), and NK cells was observed (Table 2). These cell types are of shared lineage and descend from a shared ancestor, the common lymphoid progenitor. Similarly, there is confusion among enterocytes and other epithelial cells from the large and small intestines. In the case of club cells, their primary class confusion occurred with type 2 alveolar cells. Their confusion can likely be attributed to their common functions, as both cell types are secretory epithelia featuring specialized adaptations to the respiratory tract. Evaluation of the associated clustermap (Figure S3) highlights the core similarities in scCompare phenotypic signatures that were likely the root cause of high misclassification rates for the previously mentioned cell types.

Comparison of cardiomyocyte differentiation protocols

To demonstrate the utility of scCompare to compare scRNA-seq data sets from a study design, we leveraged the data from an experiment that evaluated 2 different protocols to differentiate hiPSCs to CMs.[13] Differentiation protocol 1 used small molecules whereas protocol 2 used a cytokine in addition to the

Table 2. scCompare performance in finer categorized label transfer task.

HPA LABEL CATEGORY	# MISCLASSIFIED	TOTAL COUNT	CLASSIFICATION ACCURACY (%)	MISCLASSIFIED PHENOTYPES (>5%)
Hepatocytes	0	300	100	N/A
Basal keratinocytes	0	300	100	N/A
Exocrine glandular cells	3	300	99	N/A
Muller glia cells	3	300	99	N/A
Alveolar cells type 2	4	300	98.67	N/A
Endothelial cells	69	4200	98.36	N/A
Fibroblasts	58	1800	96.78	N/A
Macrophages	131	3600	96.36	N/A
Intestinal goblet cells	27	600	95.5	N/A
Respiratory ciliated cells	15	300	95	N/A
Smooth muscle cells	30	600	95	N/A
Cardiomyocytes	17	300	94.33	N/A
Basal prostatic cells	17	300	94.33	N/A
Nk cells	47	600	92.17	T cells: 6.5
B cells	370	3300	88.79	T cells: 9.85
Distal enterocytes	35	300	88.33	Intestinal goblet cells: 8.67
T cells	1480	11 400	87.02	Nk cells: 6.97
Plasma cells	119	900	86.78	B cells: 6.11
Proximal enterocytes	71	300	76.33	Distal enterocytes: 10.33,
				B cells: 7.67
Salivary duct cells	116	300	61.33	T cells: 32.0,
				Macrophages: 5.67
Club cells	256	600	57.33	Alveolar cells type 2: 39.5
Ductal cells	190	300	36.67	Exocrine glandular cells: 62.0
Weighted accuracy	3058	31 200	90.2	N/A
Unweighted accuracy	N/A	N/A	88.05	N/A

Abbreviation: N/A, not applicable.

small molecules. The differentiation of the hiPSCs was carried out for 90 days. As represented in Figure 4A, protocol 1 samples were taken for scRNA-seq at D0, D12, and D24, whereas protocol 2 samples were taken for scRNA-seq at D0, D14, and D26. The difference in the days of collection was due to the lag in the differentiation initiation day between the 2 protocols. We selected the scRNA-seq data from protocol 1 as the training (mapping) data set and the scRNA-seq data from protocol 2 as the testing data set. scRNA-seq analysis of the training data set generated 13 Leiden clusters (Figure 4B). The UMAP representation of the clusters shows well-formed clusters and an abundance of heterogeneity in the data which is expected in a "developmental" (differentiation) setting compared with "adult"/mature tissues. Differential expression analysis of the Leiden clusters 0 to 9 revealed marker genes *TNNT2*, *MYH6*, and *MYH7* representative of CMs. Cluster 11 includes the CM marker genes in addition to *MKI67* and *FN1* indicative of progenitor CMs (PCMs). Clusters 12, 7, and 5 all have marker genes *FN1*, but also *GRHL2* and *AFP* for the latter 2 clusters, respectively, and they represent stromal-like cells, endodermal cells, and ectodermal cells individually. The 2 remaining clusters 10 and 13 have marker genes *TRPM3, CTNNA2*, and *EGFL7*, respectively, and correspondingly, they suggest smooth muscle-like and endothelial-like cell types.

