Visualizing scRNA-Seq Data at Population Scale with GloScope

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

[1]Graduate Group in Biostatistics, University of California, Berkeley, USA

[2]Department of Statistics, University of California, Berkeley, USA.

[3]Center for Computational Biology, University of California, Berkeley, USA.

[*]Corresponding author(s). Email(s): epurdom@berkeley.edu

Contrubuting authors: hao_wang@berkeley.edu, wtorous@berkeley.edu, jorothy_gong@berkeley.edu

Abstract

Increasingly scRNA-Seq studies explore the heterogeneity of cell populations across different samples and its effect on an 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 its GloScope representation. We implement GloScope on scRNA-Seq datasets from study designs ranging from 12 to over 300 samples. These examples 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

1 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 [Crowell et al., 2020, Tiberi et al., 2020, Zhang et al., 2022]. Beyond differential expression analysis, sample-level analysis that exist are generally limited to comparisons of the relative proportions of different cell-types between groups of samples [Li et al., 2020a]. 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.

2 Results

2.1 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 [Lopez et al., 2018] 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.

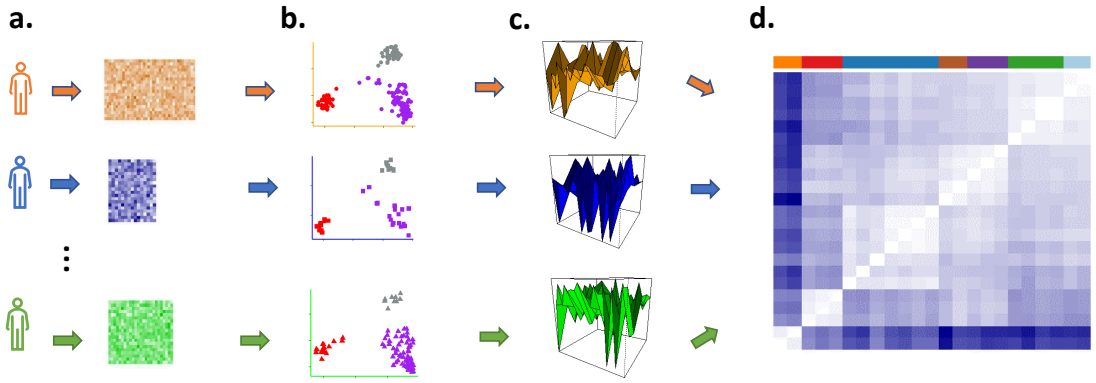


Fig. 1: Illustration of the GloScope representation of a sample's scRNA-Seq data matrix X_i as a distribution $\hat{F}_i$. (a) Each sample contributes a $g \times m_i$ matrix of gene expression values. (b) A lower dimensional latent representation is estimated across all cells and samples, resulting 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, $d(\hat{F}_i, \hat{F}_j)$.

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 Figure 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 repre-

sentation of probability distributions [Arandjelovic et al., 2005] (see Methods for details).

For example, the pairwise divergences between GloScope-represented samples can be given as input to canonical divergence analysis methods such as Multidimensional Scaling [Cox and Cox, 2001], 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 and exploratory assessment of the strength of phenotypic differences between our samples. We primarily concentrate in this work on assessing the use of the GloScope representation for the purpose of visualization and exploratory data analysis, but our representation can 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. [2008], Wang et al. [2014]).

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 [Orlova et al., 2016, Wagner et al., 2019, Li et al., 2020b, Chen et al., 2020], 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 [Orlova et al., 2016, Johnsson et al., 2016, Bruggner et al., 2014, Orlova et al., 2018]. 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. However, GloScope applied to the clusters can be used for

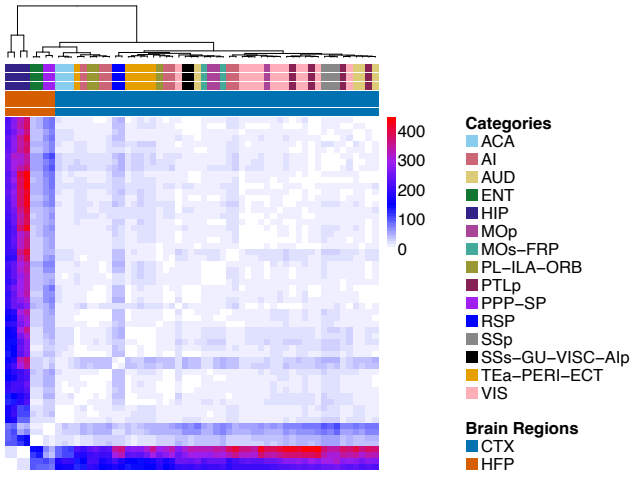
more formal hypothesis testing of significant global differences in cell-type composition.

2.2 *Visualization of patient and sample phenotypes using GloScope representations*

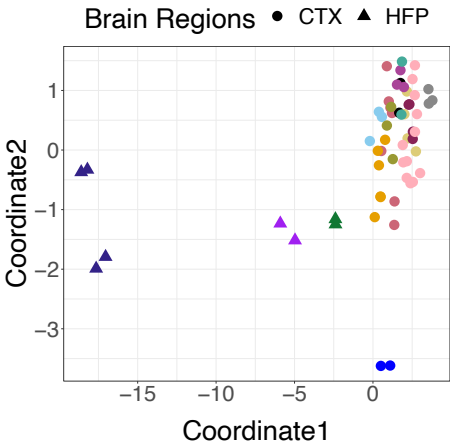
In this section we demonstrate the utility of the GloScope representation to visualize and evaluate sample-level phenotypic differences. As an initial illustration, we consider two datasets with replicate samples collected for each phenotype, where the phenotypes have well-known biological differences 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 [Yao et al., 2021]. 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 Figure 2(a), 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, Cheng et al. [2018] 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 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

