Title

Visualizing scRNA-Seq data at population scale with GloScope.

Permalink

https://escholarship.org/uc/item/6tp132qq

Journal

Genome Biology, 25(1)

Authors

Wang, Hao

Torous, William

Gong, Boying

et al.

Publication Date

2024-10-08

DOI

10.1186/s13059-024-03398-1

Peer reviewed

Genome Biology

METHOD Open Access

Visualizing scRNA-Seq data at population scale with GloScope

Hao Wang[1], William Torous[2], Boying Gong[1] and Elizabeth Purdom[2,3*]

*Correspondence:
epurdom@berkeley.edu

[1] Division of Biostatistics, University of California, Berkeley, CA, USA
[2] Department of Statistics, University of California, Berkeley, CA, USA
[3] Center for Computational Biology, University of California, Berkeley, CA, USA

Abstract

Increasingly, scRNA-Seq studies explore cell populations across different samples and the effect of sample heterogeneity on organism's phenotype. However, relatively few bioinformatic methods have been developed which adequately address the variation between samples for such population-level analyses. We propose a framework for representing the entire single-cell profile of a sample, which we call a GloScope representation. We implement GloScope on scRNA-Seq datasets from study designs ranging from 12 to over 300 samples and demonstrate how GloScope allows researchers to perform essential bioinformatic tasks at the sample-level, in particular visualization and quality control assessment.

Keywords: Single-cell sequencing data, scRNA-Seq, Density estimation, Batch effect detection and visualization

Background

Single-cell sequencing data has the potential to considerably enhance our comprehension of human health demonstrating how individual cell differences affect disease outcomes. Initially, single-cell sequencing studies examined the scope of cell diversity found in biological systems, including large projects such as the Human Cell Atlas Project. Such studies generally obtain large numbers of cells from few individual donors and focus on the shared cell type diversity. However, an increasing number of scRNA-Seq investigations target patient populations and emphasize the impact of single-cell variation on human health outcomes. These population-based scRNA-Seq studies typically involve scRNA-Seq data from larger cohorts of individuals who are selected from populations exhibiting various health-related phenotypes.

Despite the plethora of methodological advancements in scRNA-Seq, most current tools were designed for the goal of understanding the single-cell level information and lack appropriate strategies for analyzing scRNA-Seq population studies. Most of the current analyses of population scRNA-Seq data tends to consider the individual cells as the primary data unit. Existing tools that do account for population variability focus on identifying individual genes with differential expression [1–3]. Beyond differential expression analysis, sample-level analyses that exist are generally limited to comparisons of the relative proportions of different cell types between groups of samples [4]. We propose an analysis paradigm that uses the entire single-cell profile of a sample instead of focusing on cells as units. We refer to such an approach as a sample-level (or patient-level) analysis.

Our proposal is based on representing each sample as a distribution of cells. More specifically, we summarize each sample with a probability distribution describing the distribution of cells and their gene expression within the sample. Such a representation allows us to summarize the entire scRNA-profile of a sample into a single mathematical object. In this way, we synthesize the entire single-cell profile of an individual sample while maintaining information regarding the variability of the single-cells. This global representation, which we call GloScope, can be used in a wide variety of downstream tasks, such as exploratory analysis of data at the sample-level or prediction of sample phenotypes. Moreover, this representation does not require classification of sequenced cells into specific cell types (e.g., via clustering) and therefore is not sensitive to any auxiliary cell type identification procedure.

We apply the GloScope representation on a variety of published data collected on sample cohorts and demonstrate how the GloScope representation allows for visualization of important biological phenotypes and aids in detection of sample-level batch effects.

Results

Overview of the GloScope representation

If we consider trying to model individual samples, we see that the format of scRNA-Seq data when considered as data on samples (not cells) is non-standard. Most computational strategies assume each sample is measured on a shared set of features. Instead, for each sample i, we observe a matrix $X_i \in R^{g \times m_i}$, containing the gene expression measurements of that sample across all cells (g corresponds to the number of genes and m_i to the number of cells sequenced from sample i). There is no direct correspondence between the m_i cells in sample i with the m_j cells of sample j, so there is no immediate way to align data from different samples as input into a statistical model or predictive algorithm.

We propose to create a representation of each sample that does not require explicitly aligning individual cells across samples but leverages the nature of the observed data to represent each sample in a similar space. We consider the gene measurements for each of the m_i cells to be a sample from the full population of all cells of each sample. The full population of cells defines a probability distribution we designate as F_i on R^g. F_i is a representation of the sample's entire single-cell profile across all cells and importantly is a mathematical object that can be compared across samples. We do not observe F_i, but we do observe m_i samples from this distribution (the sequenced cells), allowing us to estimate F_i from the data. Thus, we transform each sample from the matrix X_i of observed gene expression measurements to an estimate of the sample's distribution, $\hat{F}_i$.

However, because gene expression data lie in a high-dimensional space, with the number of genes g in the thousands, estimating F_i directly from the cells is intractable. Thus, we assume that there exists a lower-dimensional representation or latent variable in R^d which governs the gene expression of a sample. We instead estimate the distribution of this latent variable. We do this by first estimating a lower-dimensional representation of our all our cells, for example via methods like PCA or scVI [5] applied to all the cells. This results in a matrix of reduced representation $Z_i \in R^{m_i \times d}$ corresponding to the new coordinates of each cell in this reduced space. We then estimate the distribution $\hat{F}_i$ from the m_i cells in this reduced space.

Unlike the X_i, which have different, unrelated, dimensions for each sample i, the $\hat{F}_i$ lie in the space of distributions on R^d and can be compared. As probability measures, these representations are now familiar mathematical objects and sample-level analysis can be done in the space of probability measures. There are many well-known metrics defined on the space of probability measures, such as the Wasserstein distance, and downstream analysis can be performed after choosing a metric to quantify pairwise sample differences. We call this representation of samples the GloScope representation, and we illustrate this transformation in Fig. 1. For our examples, we use the square root of the symmetrized Kullback-Leibler (KL) divergence to quantify the differences between sample distributions; while not a proper metric, this divergence can be effectively used to create a global representation of probability distributions [6] (see the "Methods" section for details).

The resulting pairwise-divergences can be used by many standard statistical or machine learning methods. We primarily concentrate in this work on the use of the GloScope representation for the purpose of visualization and exploratory data analysis. The pairwise divergences between GloScope-represented samples can be given as input to canonical divergence analysis methods such as multidimensional scaling [7], which creates coordinate system to represent the samples that capture the pairwise divergences. We will demonstrate that such a visualization enables detection of possible batch effects or outliers and exploratory assessment of the strength of phenotypic differences between our samples. We can also use the divergences to numerically quantify the separation of groups of samples using silhouette width or ANOSIM statistics (see the "Methods" section). This allows us to quantify how separated samples are due to a biological condition of interest (e.g., healthy vs diseased samples) or alternatively how separated samples

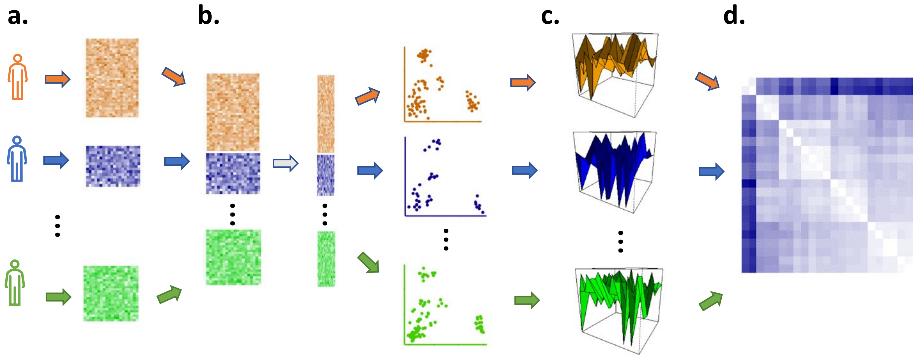

Fig. 1 Illustration of the GloScope representation of a sample's scRNA-Seq data matrix X_i as a distribution $\hat{F}_i$. **a** Each sample contributes a $g \times m_i$ matrix of gene expression values. **b** A shared, lower-dimensional latent representation is estimated across all cells and samples, resulting in each cell being represented in a lower-dimensional space **c** GloScope estimates the distribution $\hat{F}_i$ for each sample, and then **d** calculates the statistical divergence between each pair of samples, $d(\hat{F}_i, \hat{F}_j)$, resulting in a $n \times n$ matrix of all pairwise divergences

are due to a design artifact (e.g., different processing centers). Beyond EDA, our representation can also be used for other important downstream tasks, include clustering of samples, global hypothesis tests for differences between sample populations, and prediction of phenotypes (for example via kernel prediction methods, e.g., Hofmann et al. [8], Wang et al. [9]).

GloScope in the scRNA-Seq pipeline

There are many existing methods for working with scRNA-Seq data, and GloScope is designed to fit into standard pipelines and complement existing quality-control and EDA strategies. GloScope takes as input low-dimensional latent representations of the individual cells, which can come from a variety of sources. While in this paper we focus on only a few common approaches, the construction of a unified latent space is a common approach to a wide variety of preprocessing challenges, particularly in integrating data from multiple studies. This includes standard embeddings of the original cell data, like PCA or scVI; batch-correction methods like Harmony [10]; or integration methods that harmonize data processed on different gene definitions or the presence of missing genes (see [11], for a review). This flexibility allows GloScope to be integrated with a wide-variety of approaches to preprocessing the data and to be used at different stages of the pre-processing, allowing checks at each stage of whether patient-level artifacts, like processing batches, are inappropriately contributing to differences in the samples.

Using cell type composition

Our GloScope approach to creating a global representation uses the entire gene distribution F_i, which encodes both cell type composition and gene expression. However, the underlying logic of GloScope could also be applied to compare only cell type composition. Specifically, if each cell can be classified into one of K subtypes, then we observe for each sample the proportion of cells in each cell type, $\hat{\pi}_i = (\hat{\pi}_{i1}, \ldots, \hat{\pi}_{iK}) \in R^K$. $\hat{\pi}_i$ is an estimate of a probability distribution, only now a simpler discrete distribution into K groups. We can use the GloScope strategy in a similar way to globally compare samples, only now restricted to only differences in cell type composition. Comparison of cell type composition has been proposed for globally comparing single-cell samples [4, 12–15], and there has been some limited work in analysis of data from flow-cytometry using cell type compositions to globally compare samples which has similarities to using GloScope on the proportions [12, 16–18]. Unlike a full GloScope representation, applying GloScope on the cluster proportion vector requires classifying cells into subtypes before application of the method. Accurate identification of cells into subtypes is often a manual and time-consuming process, which makes this approach less useful for the exploratory data analysis that is often upstream of the subtype identification step (see the "Comparison with other quality control tools" section). However, GloScope applied to the clusters can be used for more formal hypothesis testing of significant global differences in cell type composition. In what follows, we will refer to GloScope applied to the vector cluster proportions as GloProp, as opposed to our standard implementation which calculates an estimate of the full gene expression density.

Visualization of patient and sample phenotypes using GloScope representations

In this section, we demonstrate the utility of the GloScope representation to visualize and evaluate sample-level phenotypic differences. As an initial illustration, we consider two datasets with replicate samples collected for each phenotype, where the phenotypes have well-known biological differences in cell type structure. These serve as an initial proof-of-concept of the GloScope representation.

The first dataset is scRNA-Seq data from the mouse cortex [19]. Here, the samples are cells from different regions of the brain with replication in each from three genetically identical mice. This is a dataset where we know the regions have distinct compositions of cell types and gene expressions. When we visualize these samples using the GloScope representation in Fig. 2A, we see these distinctions clearly. The samples from the two main subdivisions of the cortex, isocortex (CTX) and hippocampal formation (HPF), clearly separate. Furthermore, we see that replicate samples from the same region strongly cluster with each other, while different regions are generally well separated. Within the CTX region, we observe blocks of biologically meaningful brain region groups such as the sensory and visual area: primary somatosensory (SSp), posterior parietal association (PTLp), visual area (VIS), and the somatomotor areas: primary motor (MOp) and secondary motor (MOs). We also observe clustering of physically adjacent brain regions such as temporal association, perirhinal, and ectorhinal areas (TEa-PERI-ECT), agranular insular (AI), prelimbic, infralimbic, orbital area (PL-ILA-ORB), and anterior cingulate (ACA).