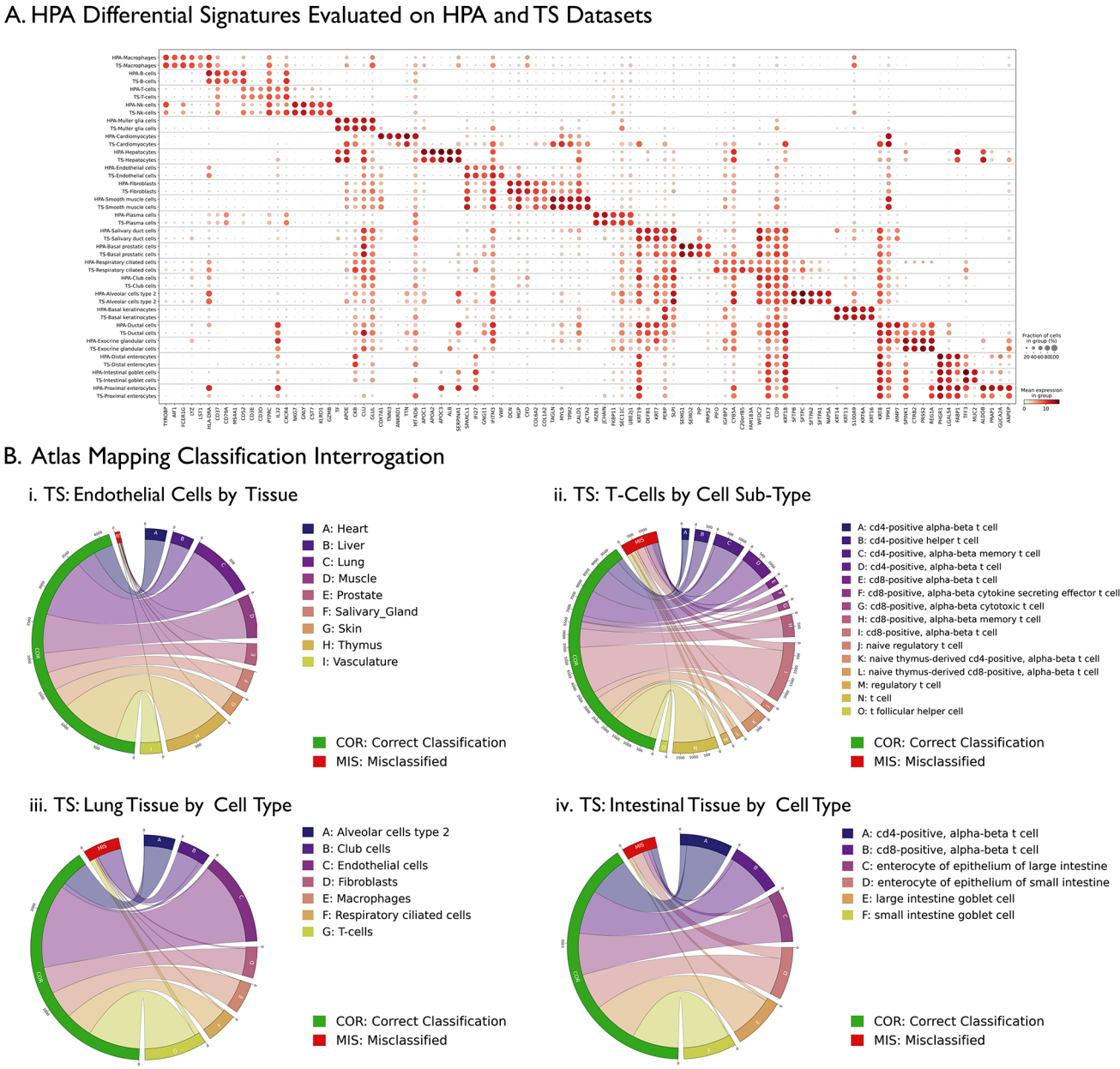
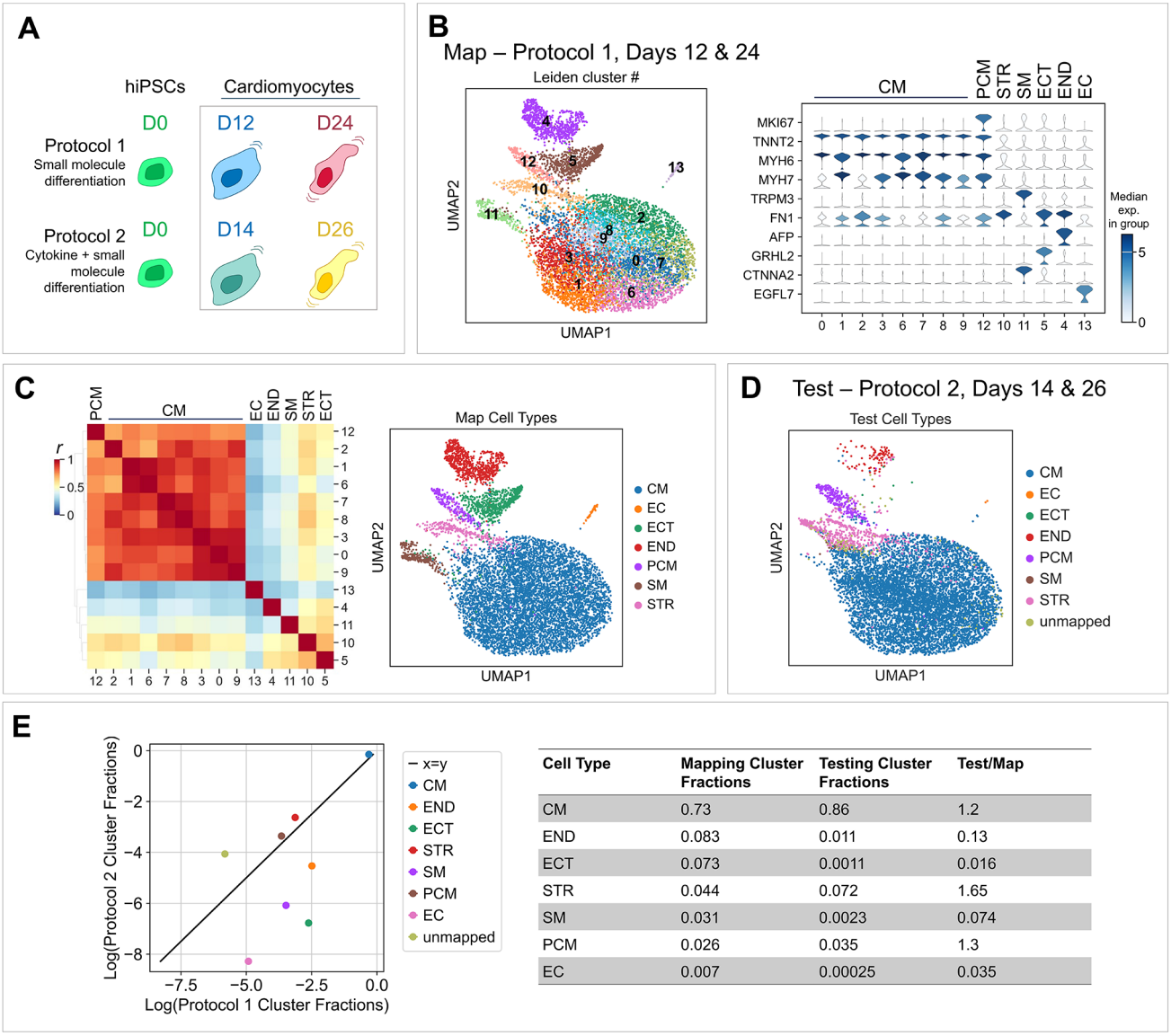


