JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 29, Number 8, 2022

Pp. 867–879
DOI: 10.1089/cmb.2021.0652

Shared Differential Expression-Based Distance Reflects Global Cell Type Relationships in Single-Cell RNA Sequencing Data

AIDAN MCLOUGHLIN[1] and HAIYAN HUANG[2]

ABSTRACT

Unsupervised cell clustering on the basis of meaningful biological variation in single-cell RNA sequencing (scRNA seq) data has received significant attention, as it assists with ontological subpopulation identification among the data. A key step in the clustering process is to compute distances between the cells under a specified distance measure. Although particular distance measures may successfully separate cells into biologically relevant clusters, they may fail to retain global structure of the data, such as relative similarity between the cell clusters. In this article, we modify a biologically motivated distance measure, SIDEseq, for use of aggregate comparisons of cell types in large single-cell assays, and demonstrate that, across simulated and real scRNA seq data, the distance matrix more consistently retains global cell type relationships than commonly used distance measures for scRNA seq clustering. We call the modified distance measure "SIDEREF." We explore spectral dimension reduction of the SIDEREF distance matrix as a means of noise filtering, similar to principal components analysis applied directly to expression data. We utilize a summary measure of relative cell type distances to better display the cell group relationships. SIDEREF visualizations more consistently reflect global structures in the data than other commonly considered distance measures. We utilize relative cell type distances and the SIDEREF distance measure to uncover compositional differences between annotated leukocyte cell groups in a compendium of *Mus musculus* scRNA seq assays comprising 12 tissues. SIDEREF and associated analysis is openly available on GitHub.

Keywords: clustering, differential expression, distance, global structure, scRNA seq.

1. INTRODUCTION

RECENT YEARS HAVE WITNESSED a fast and large expansion of single-cell sequencing technologies given their exciting ability to characterize and quantify heterogeneity across cells from tissues of interest (Tang et al., 2019). In particular, single-cell RNA sequencing (scRNA seq) data can be used to identify cellular subpopulations and study functional gene networks within such groups (Chen et al.,

Division of Biostatistics[1] and Department of Statistics,[2] University of California, Berkeley, Berkeley, California, USA.

2019; Liu et al., 2019). However, the data often suffer from challenging levels of sparsity and noise (Brennecke et al., 2013; Ding et al., 2020; Mereu et al., 2020), making it difficult to uncover stable cell groups.

A crucial step in unsupervised analysis of cellular relationships in scRNA seq data is the selected distance measure, in particular for use in cell type visualization and clustering. For example, the current iteration of the popular Seurat package (Hao et al., 2021) applies a modified Louvain community detection algorithm to a k-nearest-neighbors (KNN) graph of the cells, which is constructed from the selected cellular distance matrix, to determine clusters. Other earlier clustering approaches such as SNN-Cliq and PhenoGraph rely on a KNN graph as well (Levine et al., 2015; Xu and Su, 2015). Low dimensional visualization techniques, such as t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP), can use any distance measure (van der Maaten and Hinton, 2008; McInnes et al., 2018; Becht et al., 2019; Stuart et al., 2019).

A popular distance measure choice in applied work is to compute Euclidean distance on a low dimensional principal component analysis (PCA) embedding of the preprocessed gene expression matrix (Barres et al., 2018; Hao et al., 2021). This dissimilarity measure is successful for capturing ''local'' structure by isolating ontological cell types/groups. However, the measure may not capture relative similarity between these cell types/groups, which we refer to as the ''global'' relationships of the data. These concerns have also been levied toward the t-SNE and UMAP visualization techniques (Kobak and Linderman, 2021).

Other distance matrices specifically for scRNA seq have been developed. SIMLR learns a linear combination of Gaussian kernels to represent the cell–cell similarities, while imposing a low rank constraint on the similarity matrix (Wang et al., 2018). SIMLR still computes Euclidean distance in the kernels, which can be substituted with other measures (Kim et al., 2019). RAFSIL creates a feature set from PCA decompositions of the data and trains a random forest to learn cell–cell similarities (Pouyan and Kostka, 2018).

Other methods utilize auxiliary data such as protein–protein interaction networks and gene ontology databases (Li et al., 2021; Xu et al., 2021), or seek to learn sparse similarity values for cells arising from different subtypes (Liang et al., 2021). Deep learning autoencoders are an exciting direction to consider optimization objectives that retain scRNA seq global structure in the latent space (Lopez et al., 2018; Way and Greene, 2018; Tian et al., 2021).

Ontological cell groups can be characterized by genes that are expressed at differing levels from other cell groups, commonly referred to as differentially expressed (DE) genes (Wang et al., 2019). For each cell pair, the SIDEseq distance measure computes global concordance of their DE gene lists, where each cell's DE gene lists are derived by comparing the cell with every individual cell in the rest of the data (Schiffman et al., 2017). Its property of incorporating information from all cells (across different cell groups) when evaluating the similarity between two cells allows for the investigation of pairwise cell relationships from a global perspective, and this may better reflect the ground truth global relationships as defined by DE gene commonalities and differences between groups.

In this study, we first adapt the SIDEseq distance measure for tractable computation on data sets of several thousand cells. This is achieved by reducing the global comparison of DE gene lists with a well-selected reference set of cells in the data. We call this modified measure SIDEREF. For visualization of cell group relationships, we generate a relative group distance summary measure based on SIDEREF.

Similar to other quadratic time distance measures, SIDEREF does not scale for full cell-level distance computations in massive scRNA seq data sets; however, the cell group relative distances are reliably computed under reasonable sample sizes of each cell type. That is, we consider SIDEREF more advantageous for recovering global relationships between identified cell types, rather than for a general clustering purpose when applied to massive scRNAseq data sets.

To embed noise reduction and dimension reduction into SIDEREF, we consider taking the Euclidean distance of a spectral embedding of the SIDEREF distance matrix, though this procedure is applicable to any distance matrix.

We construct a simulation of scRNA seq data with ground truth global relationships between cell groups and show that SIDEREF successfully identifies the global structure present in the simulation, relative to other commonly used distance measures. We demonstrate that the spectral embedding can reduce noise yet also genuine signal. Furthermore, cell groups tend to grow more isolated as the dimension size of the spectral embedding increases, which is also observed as the rank constraint is increased in SIMLR.