Next, we consider skin cell samples from a study of twelve patients [20], consisting of nine healthy skin samples from the foreskin, scalp, and trunk alongside three inflamed skin samples collected from truncal psoriatic skin. We expect marked differences between cellular distributions collected at the different locations in the body due to

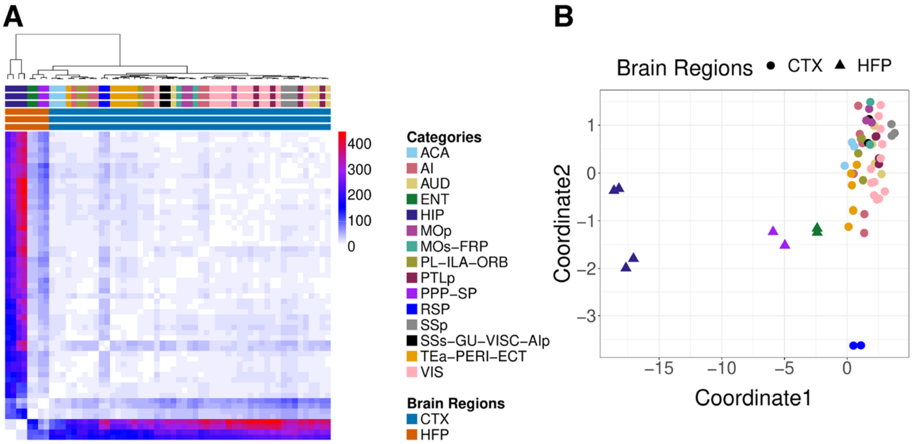

Fig. 2 Demonstration of the GloScope representation on 59 mice samples [19]. **A** Heatmap representation of the estimate of the divergences between the samples based on the GloScope representation. **B** A two-dimensional representation via MDS of the divergences shown in **A**. GloScope used the GMM estimate of the density in the first 10 PCA dimensions. The individual regions represent subregions of two main divisions of the cortex: the isocortex (CTX) and hippocampal formation (HPF). HPF is further divided into hippocampal region (HIP) and the retrohippocampal region (RHP) which is represented by the entorhinal region (ENT) and the remaining RHP, a joint dissection region of postsubiculum (POST)-presubiculum (PRE)-parasubiculum (PAR) region, subiculum (SUB), and prosubiculum (ProS) region (i.e., PPP-SP). The remaining regions are divisions of the CTX

varying proportions of cell types in certain tissues. For instance, the authors note different types of main basal keratinocytes and melanocytes dominate in scalp and trunk samples, as compared to foreskin tissues. Our visualization of the GloScope representations of this data in Fig. 3 shows a clear clustering of skin samples collected from similar locations on the body and a separation of both the foreskin and psoriasis samples from scalp and trunk samples, echoing the conclusions of the authors who identified a keratinocyte subpopulation which separates these phenotypes from the scalp and trunk control samples Cheng et al. [20].

Next we demonstrate the GloScope representation on additional datasets of patient cohorts where the samples are patients with differing disease phenotypes: (1) COVID lung atlas data from [21], which contains 27 samples, either diagnosed with COVID-19 or healthy control samples, and (2) colorectal cancer data with 99 samples (after quality control), grouped into three phenotypes: healthy, mismatch repair-proficient (MMRp) tumors, and mismatch repair-deficient (MMRd) tumors [22]. The use of GloScope on these datasets demonstrates its utility for the visualization of both sample and phenotype variability. For the COVID lung samples (Fig. 4A), we can easily see the separation between COVID-infected and healthy donors, matching the observation of Melms et al. [21] that lung samples from COVID patients were highly inflamed. For the colorectal cancer data, visualization of the GloScope representation shows healthy samples well separated from the tumor samples (Fig. 4B). Though the two types of tumors do not separate in this visualization, an analysis of similarities (ANOSIM) test of significance [23, 24] applied to their GloScope divergences between these two groups does find their representations to be significantly different ($p = 0.001$), indicating that the representation is encapsulating systematic differences between the two tumors (see "Methods" section).

Quantitative evaluation of GloScope via simulation

We use simulation experiments to quantify GloScope's efficacy at detecting various classes of single-cell differences that might be observed due to differences in samples' phenotype. We simulate sample-level data where different aspects of the single-cell composition of a sample vary depending on their group assignment; for simplicity, we

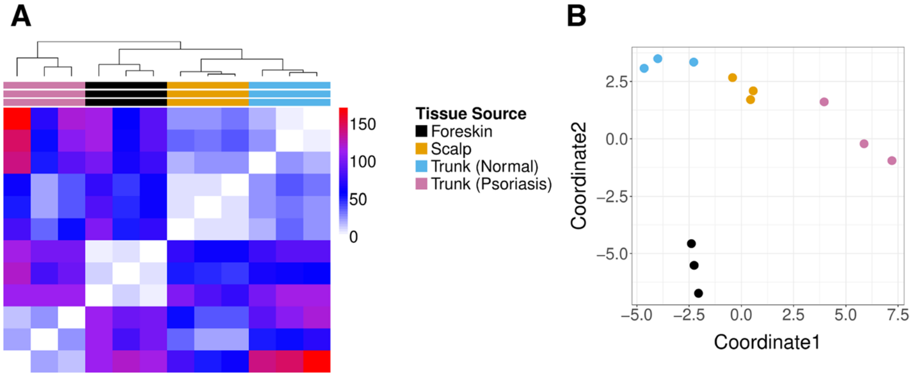

Fig. 3 GloScope representation of 12 skin rash patients collected in various locations and conditions in [20]. **A** A heatmap visualization of the estimate of the symmetrized KL divergence between the samples' GloScope representation. **B** A two-dimensional MDS representation of the divergences. The divergences were calculated using the GMM density estimation based on PCA estimation of the latent space in 10 dimensions

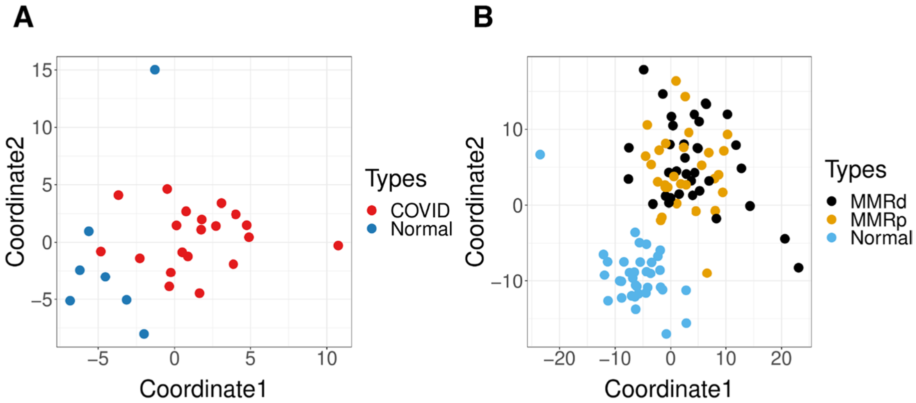

Fig. 4 Examples of MDS plot of the dissimilarities calculated from GloScope representation. **A** 27 samples of COVID lung atlas data that are either healthy samples of COVID patients from Melms et al. [21]. **B** 99 colon samples from mismatch repair-proficient (MMRp) tumors, mismatch repair-deficient (MMRd) tumors, and healthy samples from Pelka et al. [22]. The dissimilarity matrices were calculated using the GMM density estimate based on PCA estimates of the latent space in 10 dimensions

consider only two different phenotypic groups. Count matrices were generated from a pipeline modified from that presented in the R package `muscat` [25] (see the "Methods" section for details).

We focus on two basic biological scenarios that could causes phenotypic-based dissimilarity between scRNA-Seq samples which we would want the GloScope representation to accurately reflect: differential cell type composition and differential gene expression. By cell type composition, we refer to the proportion of various cell types found in a sample; for example, an inflammatory disease phenotype might result in a higher proportion of immune cells in the patient than in a healthy sample. Cell type gene expression differences (DE) refers to differences across samples in the marginal gene expression levels within cells of a certain type. For example, the IL2 gene has more expression within the T cells of inflammation tissue samples when compared to the its expression in T cells of healthy samples. Both types of differences are biologically plausible and can co-exist. We also note that in practice the distinction between these two can blur: many genes exhibiting sufficiently strong differential expression between phenotypes will result in the creation of a novel cell type for all practical purposes, thereby corresponding to differential cell type composition and vice versa.

In our simulations, we evaluate how well these two types of differences are detected by GloScope. We create datasets demonstrating either differentially expressed genes or differential cell type composition. We see that the average differences between samples in different different phenotype groups, as measured by our GloScope representation, appropriately increase in response to both increased differences in global cell composition (Fig. 5A) and increased differential gene expression (Fig. 5B). This indicates that our representation effectively reflects both types of changes. Similarly, when increased sample variability is added, both in global cell composition and gene expression, our GloScope representation correspondingly shows increased within-group variability (Fig. 5C and D).

We can use our GloScope representation to compare different choices of the design or analysis of the experiment, based on how well the two phenotypic groups separate in

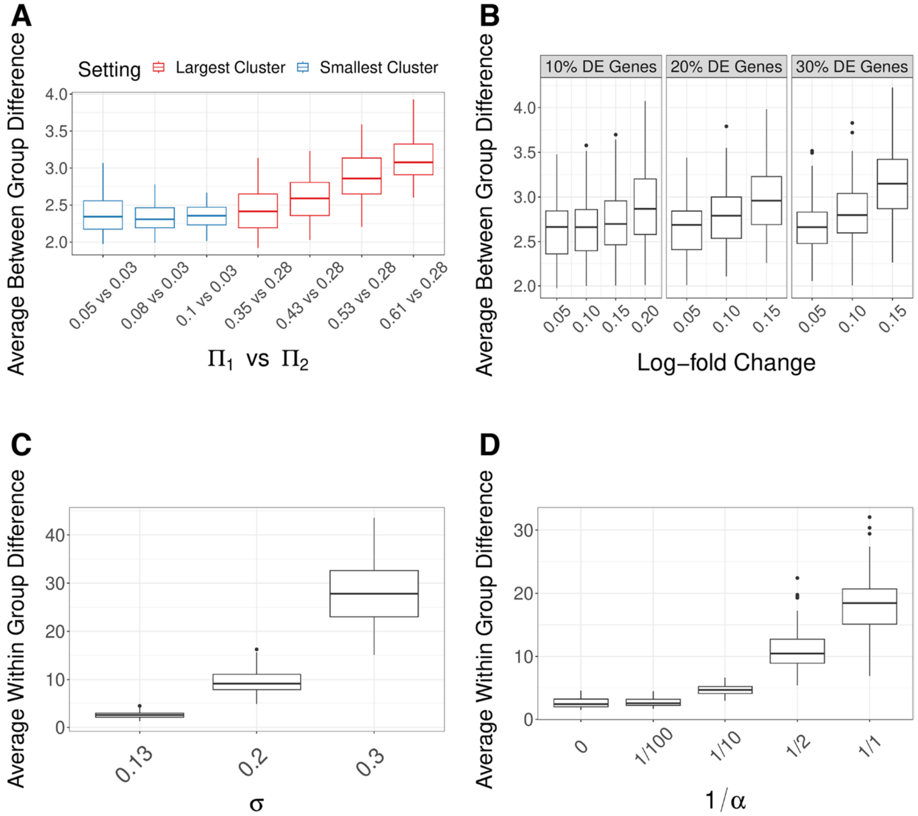

Fig. 5 GloScope captures simulated effects. **A** and **B** show how the average GloScope divergence between samples in different phenotype groups increases with **A** increased cell composition differences and **B** increased gene expression differences. The cell composition differences in **A** are color-coded as to whether the major changes were in the two groups' largest cluster or smallest cluster (the actual values of the proportion changes in the largest or smallest group, Π_1 vs Π_2, are labeled in the legends). Plots **C** and **D** shows how the average GloScope divergence between samples in the same phenotype group increases with **C** increased sample variability in gene expression differences and **D** increased cell composition differences. All box plots show these averages over 100 simulations. The dissimilarity matrices were calculated using the GMM-based GloScope representation based on PCA estimates of the latent space in 10 dimensions. For choices of kNN with scVI or PCA and GMM with scVI, see Additional file 1: Fig. S1-S4

the GloScope representation. To do so, we perform analysis of similarities (ANOSIM), a hypothesis test for differences between groups based on observed pairwise divergences on samples [23, 24]. ANOSIM takes as input divergences between samples and tests whether divergences are significantly larger between samples in different groups compared with those found within groups based on permutation testing (see the "Methods" section for more details). Evaluation of ANOSIM over many simulations gives the power of the test in different settings, resulting in a metric to compare choices in our analysis.