Figure 3. Visualizations of HPA and TS atlas mapping experimental results. (A) Differential expression analysis of HPA cells by phenotypic label with corresponding expression of scCompare-classified TS cells displayed using a dotplot. Each row denotes the similarity in expression between HPA-phenotypically labeled cells and corresponding TS-mapped cells post-scCompare analysis. Key markers for each phenotypic label that arose from the differential expression analysis are annotated here, macrophages: AIF1 (allograft inflammatory factor 1), B cells: MS4A1 (membrane spanning 4-domains A1, CD20), T cells: CD3E (CD3e molecule), NK cells: NKG7 (natural killer cell granule protein 7), Müller glia cells: GFAP (glial fibrillary acidic protein), cardiomyocytes: TNNI3 (troponin I3, cardiac type), hepatocytes: ALB (albumin), endothelial cells: VWF (von Willebrand factor), fibroblasts: COL1A2 (collagen type I alpha 2 chain), Smooth muscle cells: ACTA2 (actin, alpha 2, smooth muscle, aorta), plasma cells: JCHAIN (joining chain of multimeric IgA and IgM), salivary duct cells: KRT7/KRT19 (keratin 7/keratin 19), basal prostatic cells: PIP (prolactin-induced protein), respiratory ciliated cells: PIFO (primary cilia formation protein), club cells: SFTPA2 (surfactant protein A2), alveolar cells type 2: SFTPC (surfactant protein C), basal keratinocytes: KRT14 (keratin 14), ductal cells: KRT8/KRT19 (keratin 8/keratin 19), exocrine glandular cells: CTRB2 (chymotrypsinogen B2), distal enterocytes: FABP1 (fatty acid binding protein 1), intestinal goblet cells: MUC2 (mucin 2), and proximal enterocytes: ALDOB (aldolase B, fructose-bisphosphate). Highlighting key expression of these genes helps to validate HPA phenotype and TS mapping fidelity using canonical gene expression. (B) Atlas mapping experimental results are subdivided to further interrogate performance. Endothelial cells (i) display high classification accuracy despite a variety of sampling locations. In contrast, T cells (ii) display a higher rate of misclassification, which has a clear bias toward some sub-types showing higher rates of misclassification (eg, cd8 + alpha-beta t cells, generic t cells). This is likely due to the broader phenotypic heterogeneity that is captured within the T-cell phenotype. (iii) Lung tissue higher rates of misclassification among club cells are shown (possibly due to the commonalities between respiratory club cells and type 2 alveolar cells, as both as secretory epithelia with similar tissue-specific gene expression profiles). (iv) Cell types sampled from both the large and small intestines whereby an interesting pattern is depicted showing the confusion rate for enterocytes of the small intestine higher misclassification rates and unidirectional, as enterocytes of the large intestine are correctly classified at a higher rate.