We applied SIDEREF distance and the relative distance summary to an annotated compendium of *Mus musculus* scRNA seq data sets (Barres et al., 2018). SIDEREF distance more clearly relates ontologically similar groups of cells than competing distance measures. Furthermore, the relative distances are leveraged to identify compositional differences between broadly annotated leukocyte cell groups from different tissue sources. This demonstrates how a reliable measurement of global structure can help improve the classification of cell groups in one's data. SIDEREF and associated analysis is available on GitHub at https://github.com/aidantmcloughlin/SIDEREF.

2. METHODS

2.1. SIDEseq distance

The input for the SIDEseq measure is a $G \times N$ variable gene expression matrix with rows denoting genes and columns denoting cells. First, a DE statistic is computed for each cell pair and gene as follows:

$$\frac{\left|x_i^g - x_j^g\right|}{\sqrt{x_i^g + x_j^g}}, \tag{1}$$

where x_i^g denotes the normalized and scaled expression level of gene g in cell i. When expressions are 0 in both cells, the DE statistic is set to 0. Note that other DE statistics could be appropriate for particular data sets.

To compute the similarity measure between cells i and j, SIDEseq then indexes the top m DE statistics for each cell pair $(i,\ t)$ (or (j,t)) for cells $t \in \{1,\ \ldots\ ,\ N\} \setminus \{i,j\}$. Let $h_{i,t}^{(k)}$ and $h_{j,t}^{(k)}$ denote the gene corresponding to the top kth DE statistic of cell pair (i,t) and (j,t), respectively. Then, the SIDEseq similarity score for cell pair (i,j) is as follows:

$$S_{i,j} = \sum_{t \in \{1,\ldots,N\} \setminus \{i,j\}} \left| \left\{ h_{i,t}^{(1)},\ \ldots, h_{i,t}^{(m)} \right\} \cap \left\{ h_{j,t}^{(1)},\ \ldots, h_{j,t}^{(m)} \right\} \right|, \tag{2}$$

where the intersection set size of DE gene lists between cell pair (i,t) and cell pair (j,t) is summed for all $t \neq i,j$. These similarity scores construct a symmetric similarity matrix where diagonal elements are set to 0. To convert the matrix to a distance matrix, one subtracts the similarity scores in each off-diagonal cell from the maximum similarity score across the matrix.

Equation (2) shows how SIDEseq crucially borrows DE gene information from all cells in the data to inform the similarity in expression between cells i and j. The compared differential expression with respect to all cells in the data builds a rich summary of shared differentiating gene function for each cell pair. This enables the measure to detect subtle structures between cell subpopulations and protects each similarity score against noise artifacts. These properties help explain its demonstrated performance advantages for local clustering (Schiffman et al., 2017). However, the computation of similarity scores scales cubically with the number of cells, in contrast to square scaling of traditional distance measures.

2.2. SIDEREF modification

One natural way to alleviate the computability issue of SIDEseq is to reduce the global comparison in Equation (2) to one across a smaller reference cell set of size $C < N$, reducing the number of ranked DE gene list intersections per similarity score to C. If the reference set sufficiently represents the variety in gene expression across the data, then collecting ranked list comparisons only among the reference set should still largely capture the shared DE structures. We name this modified approach SIDEREF. Formally, given a reference set, C, the SIDEREF similarity scores are computed as

$$S_{i,j} = \frac{1}{|C \setminus \{i,j\}|} \sum_{t = \in C \setminus \{i,j\}} \left| \left\{ h_{i,t}^{(1)},\ \ldots, h_{i,t}^{(m)} \right\} \cap \left\{ h_{j,t}^{(1)},\ \ldots, h_{j,t}^{(m)} \right\} \right|. \tag{3}$$

To minimize the needed size of the reference set, SIDEREF employs cell clustering followed by stratified sampling of clusters to reach a representative cell reference set that produces more stable, more similar distance matrices to SIDEseq than simple random sampling (Supplementary Fig. S1). Specifically, the analyses in this article use k-means clustering applied to a two-dimensional UMAP embedding of a PCA

embedding of preprocessed gene expression data (Lloyd, 1982). The number of clusters and number of principal components are determined using the elbow heuristic (Pouyan and Kostka, 2018).

The clusters are sampled proportionally to their size to create cell reference sets of size 100 for all article results. The reference set size of 100 is chosen to balance compute time with consistency to the SIDEseq distance matrix (Supplementary Fig. S1). All other aspects of the SIDEseq computation remain the same for SIDEREF. In particular, the size of the DE gene list, m, is set to 300 for all experiments in this article, in accordance with robustness studies in Schiffman et al. (2017). For modest distance matrix dimensions, the compute time of SIDEseq eclipses that of SIDEREF (Supplementary Fig. S2).

2.3. Spectral embedding distance

SIDEREF integrates information from the full gene expression matrix and may show deteriorating performance for highly noisy scRNA seq assays. Spectral embedding can be applied directly to any distance matrix, including SIDEREF, as a means of dimension reduction. Specifically, given a distance matrix, X, spectral embedding of p dimensions is defined as the row-normalized matrix of eigenvectors corresponding to the p smallest nonzero eigenvalues of the spectral decomposition of the symmetric normalized graph Laplacian (Ng et al., 2001):

$$L = \mathbb{I} - D_S^{-1/2} S D_S^{-1/2}, \quad D_S = diag\{S\bar{1}\}, \tag{4}$$

where S is a conversion of the distance matrix to an adjacency matrix:

$$S = \mathbb{I} - X_{norm}, \quad X_{norm} = X/max\{X\}, \tag{5}$$

and D_S is the degree matrix of S. Note that any adjacency matrix conversion can be used in place of Equation (5). Euclidean distance is used to convert the spectral embedding to a distance matrix.

2.4. Groupwise relative distances

Low-dimensional visualization strategies such as UMAP may struggle to accurately display global cell group relationships encoded in the SIDEREF distance matrix. To meaningfully summarize such information, we compute groupwise relative distances based on an underlying distance matrix.