Using these power computations, we can see that changes in the sample variability and sample size are reflected as expected in these power calculations: increasing all of these sources of variability naturally reduces the power (Additional file 1: Fig. S5). These types of simulations, in conjunction with our GloScope representation, can be used to evaluate design choices at the sample-level, such as the number of samples needed to reach a desired power level. Unsurprisingly, differences in cell-composition in large clusters are more easily detected than similar differences in small clusters

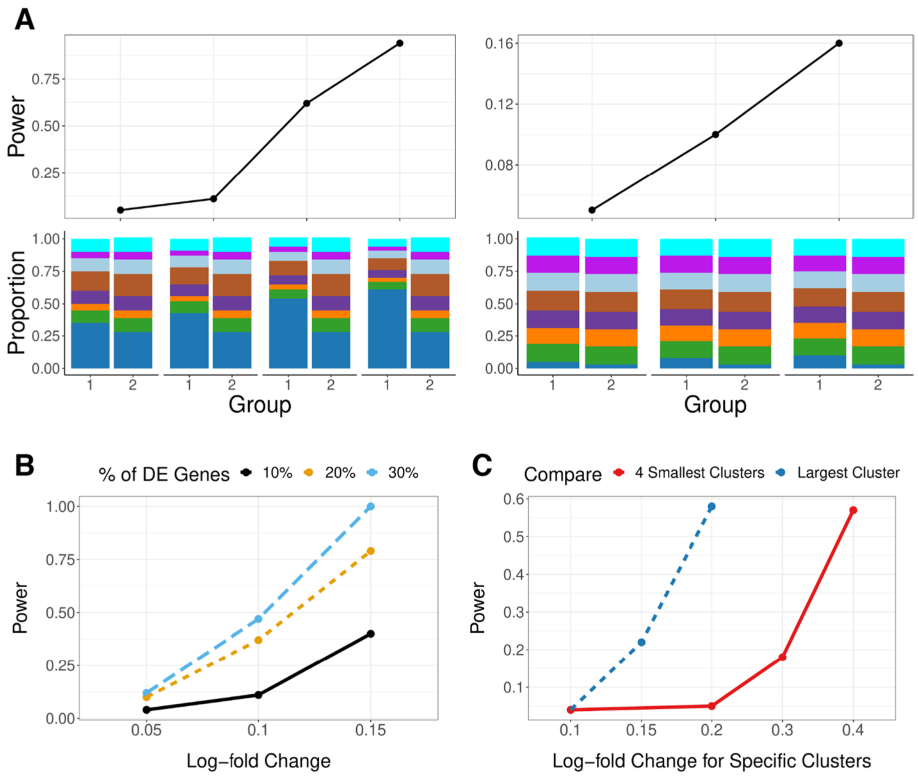

Fig. 6 ANOSIM power on simulated data (*y*-axis) under different conditions. **A** Changes in only the cell type composition (no DE genes), with major changes in the two groups' largest cluster (left) or smallest cluster (right). The cell type composition is visualized in the lower panels. **B** Increasing percentage of DE genes (ρ_{DE}) with average log fold change changing from 0.05, 0.1, and 0.15 (*x*-axis). **C** Changes of log fold changes concentrated in specific cell types/clusters (ω_k), quantified as relative to the baseline log fold change $\theta = 0.05$; the two lines correspond to whether the log fold changes were in the largest cluster (representing $\pi_k = 40\%$ proportion of cells) or for the 4 smallest cluster (representing $\pi_k = 30\%$ proportion of cells). Power calculations were done on relatively small groups to show the full range of changes ($n = 10$ samples in each group) with $m = 5000$ cells per sample; the sample level variability parameter σ is fixed at 0.13, and the sequencing depth $\lambda = 8.25$ (see the "Methods" section for details on these parameters). GloScope was calculated based on GMM density estimation with latent space representation via the first 10 dimensions of PCA

(Fig. 6A), and gene expression differences concentrated in small clusters are harder to detect than those found in large clusters (Fig. 6C).

We can also compare choices in the data analysis pipeline. For example, GloScope relies on a user-provided choice of latent variable representation of the single-cell data. We compare the choice of PCA versus scVI in a wide range of our simulation settings. The most striking difference is in detection of cell-composition differences, where scVI has much less power in detecting differences between the two phenotypic groups than PCA (Additional file 1: Fig. S6). The latent variable representations given by scVI demonstrates much greater variability between samples than the those of PCA (Additional file 1: Fig. S7), potentially resulting in less power to detect the shared phenotypic differences. On the other hand, scVI representations have more power than their PCA counterparts when the source of differences is due to log fold changes in genes (Additional File 1: Fig. S8), perhaps due to better accounting for sparse low-count data.

Finally, we can also consider choices made in implementing GloScope, in particular in the choice of estimation of the density of the latent variables Z in each sample. We consider two popular density estimation strategies: parametric Gaussian mixture models (GMMs) and non-parametric k-nearest neighbors (kNNs). We do not observe large differences in the power of these methods when varying the level of differential expression (Additional file 1: Fig. S8), but kNN is somewhat more powerful in the presence of cell type composition changes (Additional file 1: Fig. S6). Applying both methods on a wide range of datasets (Additional file 1: Fig. S9, S10) shows that, on average, the estimates of divergence from the two methods are generally monotone with moderate to strong correlations (Pearson coefficient ranging from 0.36 to 0.95); furthermore, the kNN estimates are systematically lower and appear to saturate when GMM estimates are large. While kNN-density estimation offers an asymptotically unbiased estimator of the symmetric KL divergence [26], it is known to exhibit downward finite sample bias due to underestimation of density in the tails of a distribution [27–29]. Due to these considerations, we relied on GMM estimates of density, though none of the results shown qualitatively change if kNN estimates are used instead.

GloScope representation for quality control

Finally, we demonstrate the use of GloScope for exploratory data analysis of relatively large sample cohorts and illustrate the utility of having a sample-level representation of the data for exploratory data analysis.

The first dataset is a study of COVID-19 [30] consisting of 143 samples of peripheral blood mononuclear cells (PBMC); samples in the study originated from patients that were either identified as infected with COVID-19 with varying levels of severity (COVID), negative for COVID-19 (Healthy), healthy volunteers with LPS stimulus as a substitute of an acute systemic inflammatory response (LPS), or having other disease phenotypes with similar respiratory symptoms as COVID-19 (non-COVID). Figure 7A shows these samples after applying MDS to the pairwise divergences calculated from the GloScope representation for the 143 samples of the study.

The visualization shows that both COVID patients and healthy donors are clearly separated from patients with other respiratory conditions (LPS and non-COVID). The other noticeable pattern is that the remaining patients do not show a strong separation between the COVID and Healthy phenotypes, but do appear to separate into at least two groups unrelated to these main phenotypes of interest-an observation that is further strengthened when considering the MDS representation of only the COVID patients and healthy donors (Fig. 7B). Exploration of the provided sample data from Stephenson et al. [30] shows that these groups correspond to different sequencing locations, indicating a strong batch effect due to sequencing site, with samples sequenced at the Cambridge site clearly separated from those at the New Castle (Ncl) and Sanger sites. When the individual cells are visualized (Additional file 1: Fig. S11), the distributional differences between these sequencing sites validate these differences, with cells from the Cambridge site lying in quite different spaces from cells of the same cell type from the other sequencing sites. Furthermore, Stephenson et al. [30] indicates that samples from these different sites underwent different sequencing

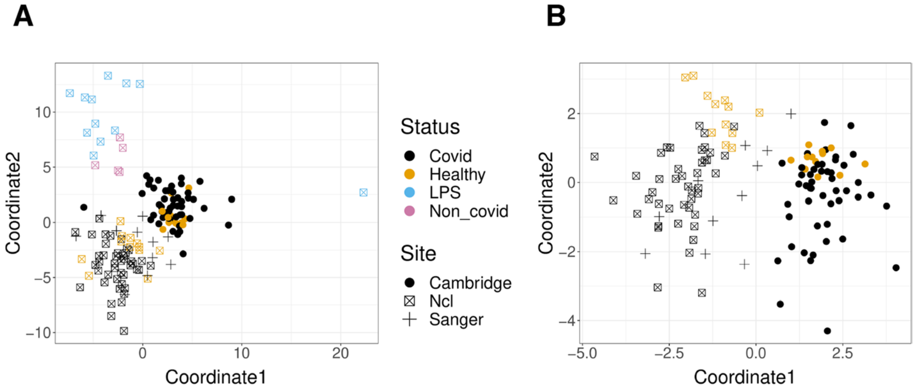

Fig. 7 GloScope representation applied to samples sequenced in Stephenson et al. [30]. Shown are the MDS representation in two dimensions of the KL divergence estimates calculated from the GloScope representation for **A** all 143 samples and **B** the subset of 126 samples that were either healthy or diagnosed with COVID-19 (MDS was rerun on the reduced subset of divergences between these 126 samples). Each point corresponds to a sample and is colored by the sample's phenotype; the plotting symbol of each sample indicates the site at which the sample was sequenced (see legend). Estimated GloScope divergences used the GMM estimate of density and latent variables were estimated with PCA in 10 dimensions. For the visualization of the full divergence matrix, see Additional file 1: Fig. S30

steps such as cell isolation and library preparations (and the original analysis in Stephenson et al. [30] corrected for potential batch effects by applying the batch correction method, Harmony [10]).

A similar analysis was applied to a systemic lupus erythematosus (SLE) dataset, with scRNA-Seq data of the PBMC cells of 261 patients; some patients had multiple samples resulting in total 336 samples [31]. Again, our GloScope representation clearly shows that there are distinct patterns among different batch sources, in addition to separation of normal samples from the other conditions (Fig. 8A). After application of Harmony to this data based on the batch, our GloScope representation shows much greater intermingling of the data from different batches (Fig. 8B). We can quantify the improvement by measuring the separation between samples within a batch compared to those in separate batches using measures such as the ANOSIM *R* statistic or silhouette width. We see the improvement due to batch correction but some loss of separation between biological conditions, which is a common trade-off when correcting for batch effects (Fig. 8C, D). This type of exploratory analyses of data is a common task in the analysis of scRNA-Seq data, and the GloScope representation provides a meaningful strategy for evaluating these types of processing choices. We further note that in addition to finding differences amongst the sequencing sites in the Lupus PBMC data, we observe further clustering of samples in Batch 4 (highlighted in Fig. 8A). These subgroups do not correspond with any patient covariates provided by the authors, but further exploration clearly show strong differences in the gene expression and cell density in certain cell types such as CD4 T cells, natural killer cells, and B cells (Additional file 1: Fig. S12, S13).