Cell Type	Mapping Cluster Fractions	Testing Cluster Fractions	Test/Map
CM	0.73	0.86	1.2
END	0.083	0.011	0.13
ECT	0.073	0.0011	0.016
STR	0.044	0.072	1.65
SM	0.031	0.0023	0.074
PCM	0.026	0.035	1.3
EC	0.007	0.00025	0.035

Figure 4. scCompare recapitulates published mapping of phenotypes between 2 directed differentiation protocols. (A) Sample schematic indicating the types of differentiation protocols and timepoints (dashed box) analyzed. (B) UMAP of Leiden clustering of "mapping" data set, protocol 1 D12 and D24, and violin plots displaying marker gene expression profiles for each Leiden cluster, with cell type indicated above. (C) Correlation dendrogram of highly variable genes and Leiden clusters in the "mapping" data set, protocol 1, and UMAP with colors indicating cell type. (D) UMAP of "testing" data set, protocol 2 D14 and D26, after mapping using the annotated protocol 1 (training) cell types, and UMAP of Pearson correlation coefficients of each labeling. (E) Scatter plot showing the relative ratios (natural log) of annotated cell types between protocol 1 and protocol 2. An $x=y$ line is plotted to aid in the visual inspection of the ratio comparisons. Dots falling to the lower right of the line represent cell types enriched in protocol 1, whereas dots falling to the upper left of the line represent cell types enriched in protocol 2. Dots that fall directly on the line indicate an equal fraction of the cell types is shared between protocols 1 and 2.
CM indicates cardiomyocyte; EC, endothelial-like; ECT, ectodermal; END, endodermal; PCM, progenitor CM; SM, smooth muscle-like; STR, stromal-like.

Figure 4C depicts a correlation heat map of the highly variable genes expressed by the single cells in the Leiden clusters and revealed a high degree of similarity between the CM cell types (clusters 0-9 and PCM cluster 11). The other clusters have moderate to low correlation to each other. The cell type annotations in the protocol 1 (mapping) data showed good uniformity of the CM clusters and relatively good individual clustering of the other cell types. However, after scCompare mapping of the phenotypic identities to the protocol 2 (test) data, there appeared to be heterogeneity in the CM cluster, very few ectodermal, endothelial-like, and smooth muscle-like cells, and less representation of endodermal cells (Figure 4D). The aforementioned reduction in cell type mapping in protocol 2 is represented in Figure 4E numerically as proportions and graphically as a scatter plot of the cluster fraction of cells with the annotated cell type in protocol 1.