Consider a partitioning of N cells into K groups, annotated $A_1, \ldots, A_K$. Let $X_{i,j}$ denote the $\{i,j\}$th entry of a given cell distance matrix. The unnormalized groupwise distance between cell groups A_i and A_j is defined as

$$d(A_i, A_j) = \frac{1}{|A_i||A_j|} \sum_{c_i \in A_i} \sum_{c_j \in A_j} X_{c_i, c_j}. \tag{6}$$

Using the groupwise distances from Equation (6), we compute the normalized groupwise relative distance between cell types A_i and A_j as

$$d_{rel}(A_i, A_j) = \frac{d(A_i, A_j) - min_k d(A_i, A_k)}{max_k d(A_i, A_k) - min_k d(A_i, A_k)}. \tag{7}$$

Note that this measure is not symmetric. When putting all the measures in a matrix (i.e., having A_i as row index group and A_j as column index group), one can think of this measure as the relative distance from the row index group, A_i (the *source* group), to the column index group, A_j (the *target* group). Note the relative distance vector for a source row must include 0 and 1, which is not necessarily true for a target column.

2.5. Performance measures

In addition to qualitatively assessing heatmaps of the groupwise relative distances, we consider a performance measure that tabulates the proportion of relative distance values that violate a ''ground truth'' hierarchy of the cell types. Specifically, consider a partitioning of the K cell groups into global groups (e.g., groups/clusters obtained by cutting the true hierarchical dendrogram at some level if such hierarchy is available), for which cell types in a global group are more similar to each other than to the other cell groups. Let G_{A_i} denote the global group of cell type A_i. The *within group maximum distance* of A_i is defined as

$$M_1(A_i) = \max_{A_k \in G_{A_i}} \max\left(d_{rel}(A_i,\ A_k),\ d_{rel}(A_k,\ A_i)\right). \tag{8}$$

Conversely, the *outside group minimum distance* of A_i is defined as

$$M_2(A_i) = \min_{A_k \notin G_{A_i}} \min\left(d_{rel}(A_i,\ A_k),\ d_{rel}(A_k,\ A_i)\right). \tag{9}$$

Using Equation (8), we define the *global group maximum violations* of a relative distance matrix as

$$\sum_{i=1}^{K} \sum_{A_j \notin G_{A_i}} \mathbb{I}\left(d_{rel}(A_i,\ A_j) \le M_1(A_i)\right) + \mathbb{I}\left(d_{rel}(A_j,\ A_i) \le M_1(A_i)\right). \tag{10}$$

Conversely, the *global group minimum violations* are defined as

$$\sum_{i=1}^{K} \sum_{A_j \in G_{A_i}} \mathbb{I}\left(d_{rel}(A_i,\ A_j) \ge M_2(A_i)\right) + \mathbb{I}\left(d_{rel}(A_j,\ A_i) \ge M_2(A_i)\right). \tag{11}$$

Evaluating one ''cross'' of the relative distance matrix at a time, Equation (10) [Equation (11)] tabulates the number of relative distances outside of global groups (within global groups) that violate the condition that cell types within global groups should have the lowest relative distances in the cross. Note that, if there exist multiple levels of global groups in the data, we can tabulate whether a violation exists for any of those levels. The results in Section 3 report Equations (10) and (11) as a percentage of distance values that are violations.

3. RESULTS

3.1. Global cell relationships under simulated scRNA seq data

To examine global structure identification of cell types with ground truth annotations, a simulation of scRNA seq data is generated using the *Splatter* software (Zappia et al., 2017). The simulated data consist of 10,000 genes and 1000 cells that are equally distributed across 20 cell types. A summary of the global

Table 1. Summary of Global Relationships Encoded in Single-Cell RNA Sequencing Simulation

Individual cell type	*Global group index*	*Global group properties*	*Global group name*
1	A	High proportion of cell type-specific DE genes; no shared DE genes	High individual DE
2	B	Low proportion of cell type-specific DE genes; no shared DE genes	Low individual DE
3	C	High proportion of cell type-specific DE genes; low proportion	High individual DE
4	C	of shared DE genes with other cell types in the global group	Low shared DE
5	C		
6	D	High proportion of cell type-specific DE genes; high proportion	High individual DE
7	D	of shared DE genes with other cell types in the global group	High shared DE
8	D		
9	E	Low proportion of cell type-specific DE genes; high proportion	Low individual DE
10	E	of shared DE genes with other cell types in the global group	High shared DE
11	E		
12	F	Low proportion of cell type-specific DE genes; low proportion	Low individual DE
13	F	of shared DE genes with other cell types in the global group	Low shared DE
14	F		
15	G	Low proportion of cell type-specific DE genes; low proportion	Low individual DE
16	G	of shared DE genes with other cell types in the global group;	Low shared DE
17	G	high variance of DE genes	High variance DE
18	H	High proportion of cell type-specific DE genes; low proportion	High individual DE
19	H	of shared DE genes with other cell types in the global group;	Low shared DE
20	H	high variance of DE genes	High variance DE

DE, differentially expressed.

structure of these 20 cell groups is provided in Table 1. Structure is encoded in the data by randomly assigning DE genes to prespecified cell groups under parameters specified in Supplementary Table S1. In particular, cell types 1 and 2 are ''single'' groups, sharing no DE genes with any other group.

Remaining cell types exist in high-level ''global groups'' of cardinality 3, with a varied number of shared DE genes in each global group. For this analysis, we assume the individual cell type identities are known, and assess the ability of SIDEREF, as well as spectral embedding, to uncover the global groups in the data.

For simulation, as well as real data in Section 3.2, cell expression data are log-normalized, gene features are subset to the 3000 most variable genes and scaled using the *Seurat* package (Hao et al., 2021).

Under groupwise relative distances, SIDEREF captures connectivity between cell types in the simulation (Fig. 1a), compared with a popular distance measure, Euclidean distance computed on the 25-dimensional PCA embedding of the data, referred to as *PCA (25 Dims.)* (Fig. 1b) (Hao et al., 2021). The Euclidean distances are weighted by eigenvalues corresponding to each PCA embedding dimension. In the figures, *x*-axis represents *target* cell types and *y*-axis represents *source* cell types. As explained in Section 2.4, the values within rows are directly comparable and show the relative degrees of connectedness between the designated source cell type and all other individual cell types.