Similar concerns are frequently explored when integrating data from different studies. We applied GloScope on the dataset of Fabre et al. [32] which integrated six lung fibrosis scRNA-Seq studies, resulting in 144 samples after quality control. Application

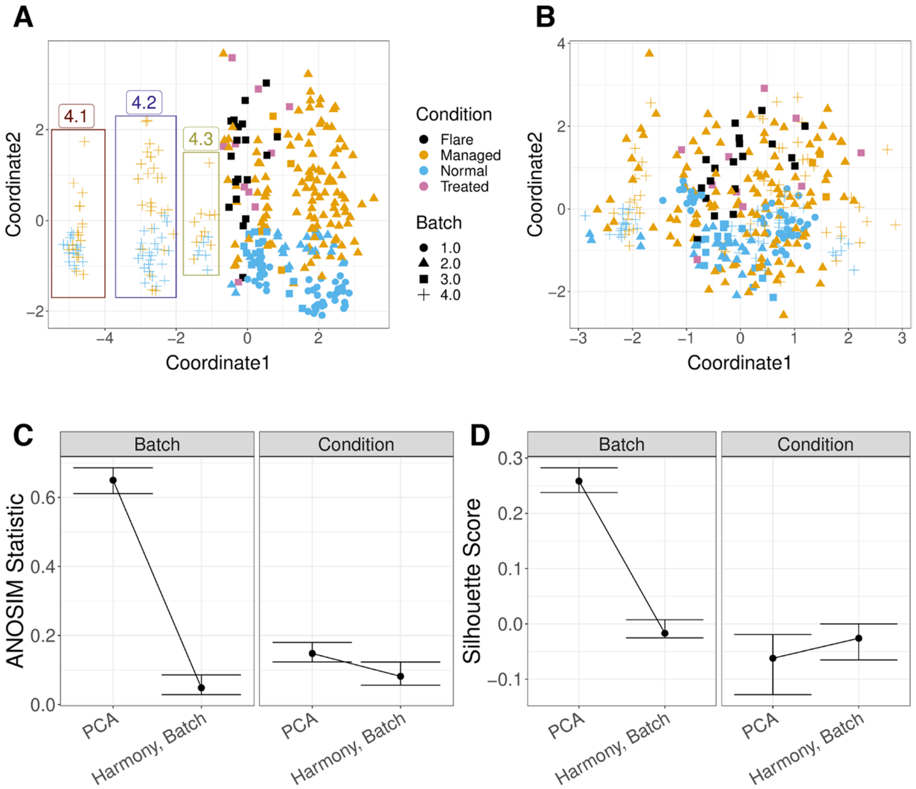

Fig. 8 GloScope representation applied to a Systemic lupus erythematosus (SLE) PBMC dataset of 336 samples [31]. Shown is the MDS of the GloScope representation applied to latent variables defined by **A** the first 10 PCA components of the original data and **B** the latent variables defined by Harmony after normalizing on processing cohort. **C** The ANOSIM statistics changing regarding capturing batch or condition signal, before and after applying batch correction (i.e., Harmony) with bootstrap confidence interval. **D** the silhouette widths changing regarding capturing batch or condition signal, before and after applying batch correction with bootstrap confidence interval

of GloScope (Fig. 9A) immediately shows one of the studies [33] as quite different from the other five; further investigation shows that the study of Adams et al. [33] has quite obvious differences in both gene expression and cell type composition than the other five studies. In particular, we observed quite obvious gene expression shifting in myeloid cells and natural killer cells in Adams et al. [33] (Additional file 1: Fig. S14), and samples collected from Adams have a higher portion of myeloid cells compared to samples from other studies (Additional file 1: Fig. S15). The remaining five studies show relatively smaller differences, but some separation is clearly visible. In addition to large batch effects, we observed a potential outlier (sample `092C_lung`), from the Adams et al. [33] study detected by the GloScope representation (Fig. 9A). Further evaluation of that outlier sample shows that `092C_lung` is missing most of the cell types except for B cells and lymphocytes (Additional file 1: Fig. S16). In contrast, a similar analysis of data from [32] which integrated studies of 50 human liver samples (after quality control) from 6 published scRNA-Seq studies of liver fibrosis shows far less distinction among the studies compared to the lung samples (Additional file 1: Fig. S17). Following application of Harmony for batch correction/integration, the

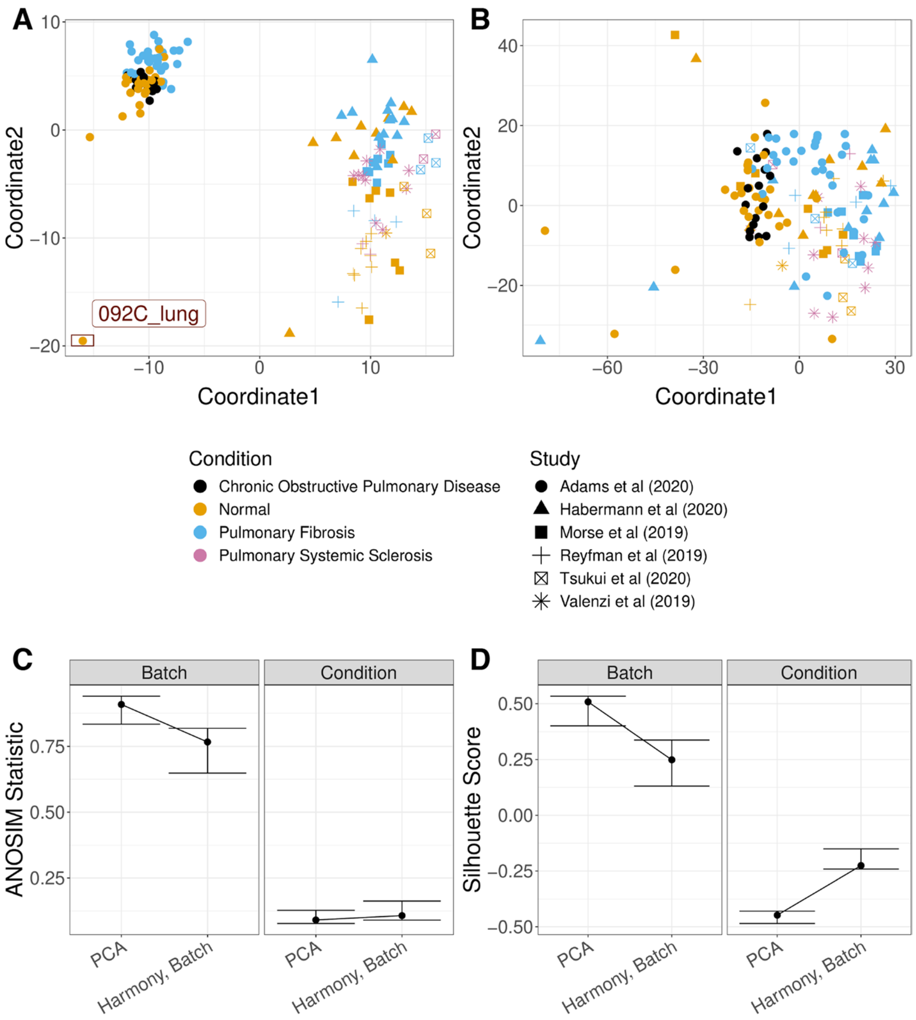

Fig. 9 GloScope representation applied to lung fibrosis samples collected in Fabre et al. [32]. Shown are the MDS representation in two dimensions of the KL divergence estimates calculated from the GloScope representation for **A** PCA embedding before batch correction and **B** PCA after applying Harmony batch correction. Each point corresponds to a sample and is colored by the sample's phenotype; the plotting symbol of each sample indicates the studies at which the sample was collected (see legend). Estimated GloScope divergences used the GMM estimate of density and latent variables were estimated with PCA in 10 dimensions. **C** and **D** visualize the ANOSIM R statistics and silhouette width, quantifying the changes of batch and biological signals before and after batch correction

GloScope shows effective integration of the lung studies and a corresponding clearer grouping of biological conditions (Fig. 9C, D).

Batch effects are common concerns with large sets of data, especially in human subject data where the samples are likely to be collected and possibly sequenced at different sites or integrated across multiple smaller studies. These examples immediately demonstrate the power of our GloScope representation for exploratory data analysis.

Comparison with other quality control tools

Existing tools for EDA and evaluation of potential quality concerns are generally focused on analysis at the level of the individual cell. Numerous metrics exist for evaluating the quality of individual cells and filtering poor cells, such as the the number of detected genes, the number of sequenced reads, or the percentage of mitochondrial DNA [34, 35]. Yet, many sources of possible artifacts are often due to variables that vary per sample or patient, such as the hospital of collection, the sequencing site, or the laboratory running the experiment. These effects have large-scale effects beyond individual cells and are best detected by comparisons of the cells as a group. However, there are limited options for detecting artifacts that vary by sample or individually poor samples.

In particular, analyses at the individual cell-level are less flexible for detecting these sample-level differences. There are metrics at the individual cell-level, such as iLISI [10] that can assess the presence of a batch effect *for known batch variables*. These are similar to our use of ANOSIM or silhouette width to quantify the separation between samples in batches, only these methods are applied to the individual cells. Such methods can highlight similar effects, such as showing an improvement in Harmony corrected data for the Stephenson et al. [30] data (Additional file 1: Fig. S18), but they are ineffective for discovering effects de novo, nor do they provide the ability to compare multiple effects, such as our visualizations of both batch and biological effects in the "GloScope representation for quality control" section.

A common exploratory visualization strategy for scRNA-Seq data consists of applying tools such as UMAP or tSNE to create a two-dimensional visualization of the individual cells. Individual cells can be color-coded by potential variables or plotted separately per sample for exploration of possible *known* artifacts, as we provided for the [30] data in Additional file 1: Fig. S11. UMAP visualizations can be helpful in retrospect for understanding the nature of the problem but are not particularly effective in discovering such effects de novo given the difficulty in visualizing sample effects for large numbers of cells. The example of the Perez et al. [31] data is illustrative, where our GloScope representation allowed us to immediately determine unexplained groupings of samples within batch 4; we were able to follow this discovery with further investigation at the individual cell-level using UMAPs to discover that there were shifts in gene expression and cell density among these subgroups GloScope identified within batch 4. These differences are not detectable in plotting all cells, and only after identifying the subgroups of patients can a UMAP help in further investigation. Furthermore, differences due to shifts in cell distributions can be tricky to see in UMAP visualizations of individual cells, due to the overplotting of cells. Even after identifying the different subgroups in batch 4 with GloScope (Fig. 8), the differences seen clearly in the GloScope representation were subtle to detect using standard UMAP visualization (Additional file 1: Fig. S13, S19, S20). This exploratory analysis of the [31] data shows the complementary nature of GloScope with other visualization tools. Similarly, outlying individual patients, as we detected in the lung samples of Fabre et al. [32] (the "GloScope representation for quality control" section), would require plotting and comparing of UMAPs of each individual sample which is simply not feasible for large cohorts.

There are some limited alternatives to GloScope available for the comparison at the sample-level, and they take different strategies for summarizing the data from a single patient which we next consider: cell-composition and pseudobulk.

Comparison with cell-composition analysis

Reducing each sample to their cell type composition has been proposed for comparing samples. A simple version of this strategy is to visualize the proportions per sample in a bar plot. Like UMAPs of individual cells, such bar plots can be useful tools for greater investigation of differences found by GloScope but do not scale for easy comparisons of large number of samples and do not aid in discovering possible differences, such as the potential subgroups of batch identified by GloScope (Additional file 1: Fig. S12, S21).