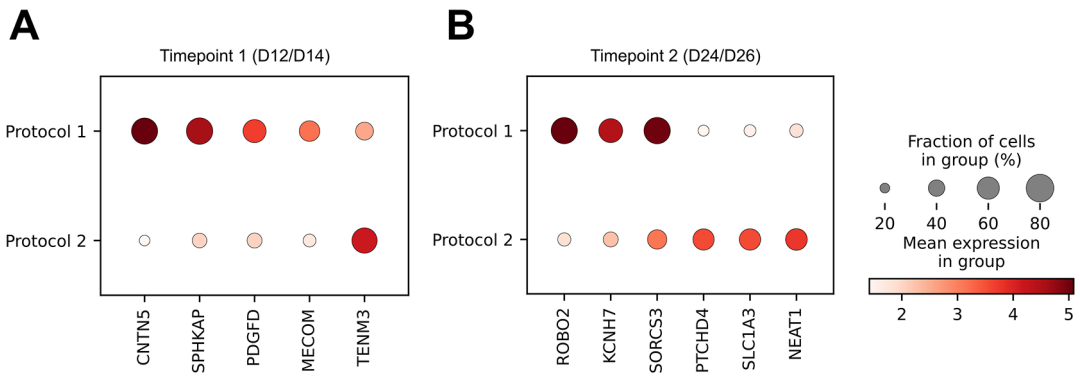


Figure 5. Dotplot visualization of marker genes between protocols 1 and 2. Differentially expressed genes between clusters of cells in (A) protocol 1 versus protocol 2 for timepoint 1 (D12/D14) and (B) protocol 1 versus protocol 2 for timepoint 2 (D24/D26). The legend denotes the size of the dots as the percentage of the cells in the groups expressing the genes, and the color bar represents the level of mean expression of the genes within the groups.

The R^2 and the R_s of the cell type fraction between protocols 1 and 2 are moderate at 0.441 and 0.393, respectively, markedly low due to the underrepresentation of endodermal END, SM, ECT, and EC cell type fractions. Single cells in protocol 2 that were not assigned a cell type were labeled as unmapped. From the higher proportion of CM cells in protocol 2 (~19%) and less off-target cell types, it can be posited that the addition of cytokines in the differentiation protocol could have led to the generation of more on target cells. Differential expression analysis for D12/D14 timepoint revealed a set of marker genes (*CNTN5*, *SPHKAP*, *MECOM, and TENM3* previously associated with iPSC-derived CM identity) that are differentially expressed in protocol 1 vs protocol 2 samples (Figure 5A). Interestingly, the same analysis for the later timepoint (D24/26) suggests a completely different set of marker genes implicated in cardiac morphogenesis (*ROBO2*), maturation (*KCNH7*), and hypoxic response (*NEAT1*) (Figure 5B).

scCompare number of genes and cluster resolution parameter assessment

A final goal of our analysis was to assess the scCompare pipeline for its sensitivity to 2 key user-defined parameter selections. To assess this, we used a subset of 10 000 cells obtained from the HPA data set and separated these into 5000 mapping and test cells. The scCompare pipeline was repeatedly performed on this data set using an increasing size selection of highly variable genes and clustering resolutions (Figure 6). Highly variable gene selection was bounded between 50 and 10 000. Clustering resolution was bounded between 0.05 and 2.5, which corresponded to 2 to 40 distinct clusters. Figure 6A colors each individual iteration by weighted-F1, a statistical measure of prediction accuracy bounded by 0 and 1. With respect to clustering resolution, the results demonstrate the impact of overclustering on performance, as greater resolutions tend to produce ambiguity between classes resulting in misclassification.

When assessing the impact of gene signature length, the results suggest that beyond a certain minimum threshold, most signature length selections tend to perform well. Figure 6B and C illustrates the same data set evaluating each individual run by map and test misclassification rates, with colors provided by input variable. These plots best capture the trend toward higher input values of clustering resolutions and signature lengths tending to produce higher test misclassification than map misclassification, an outcome that could be considered as overfitting. Figure 6D provides another view of overfitting, where points are colored by the absolute difference in map and test misclassification rates. Higher values tend to occur at the highest clustering resolutions and gene signature lengths. As a note, Figure 6D masks cases in which the misclassification rate is equally poor in test and mapping data sets, as is the case in the low gene signature length (Figure 6A). In general, our results suggest that aside from excessive clustering resolutions and gene signature lengths, our pipeline produces robust results.