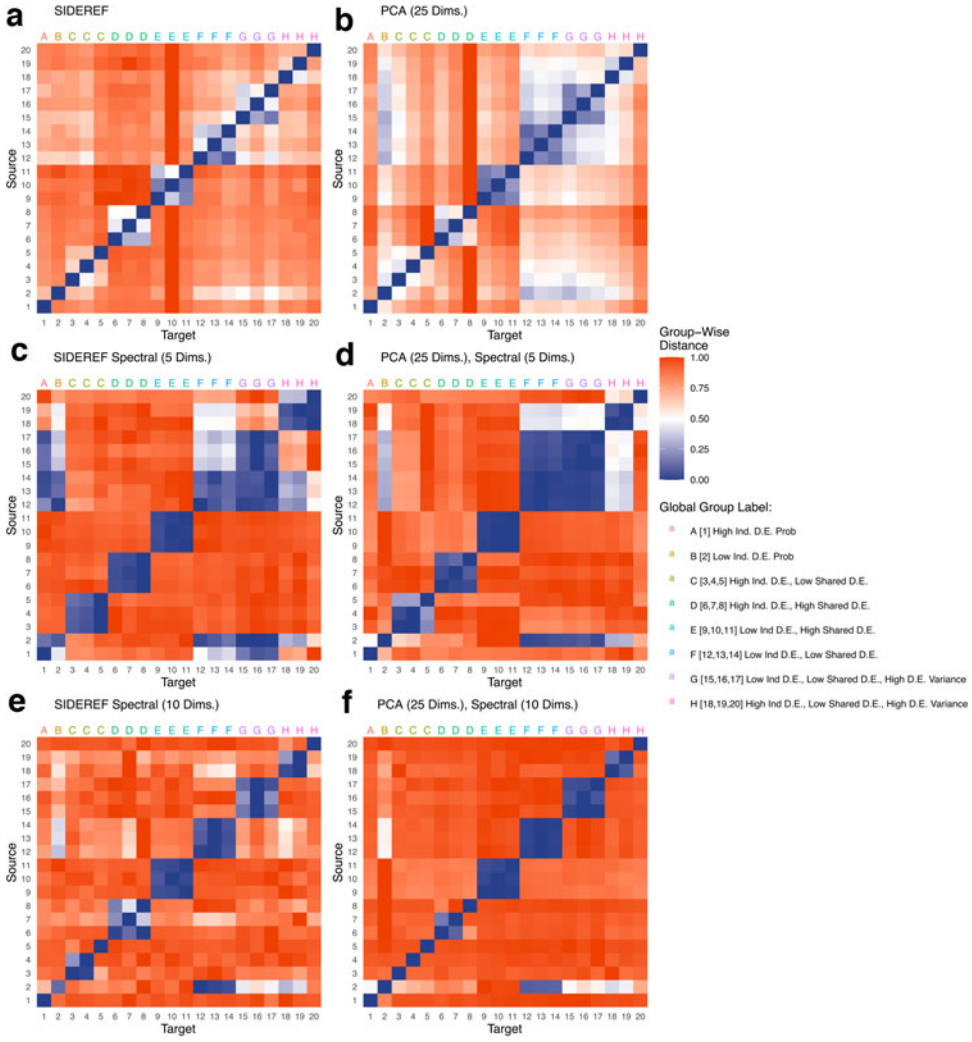

FIG. 1. Normalized groupwise relative distances for 20 cell type Splatter simulation data. Top row color coding reflects the ground truth global group assignment. The SIDEREF distance matrix with DE gene lists of size 300 is directly used in **(a)**, and is spectrally embedded in **(c)** and **(e)** to 5 and 10 dimensions, respectively. **(b)**, **(d)**, and **(f)** show analogous results for Euclidean distance applied to the 25-dimension PCA embedding. DE, differentially expressed; PCA, principal component analysis.

SIDEREF reflects lower relative distances within ground truth global groups. In addition, global groups with stronger shared DE have tighter connectivity, such as global group E. We observe global groups F and G are more defined under SIDEREF than PCA. The individual cell groups, indexed A and B, are better isolated under SIDEREF. Two-dimensional UMAP visualizations of the data will also separate the global groups under PCA (Supplementary Fig. S3).

Spectral embedding-based distance applied to the SIDEREF and PCA matrices is fairly similar under the simulation, and does not perform better than using SIDEREF directly (Fig. 1c–f). The spectral embedding dimension serves as a hyperparameter to control the number of tightly connected components among the cell groups. When the dimension is small, spectral distance fails to separate some global groups. As dimension increases, spectral distance more clearly elucidates some global groups, yet can sever the connection between cell types for other global groups.

Alternative baseline distance measures are compared with SIDEREF in Figure 2. In particular, due to the clear DE patterns of the simulated data, Spearman-based dissimilarity produces similar or slightly better simulation results to SIDEREF. However, the two measures show substantial differences in the real data. In Figure 3, SIDEREF is compared with scRNA seq-specific methods RAFSIL and SIMLR. RAFSIL struggles to clearly define global groups D and H and has noisier distance values outside of the global groups, whereas SIMLR tends to fully isolate each low-level cell type. As the rank constraint on SIMLR is increased, the connections between cell types are further severed, similar to spectral distance (Supplementary Fig. S4).

In Table 2, we report the global group violation proportions averaged over 10 seeds of the simulated data. These performance measures reflect the qualitative assessment of the heatmaps, in that SIDEREF is the second-best reflection of the simulated data, behind Spearman rank correlation. As mentioned previously, this superior performance of Spearman rank correlation is due to the clear DE patterns presented in the simulated data. In the real data, Spearman correlation reports much higher violation rates (Table 3).