The cell type proportions can also be analyzed more quantitatively-for example, the GloScope methodology can also be used for cluster proportions, which we call GloProp (see the "Overview of the GloScope representation" section). Concurrently, Joodaki et al. [15] has proposed a similar metric strategy for comparing cell type proportions named PILOT, using Wasserstein distance rather than symmetric KL divergence. These approaches require determination of cell type proportions and can only be run after clustering the individual cells. Such clustering is typically done after EDA and correction of possible batch effects, making it irrelevant for EDA. But in principle, clustering could be done earlier in the pipeline for the sole purpose of using PILOT for EDA (the discovered clusters would not be biologically meaningful until the data has been appropriately pre-processed). We do this clustering on the uncorrected data and compare PILOT and GloProp to GloScope. We see that PILOT performs much worse than GloScope or GloProp in detecting separations between the batches in all of the datasets (Additional file 1: Fig. S22). Our method for cluster proportions, GloProp, can perform similarly to that of the full GloScope representation, but the performance of both PILOT and GloProp is very sensitive to the clustering. When we vary parameters of the clustering algorithm or consider different random starts, the performance can vary dramatically, unlike GloScope (Additional file 1: Fig. S23).

Comparison with pseudo-bulk analysis

Another potential strategy for sample-level exploratory analysis is using a pseudo-bulk created from the scRNA-Seq data. This is a strategy of aggregating over each sample's cells to obtain a single observation per sample [1]; the most common is to simply sum the counts. Then, standard methods from bulk mRNA-Seq, such as PCA, can be applied at the sample level. The authors of [36] propose a strategy, MOFA, for finding lower-dimensional latent embeddings per sample based on combining pseudo-bulk measures per cell type, to better reflect cell type variability.

We create such a PCA visualization of the pseudo-bulk of several of the datasets mentioned above (Additional file 1: Fig. S24, S25). For the COVID-19 PMBC samples, for example, the pseudobulk analysis does not clearly separate out the LPS and non-COVID samples, nor is the strong batch effect due to sequencing site as clearly identified. Similarly, for the Lupus PBMC data, the pseudobulk representation does not identify the strong batch effects seen in our GloScope representation. This is borne out by the quantification of the average silhouette width or R statistic (Additional file 1: Fig. S26, S27). On the other hand, these quantification statistics show MOFA to have similar performance in detecting batches as GloScope; however, on closer examination of the visualization of the results of MOFA, we see less clear separation of the effects seen by GloScope. For example, MOFA did not show clear of a separation of all the non-COVID and LPS samples from other samples and the separation of the groupings found de novo by GloScope are attenuated and difficult to find (Additional file 1: Fig. S25).

There are other limitations to either of these pseudo-bulk strategies. The pseudo-bulk strategy, including MOFA, is based on summarizing for each gene the expression level of all the cells in a sample, usually the sum of the raw counts. However, many public datasets provide other normalized versions of the data (e.g., residuals); similarly, many batch-correction methods, like Harmony [10], provide a batch-corrected latent variable representation. None of these are obvious candidates for either of these pseudo-bulk approaches. Our GloScope representation requires as input only a latent variable representation per cell and thus is flexible to accommodate all of these types of input. This is important, for example, in evaluating the effect of batch correction methods. With GloScope, we can evaluate the data before and after batch correction with the Harmony algorithm (Fig. 8B, C, D), allowing us to confirm that the Harmony algorithm has removed much of the differences between batches. Moreover, the pseudo-bulk methods can often need normalization across samples in addition to normalization that may be done to individual cells so that they do not reflect simply the number of cells, similar to bulk RNA, which adds another layer of complexity since there are many strategies for such a normalization. GloScope summarizes the individual cells as a density, which is a measurement unaffected by the number of cells per sample.

Discussion

In this work, we demonstrated the use of GloScope for exploratory analysis, and in particular how the GloScope divergences can be used to create two-dimensional scatter plots of samples, similar to that of PCA plots of bulk mRNA-Seq data. We demonstrated the ability of the GloScope representation to detect important artifacts in the data, as well as assess batch-correction methodologies.

We also compared GloScope to the limited available strategies for summarizing the data from a single patient: cell type composition and pseudobulk. We show that these methods are not as sensitive in as diverse of settings. In particular, these approaches each focus on one aspect of the sample data (cell type proportions or gene expression) and are not sensitive to changes found in the other. GloScope uses the entire distribution of the data, thus effectively combining both cell type proportions and gene expression in a single summary. Furthermore, GloScope is far more flexible for incorporation at different stages of the analysis, whether working with raw counts or normalized data.

While we focus on the utility of the GloScope representation to visualize scRNA-Seq data at the sample level, the representation can be used more broadly with other statistical learning tools. For example, we can use the GloScope divergences between samples as input to a prediction algorithm in order to predict a phenotype. With the COVID-19 data, we apply the SVM algorithm to the GloScope divergences which results in a prediction algorithm that was able to separate the normal from the COVID samples with a 5-fold cross-validated prediction accuracy of around 0.88. This simple example serves as an illustration of the power of a global representation of the entire scRNA-Seq profile.

Finally, we note that GloScope can easily be incorporated into existing scRNA-Seq pipelines at multiple stages of analysis to assess the progress. Latent variable representation, via PCA or scVI is a standard initial step in an analysis, while many popular batch correction methods provide low-dimensional representations of corrected data. Even multi-modal integrations usually result in a low-dimensional latent space estimation. The output of all of these tasks can be provided to GloScope for evaluation of sample-level similarities, resulting in a flexible tool for exploratory analysis of the results.

Conclusions

We have presented the statistical framework GloScope that provides a global summary of each scRNA-Seq sample based on the distribution of their gene expression values across their cells. This representation allows for comparisons between the entire single-cell profile of a sample. Formal calculations of the dissimilarities between samples can be used as input to other statistical and machine learning algorithms to allow a sample-level analysis. Our representation is able to differentiate among samples from varied phenotype groups, such as COVID lung tissue samples and healthy lung tissue samples, and is shown to be a powerful tool to detect potential batch effects.

Methods

The GloScope representation

Our GloScope representation consists of representing each sample as a distribution along with a corresponding divergence or distance; we then estimate the distance or divergence between each pair of samples based on their scRNA-Seq data. This representation allows for application of kernel methods common in machine learning, which depend on the calculation of the distance between each pair of samples i, j for downstream statistical analysis.

To do this, we posit an underlying true distribution of cells F_i for each sample i, which is a continuous probability distribution on R^g, where g is the number of genes. We define a measure of divergence d on the space of probability distributions in R^g. In this work, we fix d as the symmetrized Kullback-Leibler divergence,

$$d(F_i, F_j) = KL(F_i||F_j) + KL(F_j||F_i), \tag{1}$$

which has been used in a similar manner in the case of facial recognition (e.g., [6, 37, 38]).

We do not observe the F_i directly and must instead estimate that distribution from observed data. The observations from a sample i consists of m_i sequenced cells; in order to estimate F_i, we will make the simplifying assumption that the sequenced cells are independent and identically distributed (i.i.d) draws from the sample's full population of cells, F_i. Even with this assumption, density estimation is complicated in this setting. For scRNA-Seq datasets, g is often in the range of 2000–8000 (the number of detectable genes given the sequencing depth). The number of cells per sample, m_i, can vary by experiment, and often m_i ranges lies in the range of 500 to 10,000 cells per sample. The data from each cell is high dimensional and sparse, a distributional structure known to be impactful in the analysis of scRNA-Seq data [39–43].

Defining a latent space

Even with several thousand cells per sample, it is infeasible to estimate the density in such a high-dimensional space without the assumption of an underlying lower-dimensional latent space. Therefore, for each sample i and cell c, we model a latent variable $Z_{ic} \in R^d$ and a transformation $\sigma : R^d \rightarrow R^g$. Then, our observed vector x_{ic} of gene expression counts from a cell is assumed drawn from an appropriate generative model for RNA counts with mean parameter $\sigma(Z)$, i.e., $E(x_{ic}) = \sigma(Z_{ic})$.

For a sample i, we assume that the Z_{ic} for each cell c is distributed as the latent random variable H_i. Instead of estimating F_i in R^g, GloScope instead estimates H_i in the lower-dimensional space R^d. In the "Overview of the GloScope representation" section, we denote the estimated distribution as $\hat{F}_i$ for conceptual simplicity, but a more precise notation would be $\hat{H}_i$ to clearly emphasize that we are estimating the distribution on a lower-dimensional space.

Furthermore, we note that our above heuristic states that we observe counts x_{ic} in cell c drawn from a single distribution F_i; this ignores cell-specific effects that could result in slightly different distributions for different cells, such as different sequencing depth that varies for each cell c. The latent variables Z_{ic}, however, are independent of the cell-specific effects due to the technology, which makes estimation of a single distribution, H_i, shared by all cells a coherent mathematical framework.

Estimation

The GloScope representation estimates H_i for each sample with a two-stage strategy: (1) estimation of the latent variables $Z_{ic} \in R^d$ for each cell c in sample i and (2) estimation of the density of $\hat{H}_i$ from Z_{ic} and corresponding distances $d(\hat{H}_i, \hat{H}_j)$ between samples. An advantage of estimating the latent variable samples before the density is that we can apply one of many existing dimensionality reduction techniques that account for sparse count data, such as `ZINBWave` [40] or `scVI` [5], or techniques that simultaneously remove batch effects and estimate a latent space, such as `Harmony` [10] or `fastMNN` [44].

The GloScope representation assumes that the user chooses an appropriate method for the first stage estimation of Z_{ic} (i.e., a dimensionality reduction method) and then offers two approaches for the second stage (estimation of the distances between the H_i).

The first approach fits a Gaussian mixture model to the data Z_{ic} to estimate h_i, the density associated with the distribution H_i. The mixture models are fit independently for each patient, and thus the parameters of these models, such as the number of components, can vary between patients. After a distribution $\hat{H}_i$ is fit for each patient's data, our estimate of $d(F_i, F_j)$ is calculated between patients with $d(\hat{H}_i, \hat{H}_j)$. Single-cell methods utilizing dimensionality reduction, described above, often include a regularizing assumption that the latent variables $Z \sim N(0, \Sigma)$. This Gaussian regularization in the model and the fact that many datasets are mixtures of cell type populations motivate our use of Gaussian mixture models (GMMs). We use the R package `mclust` [45] to implement the GMM estimation. As there is no closed form expression for the KL divergence between GMM distributions, we use Monte Carlo integration to approximate the

KL divergence between two GMM densities; this is based on $R = 10,000$ samples drawn from the estimated GMM distributions, again using the `mclust` package. Specifically, for R draws of x from $\hat{H}_i$, we have

$$KL(\hat{H}_i||\hat{H}_j) \approx \frac{1}{R}\sum_{u=1}^{R}\log\frac{\hat{h}_i(x_u)}{\hat{h}_j(x_u)} \tag{2}$$

We also provide a second approach that estimates $d(H_i, H_j)$ directly using a k-nearest neighbor approach without explicitly estimating the density h_i [46, 47]. Denote by $r_j(x_{i,u})$ the distance from the uth cell in sample i to its kth nearest neighbor in sample j. Then, the KL divergence can be estimated directly as

$$\widehat{KL}(H_i||H_j) = \frac{d}{m_i}\sum_{u=1}^{m_i}\log\frac{r_j(x_{i,u})}{r_i(x_{i,u})} + \log\frac{m_j}{m_i - 1} \tag{3}$$

where d is the dimension of the latent space [46, 47]. We implement this strategy using the `FNN` package to estimate the symmetrized KL divergence between sample i and sample j [48].

We would note that these density estimation methods make the assumption that cells from a sample i are drawn i.i.d. from the same distribution F_i (where F_i can be a mixture distribution to account for multiple celltypes). This is a standard assumption for density estimation, and indeed for most algorithms applied to single-cell data, such as clustering and PCA. Technical artifacts in the data, such as missing cells or bias in capturing certain cells or genes, could violate these assumptions; this would affect the density estimate and therefore result in a poor estimated density relative to the true density. However, the GloScope method is a comparative method, in the sense that its purpose is to quantify the differences between different samples. Therefore, even if the density estimates were biased due to technical artifacts, the comparisons between the samples would still be meaningful by identifying samples containing such technical artifacts relative to other samples.