Discussion

Advances in next-generation sequencing (NGS) allow for high-resolution gene expression profiling transcriptome-wide at the single-cell level. The ability to assess heterogeneity within samples provides a unique insight into the complexity of biology. Over the years, the size and scope of scRNA-seq experiments increased as the technology became more readily available. At the same time, sophisticated computational tools to analyze the data were developed. However, there is a paucity of methods to compare scRNA-seq data sets at the phenotypic level.

We developed scCompare as an analytical pipeline to compare clusters of single-cell identities to another data set for assessment of similarities and differences in phenotypic characterization (Figure 1). scCompare leverages the workflow in scanpy to filter, preprocess, cluster the data, and identify marker genes making the ease of entry extremely useful for a wide range of users in the community. For those who prefer the R environment, SeuratDisk converts Seurat objects to AnnData

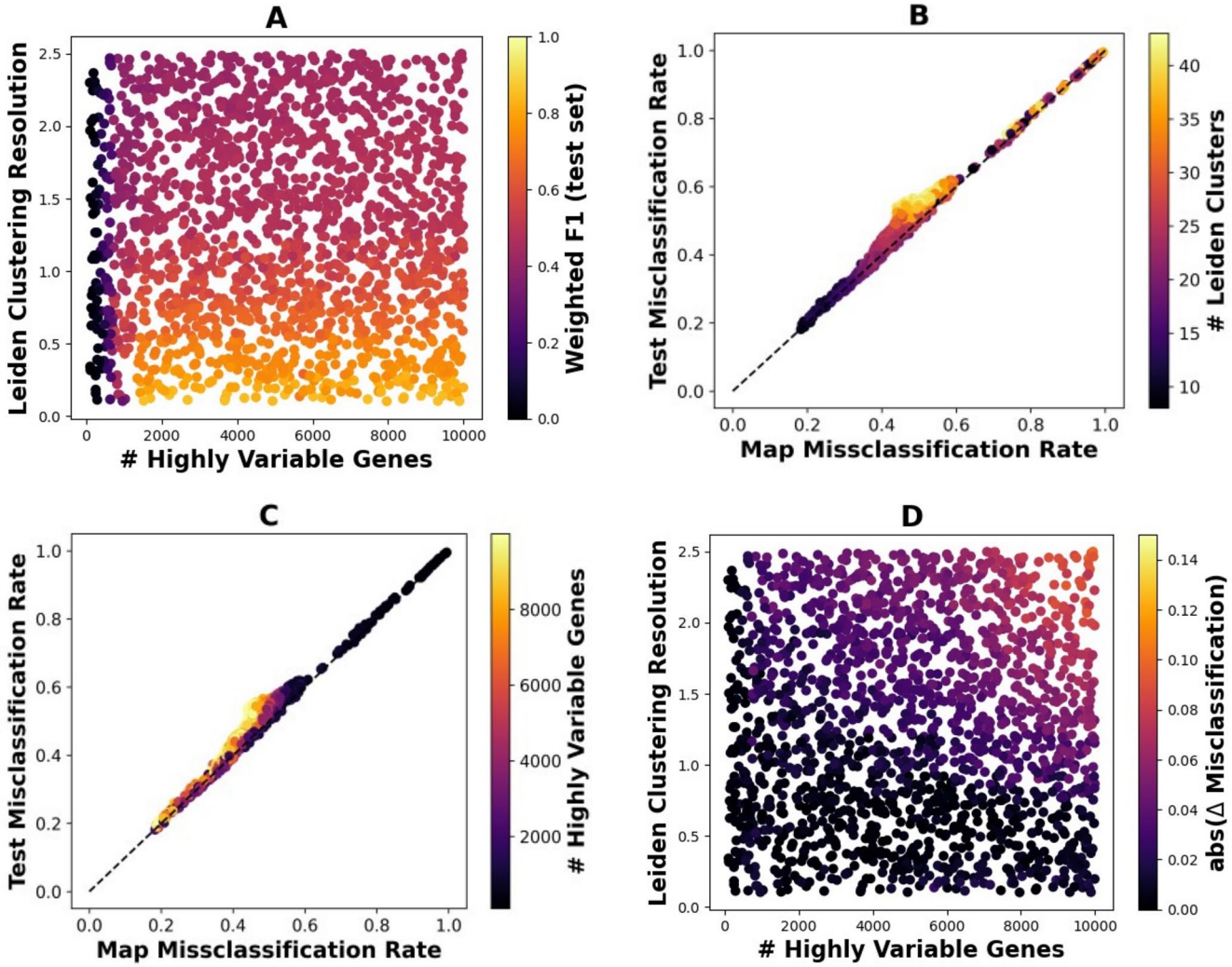


Figure 6. scCompare grid search exploring the effects of parameter optimization on results using the HPA scRNA-seq data set. (A) Each point (a subset of 10 000 cells) represents an individual run of scCompare, with its highly variable genes assigning position in X and Leiden clustering resolution assigning its position in Y. The color of each point is assigned by its classification performance as measured by class-size weighted F1 score. Panels (B) and (C) present a different view of this data, where the points are positioned by map and test misclassification rates in X and Y, respectively. The identify line acts as a visual guide, where points deviating from this line have unequal percentages of map and test set misclassification rates. The plots are colored by (B) number of Leiden clusters and (C) number of highly variable genes, respectively. Panel (D) shares the same axis as (A), but points are instead colored by the absolute difference in misclassification rate between map and test sets.

objects via the h5Seurat file format specification and is portable into scCompare. The novelty of scCompare lies in the following. Using labels in a scRNA-seq data set, bulk signatures for each label are generated by correlating highly variable genes within each label to a gene expression prototype. Based on statistical cutoffs for each distribution of bulk signatures, labels from a test scRNA-seq data set are compared with the threshold for each distribution of the labels' bulk signatures to assign phenotypic characterization or label them as unmapped. The unmapped cells provide opportunity for further investigation of similarities and differences in representation of cell identities and/or biological discovery.