3.2. *Global cell relationships in multitissue scRNA seq compendium*

We use a public compendium of *M. musculus* microfluidic droplet-based scRNA seq data. These data comprise 12 separate assays, each originating from different tissues (Barres et al., 2018; Macosko et al., 2015). In the data, each assay was separately analyzed with unsupervised clustering and marker gene detection, and the resulting cell ontology annotations are provided. In total, there are 76 unique

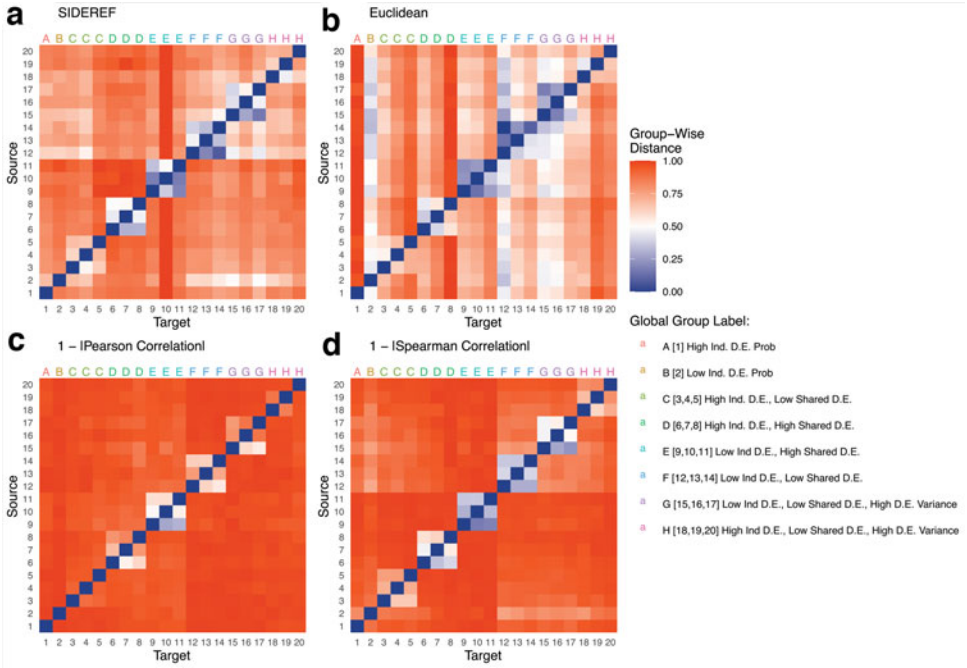

FIG. 2. Normalized groupwise relative distances for 20 cell type Splatter simulation data under alternative distance measures. Top row color coding reflects the ground truth global group assignment.

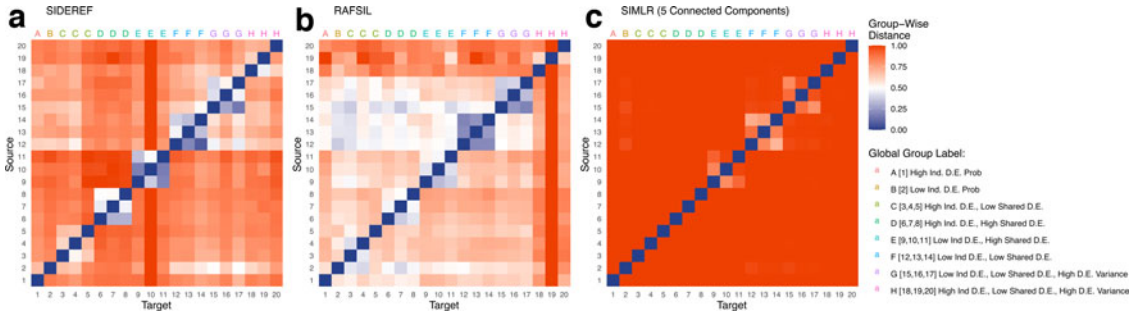

FIG. 3. Normalized groupwise relative distances for 20 cell type Splatter simulation data under scRNA seq tailored distance measures. Top row color coding reflects the ground truth global group assignment. scRNA seq, single-cell RNA sequencing.

combinations of tissue sources and cell type classifications in the data, which are used as ''ground truth'' cell types from which to analyze global relationships.

Despite the computational improvements, SIDEREF is intractable for the entire droplet data of 55,656 cells (even storing this distance matrix is a burden on memory). For aggregate analysis of cell type relations, SIDEREF groupwise distances are computed stably for reasonable sample sizes of each cell type (Supplementary Fig. S5). Specifically, from each cell type tissue combination, either 100 cells or the group's cardinality is sampled for this analysis, yielding a final data set of 6,832 cells.

Some cell type annotations appear in multiple *M. musculus* tissues, such as immune cells and endothelial cells. The same cell type in different *Tabula Muris* tissues may have slightly varying true expression profiles or batch effects, given that tissue samples are sourced from various mice. However, compared with biologically distal cell types, cells of the same cell type (even in different tissues) should clearly group together under a successful distance measure.

Each distance matrix is computed from all 76 cell types in the data. Frequently, the global group structure of the cell types is unclear. However, several specific cell types appear in various tissues, namely B cells, Pro-B cells, T cells, natural killer cells, basal cells, and endothelial cells. Other cell types that appear in multiple tissues are broadly defined, such as leukocytes, or express high diversity depending on tissue location, such as macrophages (Locati et al., 2020). To compute the global group violations tabulations from Section 2.5, we label the six initially listed cell types as ground truth global groups, and remove broadly defined cell types, other immune cell types, and progenitor cells before computing relative distances and global group violations.

Supplementary Table S2 enumerates cell type inclusion and exclusion. In addition, we consider a second level of global groups, where the four immune cell types are joined into one group, and tabulate whether a violation exists along either global group level for each relative distance value.

The global group violations for SIDEREF and competing distance measures are reported in Table 3. The right-hand two columns report violations when including the violations check for a combined immune cells global group. Under this multilevel delineation of global structure, SIDEREF performs best. In particular, the performance of Pearson correlation drops drastically when including the immune cells violation check. This indicates that Pearson correlation tightly connects very similar cell types yet severs other more distal

Table 2. Global Group Violations for Selected Distance Measures Under the Single-Cell RNA Sequencing Simulation

Method	*Global group max violations (%)*	*Global group min violations (%)*
SIDEREF	4.4	11.1
PCA (25 dimensions)	9.6	13.8
Euclidean	19.3	30.5
$1 - \lvert$ Pearson correlation $\rvert$	9.2	16.0
$1 - \lvert$ Spearman correlation $\rvert$	0.1	0.6
RAFSIL	22.1	37.8
SIMLR (five components)	40.5	31.7

Results are averaged over 10 random generations of the data.
PCA, principal component analysis.