Simulating scRNA-Seq data

To simulate population-level scRNA-Seq data for benchmarking our methodology, we follow the model introduced by [1] and implemented in the `muscat` Bioconductor package [25]. All adjustments to the `muscat` package mentioned here were based on altering the code available in the `muscat` package, Version 1.13.0 [25]. The adapted code for the simulation is available in the GitHub repository accompanying this paper [49].

Simulation model

The model of [1] is a model for simulating count data for each gene and unlike our GloScope representation does not assume any latent variable representation in generating the data. The model assumes a simple two-group setting in which each sample i may come from one of two groups, denoted by the variable $T(i) \in \{1, 2\}$. The m_i cells from sample i come from K different cell types with the proportion of cells from cell type k given by $\pi_{i,k}$, where $\sum_k \pi_{i,k} = 1$. Thus, the gene expression vector $x \in R^g$ of a cell c from sample i is assumed to follow a negative binomial mixture model:

$$F_{i,c}(x) = \sum_k \pi_k P_{NB}(\mu_{i,c,k}, \phi)(x) \tag{4}$$

where $P_{NB}(\mu_{i,c,k}, \phi)$ is a CDF on R^g representing a product distribution of independent negative binomials, i.e., each gene's expression value is independent and follows a negative binomial distribution with mean given by the j the element of the vector $\mu_{i,c,k} \in R^g$ and dispersion parameter $\phi \in R$.

The vector of gene means for cell c in sample i is parameterized as

$$\mu_{i,c} = \lambda_{i,c} e^{\beta_{i,k}} \cdot \theta_{k,j}, \tag{5}$$

where $\lambda_{i,c} \in R$ is the library size (total number of counts); $\beta_{i,k} \in R^g$ is the relative abundance of g genes in cells belonging to sample i and cell type k; $\theta_{k,j} \in R^g$ is the fold change for genes in cluster k if the sample belongs to group $j \in \{1, 2\}$. Notice, as mentioned above, that because of different sequencing depths per cell, each cell within sample i has a different mean $\mu_{i,c,k}$ governed by the sequencing-depth parameter $\lambda_{i,c}$, hence our notation $F_{i,c}$.

We make adjustments to the above model to more fully explore sample variability. To explore the effect of library size variation at both the cell and sample level, we introduce the decomposition $\lambda_{i,c} = \bar{\lambda} + \lambda_i + \delta_c$, where $\bar{\lambda}$ is the overall (average) library size, and λ_i and δ_c are variations from that due to sample or cell level differences, constrained so that $\lambda_{i,c} > 0$. We also adjusted the model to allow sample-specific proportions vectors $\pi_{i,k}$, with $\sum_k \pi_{i,k} = 1$. We define proportions per treatment group, $\Pi_j \in R^K$, for treatments $j = 1, 2$, such that $\sum_k \Pi_{j,k} = 1$ and randomly generate probability vectors π_i for sample i from a Dirichlet distribution according to its treatment group, $\pi_i \sim Dirichlet(\Pi_{T(i)} * \alpha)$, with sample level variation parameter α.

Selection of parameters

The `muscat` package also provides methods for creating these many parameters based on a few input parameters by the user and estimating the other parameters based on reference data provided by the user. We followed their strategy, with the following additions.

We chose the group fold change difference per cell type, $\theta_{k,j}$ following the schema of `muscat`, which allows for various types and size of changes between the different groups. Briefly, the simulation of $\theta_{k,j}$ is controlled by parameters (1) $\Omega \in R$, which is a user-defined average log2 fold change across all DE genes, (2) $\omega_k \in R^k$, which varies the magnitude of gene expression difference for cluster k, and (3) a proportion vector ρ which is the proportion of genes that follow six different gene expression patterns (see [1]); for simplicity, we allowed only the two most typical gene expression patterns, which are EE (equally expressed) and DE (differentially expressed) genes for our simulations, resulting in ρ effectively being a single scalar, the proportion of genes that are differentially expressed.

The selection of m_i, the number of cells per sample i, also followed the strategy of `muscat`, where the user provides a value $\bar{m}$, representing the average number of cells per sample across all samples, and the value of each individual m_i for each sample is assigned via a multinomial with equal probability and total number of cells across all samples equal to $n * \bar{m}$.

The parameters ϕ, and initial values of $\lambda_{i,c}$ and $\beta_{i,k}$ were obtained by estimating these parameters from the reference data, following the `muscat` package: after performing quality control, we used the filtered gene matrix and the `edgeR` package to estimate the parameters from the reference data.

Using our modified parameterization described above, $\bar{\lambda}$ was then chosen as the average of the $\lambda_{i,c}$ estimated from the reference samples. Sample-level sequencing depth variability λ_i were simulated as $\lambda_i \sim Unif(-\tau_\lambda, \tau_\lambda)$. Per-cell variability, δ_c, was simulated as $\delta_c \sim Unif(-\tau_\delta, \tau_\delta)$.

Finally, the selection of $\beta_{i,k}$ used in our simulation diverged from `muscat` package strategy. The `muscat` estimates of $\beta_{i,k}$ created overly large differences between the treatment groups and samples (Additional file 1: Fig. S28); furthermore, their strategy recycles the same set of parameters $\beta_{i,k}$ if the simulated sample sizes are larger than provided reference sample sizes (i.e., the same value of $\beta_{i,k}$ would be given to multiple simulated samples), resulting in unintended batches of samples. Instead, we estimated $\hat{\beta}_{i,k}$ from the reference data using the `muscat` strategy and chose a single sample i^* whose initial estimates $\hat{\beta}_{i,k}$ were representative. We then set $\hat{\beta}_k = \hat{\beta}_{i^*,k}$ and created individual $\beta_{i,k}$ with variation per sample by adding noise to $\hat{\beta}_k$, $\beta_{i,k} = \hat{\beta}_k/2 + \xi_{i,k}$, where $\xi_{i,k} \sim N(0, \sigma_\xi)$. σ_ξ controled the degree of sample-level variation.

Additional file 1: Fig. S29 shows the effect of changing different parameters (σ and log fold change), visualized using UMAP on an illustrative example.

Simulation settings

In following the above strategy of selecting parameters, we randomly chose 5 COVID samples from the COVID-19 PBMC dataset, [30]. After estimating ϕ and $\hat{\beta}_k$ as described above from the reference samples, the values were fixed for all simulations. The value $\bar{m}$ was chosen as 5000, which is similar to the average cell per samples in several datasets (e.g., [21, 22, 30]). The default value for α to control the sample level cluster proportion variability was set to be 100, except where explicitly noted, which keeps the variation in cluster proportions to be relatively small among samples (see Fig. 5D).

Once these parameters were fixed, the following user-defined parameters were set differently for different simulation settings: n (the number of samples in a single group), the vector group proportions Π_j ($j = 1, 2$), average library size $\bar{\lambda}$, and the DE parameters Ω, ω, and ρ. With these global parameters chosen for a simulation setting, the remaining sample-specific parameters are generated anew in each simulation:

1. For each cell type k, n values of $\beta_{i,k}$ as described above based on $\hat{\beta}_k$,
2. For each cell type k, a single vector $\theta_{k,j} \in R^G$ for the population log fold change between groups, based on the parameters Ω, ω, and ρ,
3. For each sample i, a single value λ_i and m_i values of δ_c, one for each of the m_i cells from each sample. This results in m_i values of $\lambda_{i,c} = \bar{\lambda} + \lambda_i + \delta_c$ for each sample (note that some simulations set λ_i and/or δ_c to 0 for all c and i).

Combining these parameters result in the $\mu_{i,c,k}$ needed for each sample in a single simulation, and then the cell counts for each sample i are simulated from $F_{i,c}$.

Additional files 2–7 provide the different parameter settings that were run and their resulting power and average ANOSIM values corresponding to the figures shown here.

Numerical metrics for evaluating simulations

In order to quantify how well our representation was able to differentiate sample groups in different settings, we implemented a simple hypothesis test for comparing the two groups based on our estimated distances from our GloScope representation. We relied on the analysis of similarities (ANOSIM) test, which is a non-parametric test based on a metric of dissimilarity, to evaluate whether the between group distance is greater than the within group distance. We used the function *anosim* in the R package `vegan` to perform the test [23, 24]. The test statistic is calculated as:

$$R = \frac{r_B - r_W}{N/2(N/2-1)/4} \quad (6)$$

where r_B is the mean of rank similarities of pairs of samples from different groups, r_W is the mean of rank similarity of pairs within the same groups, and N is the total number of samples. The test statistics ranges from -1 to 1. Strong positive test statistics means greater between group distances than the within groups, strong negative test statistics means the opposite and may represent wrong group assignments, and test statistics near zero indicate no differences. Finally, p-values are calculated based on a null permutation distribution: the distribution of R recalculated after randomly shuffling the samples' group assignment. The p values are calculated as the proportion of times that the permuted-derived statistics are larger than the original test statistic.

We used the results of ANOSIM to calculate the power in different simulation settings, creating a quantitative metric for evaluating the sensitivity of the GloScope representation in different scenarios. For a choice of input parameters, we repeated the simulation 100 times. For each simulation, we calculated the pairwise distances between all $2n$ samples, then used ANOSIM p-values to determined whether we would reject the null hypothesis. Finally, we calculated the power as the proportion of the 100 simulations' test statistics that have p values smaller than $\alpha = 0.05$.

Data processing procedures

This section details the steps undertaken to estimate GloScope representations of samples from publicly available scRNA-Seq data. These steps broadly consisted of ensuring the data we used had quality control matching the corresponding paper, estimating the cells' latent embeddings, and applying the GloScope methodology. For most datasets, we performed the first two steps with data structures and functions from the R package `Seurat`. For the larger Lupus PBMC and mouse brain datasets, we instead utilized the `SingleCellExperiment` data structure and applied functions from other packages. Code for running these analyses, as well as text files containing data sources and specific processing choices, is available in the following GitHub repository: https://github.com/epurdom/GloScope_analysis [49].

Quality control verification

The UMI count data and cell annotations from each sample-level scRNA-Seq study were downloaded from its publicly accessible source (indicated in the code). We checked whether data provided already had the quality control steps described in its respective paper. These steps can include removing cells with extreme expression values and filtering certain gene sets, such as mitochondrial genes. Only the data provided from Ledergor et al. [50] did not appear to have the stated steps of the manuscript already applied, and we reproduced the cell-wise quality control procedure described in that paper's Methods section. We also removed genes expressed in less than 10 cells (except for the data from Stephenson et al. [30] which provided only PCA embeddings that we used directly, and Fabre et al. [32], where the preprocessed and normalized data is provided).

Latent space estimation

In this paper, we present results based on using 10-dimensional latent embeddings, calculated with either scVI or PCA. To calculate scVI embeddings, we used the entire UMI count matrix as input after the aforementioned verification steps. To calculate PCA embeddings, we used a subset of only the 2000 most highly variable genes. To select which genes to include, we first log-normalized the counts within each cell; this was implemented with *logNormCounts* from `Seurat` or *logNormCounts* from `scuttle` for `SingleCellExperiment` objects. Then, we fit a LOESS curve to predict each gene's log-scale variance from its log-scale mean; that regression was implemented with the *vst* method of *FindVariableFeatures* in `Seurat` and with *modelGeneVar* from the `scran` package for `SingleCellExperiment` objects. The *FindVariableFeatures* function of `Seurat` also selects the 2000 highly variable genes based on large residuals in the LOESS regression. That exact selection rule is not available for `SingleCellExperiment` objects, so we instead applied a similar procedure implemented by *getTopHVGs* from `scran`. This alternative only differs in a truncation step and is commonly used in other scRNA-Seq analyses [51]. Each of the 2000 selected genes was centered and scaled to zero mean and unit variance before running PCA. In `Seurat` objects, this was done via a two-step procedure with calls to *ScaleData* and *RunPCA*. However, for `SingleCellExperiment` objects, we standardized each gene and ran PCA with a single call to *runPCA* from `scater`.