In comparison with scVI, a deep learning probabilistic model for perusing unexplored biological diversity for scRNA-seq data, scCompare matched or outperformed the tool in higher accuracy and specificity for most of the cell types (Table 1 and Figure 2). More importantly, in contrast with scVI, scCompare has a flexible utility in that it can compare atlases of scRNA-seq data (Table 2) (Figure 3) or, in the case of CM differentiation protocols, reveal an unmapped cluster of single cells (Figure 4). scCompare also allows for the identification of "unmapped" cells based on their dissimilarity to computed signatures by setting statistical cutoffs. We demonstrated the utility of this feature in our comparison of CM differentiation protocols, which revealed an unmapped group of cells which may have some biological relevance (Figures 4 and 5).

It is clear that scRNA-seq data are large, complex and requires a fair amount of bioinformatics expertise, computational savviness, and biological intuition to mine the data effectively. scCompare provides a fairly straightforward analysis pipeline for novices

to use. In fact, our assessment of 2 key parameters used in the scCompare pipeline gives guidance to users on how to optimize analysis results (Figure 6). However, there are a few caveats to consider when using scCompare to glean biological insight from a comparison of 2 scRNA-seq data sets: (1) the mapping data set requires identities of the cells or biological expertise to phenotypically label them; (2) often times the mapping data set has a hierarchical structure (Figure S3) to the cell identities, and as such, the results will likely differ based on the level of phenotypic annotation used; thus, it is important to evaluate the correlation dendrogram produced during the scCompare pipeline run to identify phenotypic signatures that are highly correlated as they may lead to misclassification of the test data set (Figure S3); (3) the statistical cutoffs are user-defined and can be adjusted to refine the phenotypic characterization in the test data set; (4) cells may be misclassified or unlabeled if they have significantly higher sparsity than the cells used to generate the signatures; and (5) it is plausible that the cell cycle may be confounding and cells may map to cell cycle signatures on the sole basis of that phenotypic property, despite, for example, originating from a totally different germ layer.

Given the utility and practicality of scCompare, we envision that the tool will be of value to the basic science research community, biotechnology, and pharmaceutical industries, when needing to compare large scRNA-seq data sets.