Table 3. Global Group Violations for Selected Distance Measures Under the *Tabula Muris* Data

	Single level of global groups		*Immune level of global groups included*	
Method	*Global group maximum violations (%)*	*Global group minimum violations (%)*	*Global group maximum violations (%)*	*Global group minimum violations (%)*
SIDEREF	11.3	26.2	22.2	44.5
PCA (25 dimensions)	15.2	24.2	28.8	56.1
Euclidean	41.5	84.0	51.5	83.1
$1 - \lvert$ Pearson correlation $\rvert$	8.2	25.4	46.0	54.7
$1 - \lvert$ Spearman correlation $\rvert$	19.2	56.1	42.5	72.0
RAFSIL	23.0	54.9	46.0	72.9
SIMLR (25 components)	26.4	33.2	35.2	47.7

The left-hand column considers one level of global grouping based on matching narrow cell type annotations. The right-hand column includes an additional level of global grouping of all immune system cells.

relations. Supplementary Table S3 reports the violation measures when filtering the data to only cell types that belong to the global groups. SIDEREF performs favorably under this scenario as well.

Groupwise relative distances under SIDEREF show expected global group structure for annotated leukocyte and B cells, along with three endothelial cell types and two otherwise unrelated cell types (Fig. 4a). These cell types were manually selected to offer a diverse mix of cell function and to compare broadly defined leukocytes against narrowly defined B cells. The matrix is filtered to the 14 cell types before computing groupwise relative distances.

In Section 3.3, we show the kidney and thymus leukocytes show strong relation to the B cells, suggesting these six cell types should form a global group. The other leukocyte groups still consist of immune cells, and should be closer to other immune cells than the endothelial cells, basal cells, and pneumocytes. Finally, the global structure should show a global group of leukocytes and of endothelial cells, though the endocardial cell type is a specialized endothelial cell, so that the lung and limb endothelial cells should be most connected of the three. The other two cell types (basal cell of epidermis, type II pneumocyte) should be fully isolated. SIDEREF distance reflects this structure.

Similar to the simulation, spectral embedding-based distance of SIDEREF controls the number of connected components (Fig. 4b, c). PCA distance suggests B cell enrichment of bladder leukocytes under source to target distance, indicating the B cell groups lie closer to the bladder leukocytes (''source'') than do the other leukocytes. PCA also fully isolates liver leukocytes under target–source distance, and does not clearly differentiate endocardial cells from endothelial cells (Fig. 4d). Euclidean and correlation-based distances show clear deficiencies (Fig. 4e, g, and h). RAFSIL successfully tracks kidney and thymus B cell enrichment, but does not clearly indicate the other global structures (Fig. 4f). SIMLR fully isolates most low-level cell types (Fig. 4i).

A practitioner may be interested in focusing on one direction of distance or simply aggregating both directions. To this end, the open software allows for symmetrized heatmaps and selected bipartite networks (Supplementary Fig. S6).

3.3. SIDEREF uncovers compositional differences between leukocyte cell groups

Leukocytes broadly include several immune cell types such as B lymphocytes, which are also present in several tissues in the data. A strong relative distance measure should correctly identify whether heterogeneous leukocyte cell groups are more similar to homogeneous B cell groups. When SIDEREF groupwise relative differences are computed between leukocytes and B cells in the *Tabula Muris* data, two of five organs containing leukocytes, the kidney and the thymus, show clear proximity to the B cell groups (Fig. 4a). The 25-dimensional PCA (Fig. 4d) suggests the bladder cells are intermediately related to the B cells.

The finding of SIDEREF is validated through gene set enrichment (GSE) analysis conducted using the *fgsea* package (Subramanian et al., 2005; Korotkevich et al., 2021). Specifically, the Wilcoxon Rank Sum Test is run between B cells and nonleukocyte cells in the droplet data to generate associated DE p values for each gene (Hao et al., 2021). We subset the lowest p values from this list to generate the B cell DE gene