Application of GloScope

After obtaining each cell's latent representation via PCA or scVI, we fit sample-level densities with GMM or kNN and the KL divergence between samples was estimated. This produces the GloScope representations, and these steps are implemented by *gloscope* function in our `GloScope` R package, which accompanies this paper and is available in the accompanying Bioconductor package `GloScope` (https://www.bioconductor.org/packages/release/bioc/html/GloScope.html, version 1.2.0 [52].

To run GloScope, we first had to determined which cells constituted a single sample in each dataset, based on the provided metadata. In some studies, each patient only provided one sample, and in others, a single patient provided multiple samples, for instance from affected and healthy regions. Considering this, we ran GloScope with tissue samples as the unit of analysis. The sole exception to this choice is the PBMC data from Lupus patients in Perez et al. [31]; this study processed some tissue samples in multiple processing cohorts, and we therefore used the cross of sample and cohort as our unit of analysis.

Before applying GloScope, we confirmed that the tissue sample identifier associated with each cell matches the reported study design. The original melanoma data from Jerby-Arnon et al. [53] had duplicate encodings, which we standardized, and one sample with a mislabeled phenotype. For multiple myeloma cells from Ledergor et al. [50], we parsed concatenated strings into patient, observation period, and phenotype indicators. Two colorectal tumor samples from Pelka et al. [22] were sequenced with two technologies, and we only considered the replicates using the newer technology.

For datasets other than Fabre et al. [32], we also chose to remove samples with less than 50 cells. This excluded 2 samples from Ledergor et al. [50] (AB3178 and AB3195) and 1 from Pelka et al. [22] (C119N). For data in Fabre et al. [32], we observed extremely small cell number per sample (minimum 2 cells) and chose to remove samples with less than 500 cells, resulting in removing 4 samples in Lung study (137CO_lung, 244C_lung, 8CO_lung, and 084C_lung) and 15 samples in Liver study (P13_healthy_liver, P4_Tumor_liver, P8_healthy_liver, P10_Tumor_liver, P11_Tumor_liver, P11_healthy_liver, P5_healthy_liver, P7_healthy_liver, HN_healthy_liver, P12_healthy_liver,P1_healthy_liver,P3_healthy_liver and P9_healthy_liver). We noted that one sample from Ledergor et al. [50] (AB3461) had extreme divergences with other samples. We removed all cells from this sample for the results presented in this paper.

Computational time

In Additional file 1: Table S1, we provide the timing for how long it took to run to GloScope compared with other steps in the pipeline. We would note that there are two major computational tasks: the estimation of the density per sample and the estimation of the divergence between pairs of samples. However, because the calculation of the densities is more time intensive than the estimation of the pairwise divergences, the total time to run GloScope is roughly linear in the number of samples, not quadratic for the sample sizes reasonably encountered in scRNA-Seq studies.

Pseudobulk comparison experiments

For datasets where raw count data are available, we performed pseudobulk analysis by summing each sample's cell entries across each genes using *rowSums*, a base R function. After obtaining the pseudobulk data, we processed it using functions from the Seurat package. We log-normalized the data using *NormalizeData* and selected the top 2000 highly variable feature using the *FindVariableFeatures* function with default arguments. Counts from the selected genes were scaled using *ScaleData*, and a PCA embedding was obtained with *RunPCA*.

To perform MOFA analysis, we followed the vignettes provided in the R package MOFAcellulaR [36]. For filtering cells for downstream analysis, we set *ncells* to be 1. As we used the pre-processed data, we did not filter on sample and set *nsamples* threshold to be 0. For the final output, we set *num_factors* to be 5.

Comparison to cluster composition

In the "Comparison with other quality control tools" section, we compare our methods to those that summarize the samples using cluster composition. We call it "cluster composition" rather than "celltype composition" to emphasize that this is clustering done before standard batch corrections, and thus the results are not biologically meaningful. To do this comparison, we calculate cluster assignments for the individual cells based on a variety of different clustering choices: different algorithms and various choices of parameter settings and random starts for those algorithms. For each clustering assignment generated, we calculate the proportion of cells that are classified into each cluster and use those cluster proportions as input to PILOT [15] and GloProp, our implementation of GloScope for proportions.

To generate cluster assignments for individual cells, de novo clustering is preformed on lower-dimensional cell representations. In our implementation, we focus on PCA to match our input into GloScope. We use two of the most utilized clustering algorithms in scRNA-Seq for this task: the Louvain algorithm [54] and its refinement in the Leiden algorithm [55]. Leiden algorithm is recommended by recent review studies [56, 57] and has become the default clustering algorithm for popular software tools such as `scanpy` [58] and `Monocle3` [59]; furthermore, the Louvain algorithm codebase is no longer maintained. We present only results from Leiden, but the results from Louvain were qualitatively similar.

Both algorithms expect a k-nearest neighbor graph between cells as the input. In these experiments, we present results utilizing a k-nearest neighbor graph with $k = 20$, the default in the `Seurat` package [60]. Clustering result initialized by graphs with $k = 5$ and $k = 100$ were qualitatively similar. Both algorithms also have a resolution parameter as the primary parameter, a value which is correlated with the number of clusters ultimately identified; larger values of this parameter lead to fewer unique clusters. For the Leiden algorithm implemented with `igraph` [61], the default resolution parameter is *1*, which is also the default in `scanpy` and similar to the default of 0.8 in `Seurat`; however, this value led to each cell being placed in its own cluster. Instead, we followed the example of the *cluster_cells* function in `Monocle3` and set our default resolution parameter to be 10^{-5}; this parameter choice gave rise to a similar number of clusters as the default parameter of the Louvain algorithm as implemented in `Seurat`. We also repeated this experiment with resolution values in $\{10^{-4}, 5^{-5}, 5^{-6}\}$ (Additional file 1: Fig. S23); these values were selected to give an average number of unique clusters ranging from tens to hundreds, a representative range of cell types identified in single-cell studies. For each parameter value, we considered results from 20 different random seeds.

Numerical metrics for evaluating performance

To evaluate different methods' performance on detecting batch effects or biological signals, we used the two metrics for quantification:

1. ANOSIM R statistic, Eq. 6 above,
2. Average silhouette width using *silhouette* in R package `cluster`.

After obtaining the above two values from the calculated distance matrix D, we calculated bootstrap confidence intervals for each of the metrics. To do so, we defined the unique combinations of batch and biological condition. For each unique combination, we repeatedly sampled with replacement from samples in that combination; the union of the sampled samples from each combination resulted in a single full bootstrap sample. After obtaining the bootstrap sample for each run, we obtained the bootstrap distance matrix D_{boot} from the original distance matrix D by subsetting to the bootstrap sample ids. Finally, we calculate the two metrics based on D_{boot}. We repeated this for $B = 100$ bootstrap samples. For each of the metrics, we calculated percentile bootstrap confidence intervals by taking the 2.5% and 97.5% quantiles from the empirical distribution of the bootstrap distribution of the metrics.

Prediction of phenotype with GloScope representations

After using GloScope to obtain the symmetrized KL divergence matrices of COVID PBMC samples [30], we obtained their MDS embeddings with 10 dimensions. Sixty percent of the data points were reserved for training, and the remaining 40% were used for testing purpose. We applied SVM to classify sample's phenotype using the package `e1071` (i.e., COVID vs healthy) and 5-fold cross validation to tune the hyperparameters cost and γ [62, 63]. Finally, we used the test sets to assess the prediction algorithm by counting the prediction accuracy rate.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03398-1.

Additional file 1. Supplementary table and figures. Contains supplementary table S1 and figures S1-S30.

Additional file 2. Table for power and ANOSIM statistics calculated based on 100 simulation to detect between group difference when changing gene level variability β_k.

Additional file 3. Table for power and ANOSIM statistics calculated based on 100 simulation, to detect between group difference when changing cluster proportion only.

Additional file 4. Table for power and ANOSIM statistics calculated based on 100 simulation, to detect between group difference when changing log fold change and percentage of DE genes.

Additional file 5. Table for power and ANOSIM statistics calculated based on 100 simulation, to detect between group difference when changing log fold change concentrated in certain clusters.

Additional file 6. Table for power and ANOSIM statistics calculated based on 100 simulation, to detect between group difference when changing library sizes λ.

Additional file 7. Table for power and ANOSIM statistics calculated based on 100 simulation, to detect between group difference when changing sample size n.

Additional file 8. Review history.

Acknowledgements
None.

Review history
The review history is available as Additional file 8.

Peer review information
Veronique van den Berghe was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
HW, WT, BG, and EP contributed to the development and modeling work described in the manuscript. HW, WT, and EP wrote the main manuscript text. HW and WT developed figures and/or tables for the manuscript. All authors reviewed the manuscript and provided critical editing to main manuscript text. The authors read and approved the final manuscript.

Availability of data and materials

The SingleCellExperiment object with the quality-controlled UMI counts for the Allen mouse study [19] was downloaded using the R package AllenInstituteBrainData [64]. The raw count data for the skin rash study [20] was downloaded from the European Genome-Phenome Archive with associated Study ID EGAS00001002927 [65]. The processed data for the lung atlas study [21] was downloaded from the Broad Institute's Single Cell Portal [66]. The raw counts for the colon cancer study [22] were downloaded from the Gene Expression Omnibus (GEO) with ascension ID GSE178341 [67]. The h5ad file containing the raw and processed counts for the COVID PBMC study [30] was downloaded from Array Express under accession number E-MTAB-10026 [68]. The UMI counts for the Lupus PBMC study [31] were downloaded from the CellxGene portal [69]. The raw and processed counts for Lung and Liver fibrosis study [32] were downloaded from the Broad Institute's Single Cell Portal [70, 71].

Code for running these analyses, generating figures, and text files detailing dataset download source and specific processing choices are available in the following Zenodo repository: https://doi.org/10.5281/zenodo.13368089 [49]. The source code is under the Creative Commons Attribution 4.0 International license. The GloScope implementation is available in an accompanying Bioconductor package GloScope (https://bioconductor.org/packages/release/bioc/html/GloScope.html) [52], with all analyses performed on version 1.2.0.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 6 July 2023 Accepted: 20 September 2024

Published online: 08 October 2024

References

1. Crowell HL, Soneson C, Germain PL, Calini D, Collin L, Raposo C, et al. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. Nat Commun. 2020;11(1). https://doi.org/10.1038/s41467-020-19894-4.
2. Tiberi S, Crowell HL, Samartsidis P, Weber LM, Robinson MD. *distinct*: a novel approach to differential distribution analyses. Ann Appl Stat. 2023;17(2). https://doi.org/10.1214/22-aoas1689.
3. Zhang M, Liu S, Miao Z, Han F, Gottardo R, Sun W. IDEAS: individual level differential expression analysis for single-cell RNA-seq data. Genome Biol. 2022;23(1). https://doi.org/10.1186/s13059-022-02605-1.
4. Li CMC, Shapiro H, Tsiobikas C, Selfors LM, Chen H, Rosenbluth J, et al. Aging-associated alterations in mammary epithelia and stroma revealed by single-cell RNA sequencing. Cell Rep. 2020;33(13):108566. https://doi.org/10.1016/j.celrep.2020.108566.
5. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053–8. https://doi.org/10.1038/s41592-018-0229-2.
6. Arandjelovic O, Shakhnarovich G, Fisher J, Cipolla R, Darrell T. Face recognition with image sets using manifold density divergence. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1. IEEE; 2005. pp. 581–8. https://doi.org/10.1109/cvpr.2005.151.
7. Cox M, Cox T. Multidimensional scaling. In: Handbook of Data Visualization. Springer Handbooks Comp. Statistics. Berlin: Springer; 2008. https://doi.org/10.1007/978-3-540-33037-0_14.
8. Hofmann T, Schölkopf B, Smola AJ. Kernel methods in machine learning. Ann Stat. 2008;36(3). https://doi.org/10.1214/009053607000000677.
9. Wang X, Xing EP, Schaid DJ. Kernel methods for large-scale genomic data analysis. Brief Bioinforma. 2014;16(2):183–92. https://doi.org/10.1093/bib/bbu024.
10. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289–96. https://doi.org/10.1038/s41592-019-0619-0.
11. Forcato M, Romano O, Bicciato S. Computational methods for the integrative analysis of single-cell data. Brief Bioinforma. 2021;22(1):20–9. https://doi.org/10.1093/bib/bbaa042.
12. Orlova DY, Zimmerman N, Meehan S, Meehan C, Waters J, Ghosn EEB, et al. Earth mover's distance (EMD): a true metric for comparing biomarker expression levels in cell populations. PLoS ONE. 2016;11(3):e0151859. https://doi.org/10.1371/journal.pone.0151859.
13. Wagner J, Rapsomaniki MA, Chevrier S, Anzeneder T, Langwieder C, Dykgers A, et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. Cell. 2019;177(5):1330-1345.e18. https://doi.org/10.1016/j.cell.2019.03.005.
14. Chen WS, Zivanovic N, van Dijk D, Wolf G, Bodenmiller B, Krishnaswamy S. Uncovering axes of variation among single-cell cancer specimens. Nat Methods. 2020;17(3):302–10. https://doi.org/10.1038/s41592-019-0689-z.
15. Joodaki M, Shaigan M, Parra V, Bülow RD, Kuppe C, Hölscher DL, et al. Detection of PatIent-Level distances from single cell genomics and pathomics data with Optimal Transport (PILOT). Mol Syst Biol. 2023;20(2):57–74. https://doi.org/10.1038/s44320-023-00003-8.