Acknowledgements

The authors thank Nick Stitt and Amanda Ferreira for their critique of the article and the scCompare pipeline.

Author Contributions

DCW conceived of the scCompare methodology, developed the scCompare code and served as lead architect, devised the comparison analyses, performed analyses, and wrote portions of the manuscript. ET and MA performed analyses and wrote portions of the manuscript. PRB contributed to the statistical applications in scCompare and wrote significant portions of the manuscript. BB and TGR contributed to the creation and maintenance of the scCompare repository. TGR also performed marker gene analyses for the CM differentiation data. JL interpreted the results of the scCompare analysis of the CM differentiation protocols and wrote a portion of the manuscript related to it.

Data and Programming Code Availability Statements

The HPA data are publicly available at https://www.proteinatlas.org/about/download. The TS gene count data are publicly available at the Gene Expression Omnibus under accession GSE201333. The 3k PBMCs data are publicly available at https://s3-us-west-2.amazonaws.com/10x.files/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz. The 68k PBMCs data are publicly available at https://www.10xgenomics.com/resources/datasets/fresh-68-k-pbm-cs-donor-a-1-standard-1-1-0. The CM data are publicly available at https://open.quiltdata.com/b/allencell/packages/aics/wtc11_hipsc_cardiomyocyte_scrnaseq_d0_to_d90. The scCompare repository is https://github.com/bluerocktx/bfx-scCompare.

ORCID iD

Pierre R Bushel https://orcid.org/0000-0001-5188-8693

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

REFERENCES

1. Tang F, Barbacioru C, Wang Y, et al. MRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*. 2009;6:377-382. doi:10.1038/nmeth.1315
2. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep*. 2019;9:5233. doi:10.1038/s41598-019-41695-z
3. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. Published September 17, 2020. Accessed September 21, 2023. http://arxiv.org/abs/1802.03426
4. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19:15. doi:10.1186/s13059-017-1382-0
5. Choi YH, Kim JK. Dissecting cellular heterogeneity using single-cell RNA sequencing. *Mol Cells*. 2019;42:189-199. doi:10.14348/molcells.2019.2446
6. Ryu Y, Han GH, Jung E, Hwang D. Integration of single-cell RNA-seq datasets: a review of computational methods. *Mol Cells*. 2023;46:106-119. doi:10.14348/molcells.2023.0009
7. Tran HTN, Ang KS, Chevrier M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol*. 2020;21:12. doi:10.1186/s13059-019-1850-9
8. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. *Nat Methods*. 2018;15:1053-1058. doi:10.1038/s41592-018-0229-2
9. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347:1260419. doi:10.1126/science.1260419
10. Tabula Sapiens Consortium, Jones RC, Karkanias J, et al. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science*. 2022;376:eabl4896. doi:10.1126/science.abl4896
11. 10X Genomics 3k PBMCs from a Healthy Donor. https://s3-us-west-2.amazonaws.com/10x.files/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz
12. Zheng GXY, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8:14049. doi:10.1038/ncomms14049
13. Grancharova T, Gerbin KA, Rosenberg AB, et al. A comprehensive analysis of gene expression changes in a high replicate and open-source dataset of differentiating hiPSC-derived cardiomyocytes. *Sci Rep*. 2021;11:15845. doi:10.1038/s41598-021-94732-1
14. Rosenberg AB, Roco CM, Muscat RA, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*. 2018;360:176-182. doi:10.1126/science.aam8999
15. Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. *Nucleic Acids Res*. 2022;50:D988-D995. doi:10.1093/nar/gkab1049
16. Satopaa V, Albrecht J, Irwin D, Raghavan B. Finding a "Kneedle" in a haystack: detecting knee points in system behavior. Paper presented at: 2011 31st International Conference on Distributed Computing Systems Workshops; June 20-24, 2011:166-171; Minneapolis, MN. doi:10.1109/ICDCSW.2011.20
17. Iglewicz B, Hoaglin DC. *How to Detect and Handle Outliers*. ASQ Quality Press; 1993.
18. Kapoor T, Corrado M, Pearce EL, Pearce EJ, Grosschedl R. MZB1 enables efficient interferon α secretion in stimulated plasmacytoid dendritic cells. *Sci Rep*. 2020;10:21626. doi:10.1038/s41598-020-78293-3