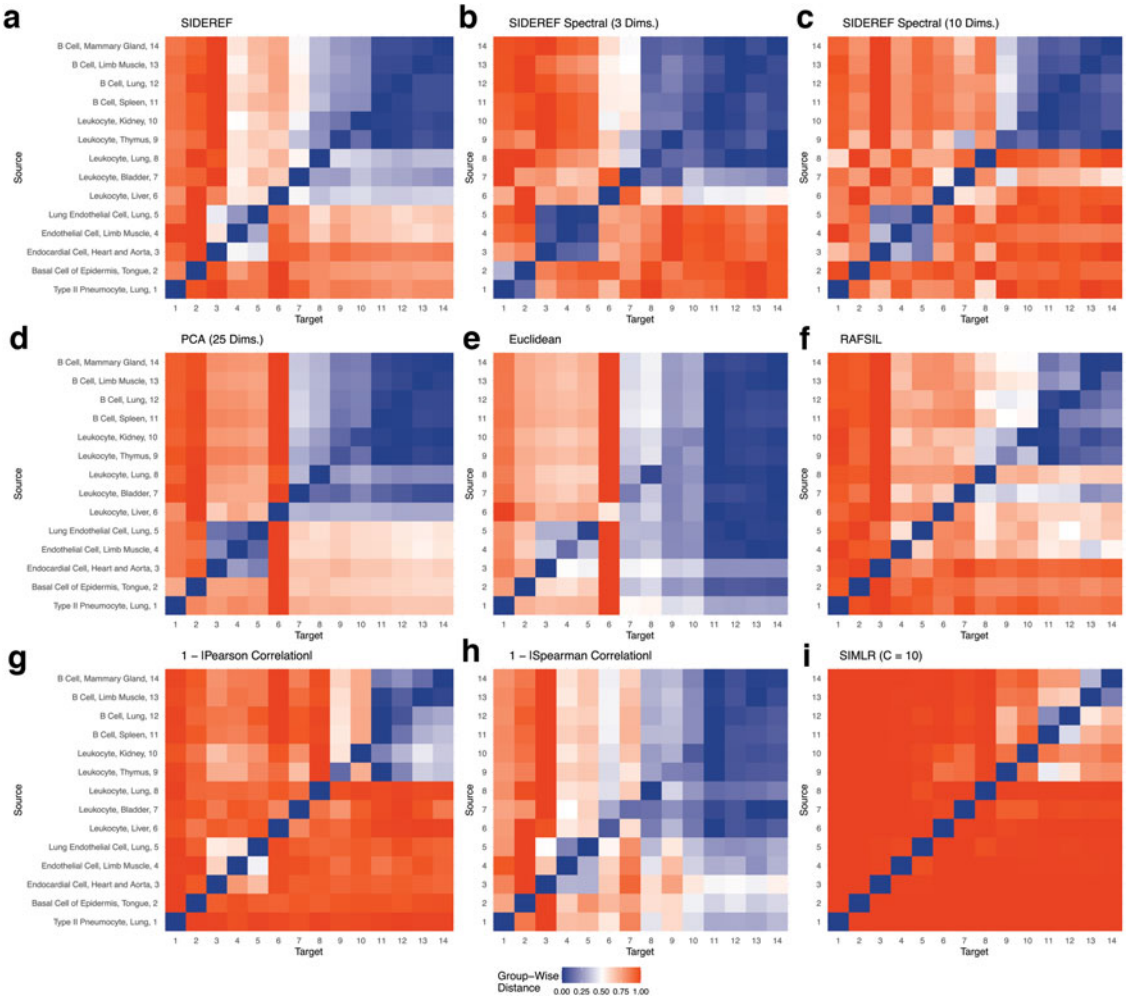

FIG. 4. Normalized groupwise relative distances for select cell group samples in *Mus musculus* droplet scRNA seq data. The SIDEREF distance matrix with DE gene lists of size 300 is directly used in **(a)**, and is spectrally embedded in **(b)** and **(c)** to 3 and 10 dimensions, respectively. Alternative distance matrices are presented in **(d–i)**.

lists. Similarly, a Wilcoxon Rank Sum Test generates p values associated with each gene for the leukocyte cell groups against non-B cells.

Given a list of leukocyte gene statistics and a B cell DE gene list, the GSE algorithm computes a running sum statistic, which increases (proportional to the gene p value ranking) if the gene is in the B cell DE gene list, and decreases by the proportion of genes outside of the B cell DE gene list otherwise. The enrichment score (ES) is the maximum value of the running sum statistic. A null distribution of ES is computed using random gene lists that are the size of the B cell DE gene list.

The normalized ES is the ES divided by the mean of the null distribution. The kidney and thymus leukocyte groups have enriched B cell expression profiles relative to other leukocytes (Fig. 5; Table 4). Note that as the cardinality of the B cell DE gene list increases, we may include more genes that have shared expression profiles with broad immune cell groups.

4. DISCUSSION

Though isolating functionally similar cell types is a primary focus of unsupervised learning in scRNA seq analysis, it is of interest to capture global relationships between cell groups. This study modifies a biologically motivated distance measure, SIDEseq, for use on large scale scRNA seq data, and explores its ability to better identify multigroup structures present in simulated and real scRNA seq data over common baseline distances.

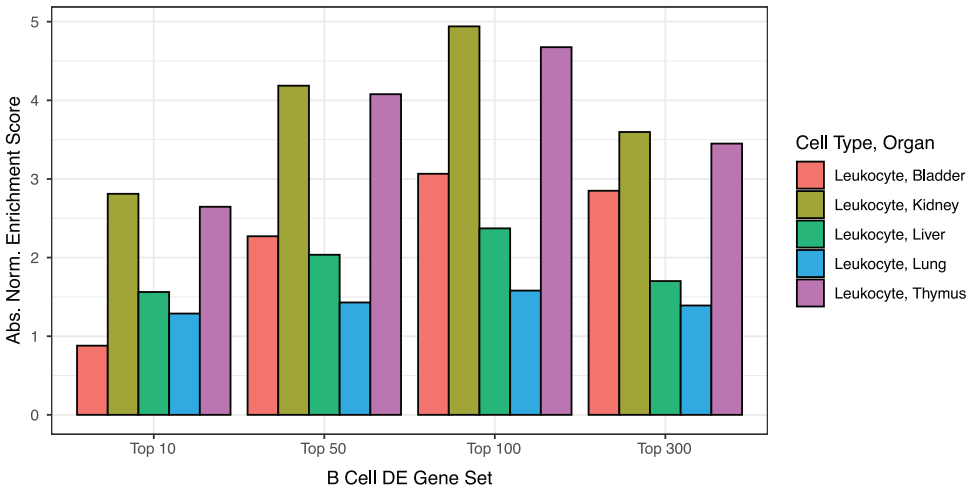

FIG. 5. GSE analysis absolute NESs for ranked DE gene lists of leukocyte cell samples in *Mus musculus* droplet scRNA seq data. Gene sets consist of the top DE genes between B cells and the nonleukocyte cells. GSE, gene set enrichment.

Computing normalized relative distances between cell types under the SIDEREF measure reveals the cell type relationships. SIDEREF relative distances are found to better respect global group relationships in simulated data and the *Tabula Muris* compendium than competing measures. This analysis framework discovers significant B cell compositional differences between leukocyte cell groups in a public scRNA seq compendium.

Spectral embedding of SIDEREF distance is introduced as both a potential noise reduction method and a means to control the number of tightly connected groups. It is shown to be quite sensitive to the dimension of the embedding. It is important to investigate specifically how cell group connectivity changes as embedding dimension changes, and potentially leverage this information for hierarchical clustering. Standard hierarchical clustering of cell types into global groups proves difficult, as the amount of DE genes specific to a cell type may dominate the amount of DE genes shared with other cell types.