16. Johnsson K, Wallin J, Fontes M. BayesFlow: latent modeling of flow cytometry cell populations. BMC Bioinformatics. 2016;17(1). https://doi.org/10.1186/s12859-015-0862-z.
17. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. Proc Natl Acad Sci. 2014;111(26). https://doi.org/10.1073/pnas.1408792111.
18. Orlova DY, Meehan S, Parks D, Moore WA, Meehan C, Zhao Q, et al. QFMatch: multidimensional flow and mass cytometry samples alignment. Sci Rep. 2018;8(1). https://doi.org/10.1038/s41598-018-21444-4.
19. Yao Z, van Velthoven CTJ, Nguyen TN, Goldy J, Sedeno-Cortes AE, Baftizadeh F, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. Cell. 2021;184(12):3222-3241.e26. https://doi.org/10.1016/j.cell.2021.04.021.
20. Cheng JB, Sedgewick AJ, Finnegan AI, Harirchian P, Lee J, Kwon S, et al. Transcriptional programming of normal and inflamed human epidermis at single-cell resolution. Cell Rep. 2018;25(4):871–83. https://doi.org/10.1016/j.celrep.2018.09.006.
21. Melms JC, Biermann J, Huang H, Wang Y, Nair A, Tagore S, et al. A molecular single-cell lung atlas of lethal COVID-19. Nature. 2021;595(7865):114–9. https://doi.org/10.1038/s41586-021-03569-1.
22. Pelka K, Hofree M, Chen JH, Sarkizova S, Pirl JD, Jorgji V, et al. Spatially organized multicellular immune hubs in human colorectal cancer. Cell. 2021;184(18):4734-4752.e20. https://doi.org/10.1016/j.cell.2021.08.003.
23. Clarke KR. Non-parametric multivariate analyses of changes in community structure. Aust J Ecol. 1993;18(1):117–43. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x.
24. Somerfield PJ, Clarke KR, Gorley RN. Analysis of similarities (ANOSIM) for 2-way layouts using a generalised ANOSIM statistic, with comparative notes on Permutational Multivariate Analysis of Variance (PERMANOVA). Austral Ecol. 2021;46(6):911–26.
25. Crowell H, Germain PL, Soneson C, Sonrel A, Robinson MD. muscat: multi-sample multi-group scRNA-seq data analysis tools. Bioconductor. 2022. Version 1.13.0. https://code.bioconductor.org/browse/muscat/RELEASE_3_13/. Accessed 5 May 2022.
26. Singh S, Póczos B. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS'16. Red Hook: Curran Associates Inc.; 2016. pp. 1225–33.
27. Noshad M, Moon KR, Sekeh SY, Hero AO. Direct estimation of information divergence using nearest neighbor ratios. In: 2017 IEEE International Symposium on Information Theory (ISIT). IEEE; 2017. pp. 903–7. https://doi.org/10.1109/isit.2017.8006659.
28. Wang Q, Kulkarni SR, Verdú S. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. IEEE Trans Inf Theory. 2009;55(5):2392–405. https://doi.org/10.1109/tit.2009.2016060.
29. Zhao P, Lai L. Minimax optimal estimation of KL divergence for continuous distributions. IEEE Trans Inf Theory. 2020;66(12):7787–811. https://doi.org/10.1109/tit.2020.3009923.
30. Stephenson E, Reynolds G, Botting RA, Calero-Nieto FJ, Morgan MD, Tuong ZK, et al. Single-cell multi-omics analysis of the immune response in COVID-19. Nat Med. 2021;27(5):904–16. https://doi.org/10.1038/s41591-021-01329-2.
31. Perez RK, Gordon MG, Subramaniam M, Kim MC, Hartoularos GC, Targ S, et al. Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. Science. 2022;376(6589). https://doi.org/10.1126/science.abf1970.
32. Fabre T, Barron AMS, Christensen SM, Asano S, Bound K, Lech MP, et al. Identification of a broadly fibrogenic macrophage subset induced by type 3 inflammation. Sci Immunol. 2023;8(82). https://doi.org/10.1126/sciimmunol.add8945.
33. Adams TS, Schupp JC, Poli S, Ayaub EA, Neumark N, Ahangari F, et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. Sci Adv. 2020;6(28). https://doi.org/10.1126/sciadv.aba1983.
34. Osorio D, Cai JJ. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. Bioinformatics. 2020;37(7):963–7. https://doi.org/10.1093/bioinformatics/btaa751.
35. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016;17(1). https://doi.org/10.1186/s13059-016-0888-1.
36. Ramirez Flores RO, Lanzer JD, Dimitrov D, Velten B, Saez-Rodriguez J. Multicellular factor analysis of single-cell data for a tissue-centric understanding of disease. eLife. 2023;12. https://doi.org/10.7554/elife.93161.
37. Wang G, Qi J. PET image reconstruction using kernel method. IEEE Trans Med Imaging. 2015;34(1):61–71. https://doi.org/10.1109/tmi.2014.2343916.
38. Wang S, Jiang Y, Chung FL, Qian P. Feedforward kernel neural networks, generalized least learning machine, and its deep learning with application to image classification. Appl Soft Comput. 2015;37:125–41. https://doi.org/10.1016/j.asoc.2015.07.040.
39. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16(1). https://doi.org/10.1186/s13059-015-0805-z.
40. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert JP. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018;9(1). https://doi.org/10.1038/s41467-017-02554-5.
41. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun. 2019;10(1). https://doi.org/10.1038/s41467-018-07931-2.
42. Van den Berge K, Perraudeau F, Soneson C, Love MI, Risso D, Vert JP, et al. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. Genome Biol. 2018;19(1). https://doi.org/10.1186/s13059-018-1406-4.
43. Jiang R, Sun T, Song D, Li JJ. Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome Biol. 2022;23(1). https://doi.org/10.1186/s13059-022-02601-5.
44. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36(5):421–7. https://doi.org/10.1038/nbt.4091.

45. Scrucca L, Fop M, Murphy B T, Raftery E Adrian. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R Journal. 2016;8(1):289. https://doi.org/10.32614/rj-2016-021.
46. Wang Q, Kulkarni S, Verdu S. A nearest-neighbor approach to estimating divergence between continuous random vectors. In: 2006 IEEE International Symposium on Information Theory. IEEE; 2006. pp. 242–6. https://doi.org/10.1109/isit.2006.261842.
47. Boltz S, Debreuve E, Barlaud M. High-dimensional statistical measure for region-of-interest tracking. IEEE Trans Image Process. 2009;18(6):1266–83. https://doi.org/10.1109/tip.2009.2015158.
48. Beygelzimer A, Kakadet S, Langford J, Arya S, Mount D, Li S. FNN: fast nearest neighbor search algorithms and applications. R. Version 1.1.4, 2024. https://CRAN.R-project.org/package=FNN.
49. Wang H, Torous W, Purdom E. GloScope analysis. Zenodo. 2024. https://doi.org/10.5281/zenodo.13368089. Accessed 23 Aug 2024.
50. Ledergor G, Weiner A, Zada M, Wang SY, Cohen YC, Gatt ME, et al. Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. Nat Med. 2018;24(12):1867–76. https://doi.org/10.1038/s41591-018-0269-2.
51. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating single-cell analysis with Bioconductor. Nat Methods. 2019;17(2):137–45. https://doi.org/10.1038/s41592-019-0654-x.
52. Wang H, Torous W, Gong B, Purdom E. GloScope. Bioconductor. 2023. https://doi.org/10.18129/B9.bioc.GloScope.
53. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. Cell. 2018;175(4):984-997.e24. https://doi.org/10.1016/j.cell.2018.09.006.
54. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10):P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008.
55. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9(1). https://doi.org/10.1038/s41598-019-41695-z.
56. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. Nat Rev Genet. 2023;24(8):550–72. https://doi.org/10.1038/s41576-023-00586-w.
57. Zhang S, Li X, Lin J, Lin Q, Wong KC. Review of single-cell RNA-seq data clustering for cell-type identification and characterization. RNA. 2023;29(5):517–30. https://doi.org/10.1261/rna.078965.121.
58. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:1–5.
59. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566(7745):496–502.
60. Hao Y, Hao S, Andersen-Nissen E, Mauck WM III, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.
61. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Systems:1695. https://igraph.org.
62. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
63. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R. 2023. Version 1.7-14. https://CRAN.R-project.org/package=e1071.
64. Righelli D. AllenInstituteBrainData R package. GitHub. https://github.com/drighelli/AllenInstituteBrainData. Accessed 14 June 2022.
65. Cheng JB, Harirchian P. RNA-seq analysis of human skin. European Genome-phenome Archive. Accession No. EGAS00001002927. https://ega-archive.org/studies/EGAS00001002927. Accessed 10 July 2022.
66. Izar B. A molecular single-cell lung atlas of lethal COVID-19. Single Cell Portal. https://singlecell.broadinstitute.org/single_cell/study/SCP1219. Accessed 8 July 2022.
67. Pelka K, Chen JH, Anderson AC, Rozenblatt-Rosen O, Regev A, Hacohen N. A single cell atlas of MMRd and MMRp colorectal cancer. Gene Expr Omnibus. Accession No. GSE178341. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178341. Accessed 13 Mar 2023.
68. Morgan M. Deciphering the molecular immune response to COVID-19 using single cell multi-omics. Array Express. Accession No. E-MTAB-10026. http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10026/. Accessed 22 Mar 2021.
69. Ye CJ. Single-cell RNA-seq reveals the cell-type-specific molecular and genetic associations to lupus. CellxGene portal. https://cellxgene.cziscience.com/collections/436154da-bcf1-4130-9c8b-120ff9a888f2. Accessed 2 Apr 2023.
70. Fabre T, Barron AMS, Christensen SM, Asano S, Bound K, Lech MP, et al.. Identification of a broadly fibrogenic macrophage subset induced by type 3 inflammation: human lung fibrosis scRNAseq atlas. Single Cell Portal. https://singlecell.broadinstitute.org/single_cell/study/SCP2155. Accessed 11 Apr 2024.
71. Fabre T, Barron AMS, Christensen SM, Asano S, Bound K, Lech MP, et al.. Identification of a broadly fibrogenic macrophage subset induced by type 3 inflammation: human liver fibrosis scRNAseq atlas. Single Cell Portal. https://singlecell.broadinstitute.org/single_cell/study/SCP2154. Accessed 11 Apr 2024.

Publisher's Note