Table 4. B Cell Gene Set Enrichment Analysis *p*-Values in Leukocyte Cell Groups

Pathway	*Cell type, organ*	*GSE* p-*value*
B cells top 10 DE genes	Leukocyte, bladder	5.7E-01
	Leukocyte, kidney	2.5E-07
	Leukocyte, liver	3.7E-02
	Leukocyte, lung	1.5E-01
	Leukocyte, thymus	5.3E-09
B cells top 50 DE genes	Leukocyte, bladder	1.2E-05
	Leukocyte, kidney	9.6E-26
	Leukocyte, liver	5.2E-05
	Leukocyte, lung	2.9E-02
	Leukocyte, thymus	2.1E-36
B cells top 100 DE genes	Leukocyte, bladder	1.2E-15
	Leukocyte, kidney	1.7E-55
	Leukocyte, liver	1.3E-10
	Leukocyte, lung	3.9E-04
	Leukocyte, thymus	1.1E-67
B cells top 300 DE genes	Leukocyte, bladder	3.8E-23
	Leukocyte, kidney	4.0E-42
	Leukocyte, liver	4.0E-06
	Leukocyte, lung	2.9E-04
	Leukocyte, thymus	2.8E-41

GSE, gene set enrichment.

ACKNOWLEDGMENTS

The authors thank the reviewers from *The Journal of Computational Biology* for very helpful feedback as we iterated versions of this article. The authors thank Dr. Elizabeth Purdom and Dr. Koen Van den Berge for discussions about the results, and Dr. Nancy Zhang for guidance on scRNA seq preprocessing.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

This study was partially supported by NIH R01 GM134307-01.

SUPPLEMENTARY MATERIAL

Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Figure S4
Supplementary Figure S5
Supplementary Figure S6
Supplementary Table S1
Supplementary Table S2
Supplementary Table S3

REFERENCES

Barres, B., Beachy, P., Chan, C., et al. 2018. Single-cell transcriptomics of 20 mouse organs create a *Tabula Muris*. *Nature* 562, 367–372.

Becht, E., McInnes, L., Healy, J., et al. 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol*. 37, 38–44.

Brennecke, P., Anders, S., Kim, J., et al. 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* 10, 1093–1095.

Chen, G., Ning, B., and Shi, T. 2019. Single-cell RNA-Seq technologies and related computational data analysis. *Front. Genet*. 10, 317.

Ding, J., Adiconis, X., Simmons, S., et al. 2020. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol*. 38, 737–746.

Hao Y., Hao, S., Andersen-Nissen, E., et al. 2021. Integrated analysis of multimodal single-cell data. *Cell* 184, 3573–3587.

Kim, T., Chen, I. R., Lin, et al. 2019. Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief. Bioinformatics* 20, 2316–2326.

Kobak, D., and Linderman, G. 2021. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat. Biotechnol*. 39, 156–157.

Korotkevich, G., Sukhov, V., Budin, N., et al. 2021. Fast gene set enrichment analysis. *BioRxiv* 060012.

Levine, J. Simonds, E., Bendall, S., et al. 2015. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* 162, 184–197.

Li, J., Jiang, W., Han, H., et al. 2021. ScGSLC: An unsupervised graph similarity learning framework for single-cell RNA-seq data clustering. *Comput. Biol. Chem*. 90, 107415.

Liang, Z., Li, M., Zheng, R., et al. 2021. SSRE: Cell type detection based on sparse subspace representation and similarity enhancement. *Genomics Proteomics Bioinformatics* 19, 282–291.

Liu, K., Theusch, E., Zhou, Y., et al. 2019. GeneFishing to reconstruct context specific portraits of biological processes. *Proc. Natl. Acad. Sci. U. S. A*. 116, 18943–18950.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* 28.2, 129–137.
Locati, M., Curtale, G., and Mantovani, A. 2020. Diversity, mechanisms, and significance of macrophage plasticity. *Annu. Rev. Pathol*. 15, 123–147.
Lopez, R., Regier, J., Cole, M., et al. 2018. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* 15, 1053–1058.
Macosko, E., Basu, A., Satija, R., et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214.
McInnes, L., Healy, J., Saul, N., et al. 2018. UMAP: Uniform manifold approximation and projection. *J. Open Sour. Softw*. 3, 861.
Mereu, E., Lafzi, A., Moutinho, C., et al. 2020. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol*. 38, 747–755.
Ng, A., Jordan, M., and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. *Adv. Neural Inf. Process. Syst*. 14.
Pouyan, M.B. and Kostka, D. 2018. Random forest based similarity learning for single cell RNA sequencing data. *Bioinformatics* 34, i79–i88.
Schiffman, C., Lin, C., Shi, F., et al. 2017. SIDEseq: A cell similarity measure defined by shared identified differentially expressed genes for single-cell RNA sequencing data. *Stat. Biosci*. 9, 200–216.
Stuart, T., Butler, A., Hoffman, P., et al. 2019. Comprehensive integration of single-cell data. *Cell* 177, 1888–1902.
Subramanian, A., Tamayo, P., Mootha, V., et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A*. 102, 15545–15550.
Tang, X., Huang, Y., Lei, J., et al. 2019. The single-cell sequencing: New developments and medical applications. *Cell Biosci*. 9, 53.
Tian, T., Zhang, J., Lin, X., et al. 2021. Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data. *Nat. Commun*. 12, 1873.
van der Maaten, L.J.P., and Hinton, G.E. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res*. 9, 2579–2605.
Wang, B., Ramazzotti, D., De Sano, L., et al. 2018. SIMLR: A tool for large-scale genomic analyses by multi-kernel learning. *Proteomics* 18, 2.
Wang, T., Li, B., Nelson, C., et al. 2019. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* 20, 40.
Way, G.P., and Greene, C.S. 2018. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pacific Symp. Biocomput*. 23, 80–91.
Xu, C., and Su, Z. 2015. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31, 1974–1980.
Xu, Y., Li, H.-D., Pan, Y., et al. 2021. A gene rank based approach for single cell similarity assessment and clustering. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 18, 431–442.
Zappia, L., Phipson, B., and Oshlack, A. 2017. Splatter: Simulation of single-cell RNA sequencing data. *Genome Biol*. 18, 174.

Address correspondence to:
Dr. Haiyan Huang
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
USA

E-mail: hyh0110@berkeley.edu