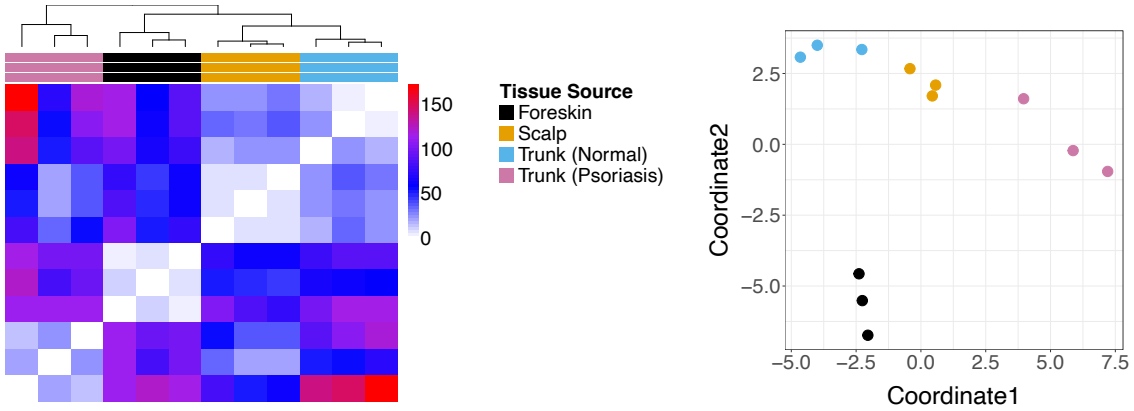


(a) Heatmap of KL pairwise dissimilarities



(b) 2D MDS plot of dissimilarity matrix

Fig. 2: Demonstration of the GloScope representation on scRNA-Seq data collected from different regions of mice cortex from 59 mice samples [Yao et al., 2021]. (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.



(a) Heatmap of KL pairwise distance

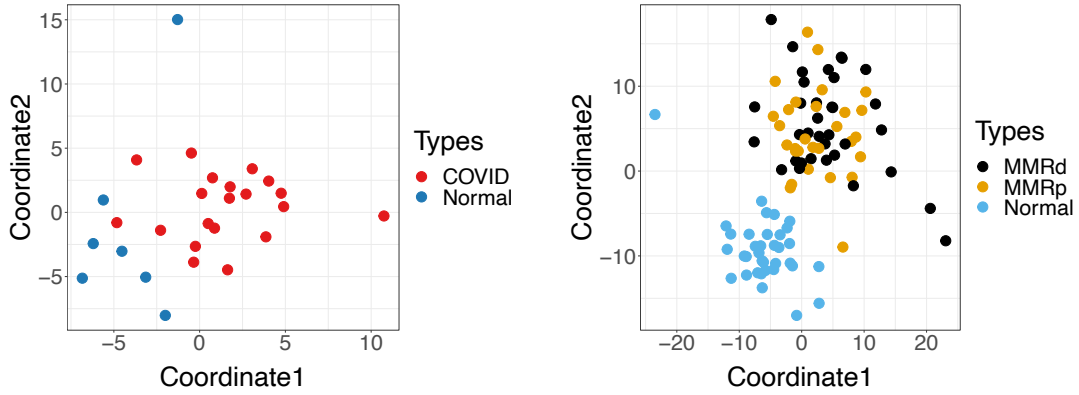
(b) 2D MDS plot of patients' skin samples distance matrix

Fig. 3: GloScope representation of skin rash ScRNA-Seq data collected from 12 patients in various locations and conditions in Cheng et al. [2018]. (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.

data in Figure 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. [2018].

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 Melms et al. [2021], 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 [Pelka et al., 2021]. The use of GloScope on these datasets demonstrates its utility for the visualization of both sample and phenotype variability. For the COVID lung samples (Figure 4(a)), we can easily see the separation between COVID-infected and healthy donors, matching the observation of Melms et al. [2021] 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 (Figure 4(b)). Though the two types of tumors do not separate in this visualization, an Analysis of Similarities (ANOSIM, Clarke [1993]) test of significance 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).



(a) Lung tissue [Melms et al., 2021] (b) Colon cancer [Pelka et al., 2021]

Fig. 4: A two dimensional MDS representation of the dissimilarities between samples from (a) 27 samples of COVID lung atlas data that are either healthy samples of COVID patients from Melms et al. [2021]; (b) 99 samples colon samples from mismatch repair-proficient (MMRp) tumors, mismatch repair-deficient (MMRd) tumors and healthy samples from Pelka et al. [2021]. The dissimilarity matrices were calculated using the GMM density estimate based on PCA estimates of the latent space in 10 dimensions.

2.3 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 consider only two different phenotypic groups. Count matrices were generated from a pipeline modified from that presented in the R package `muscat` [Crowell et al., 2020] (see Methods for details).

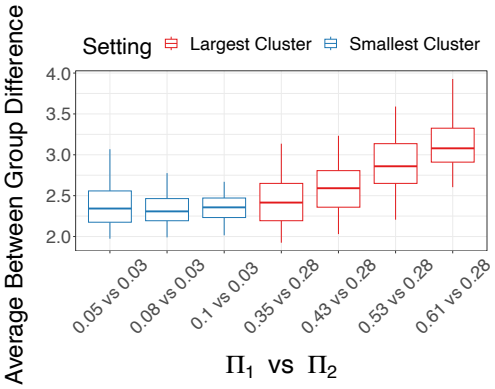
We focus on two basic biological scenarios that could causes phenotypic-based dissimilarity between scRNA-Seq samples which we would want the GloScope representation to accurately reflect: differential cell-type composition and differential gene expression. By

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.

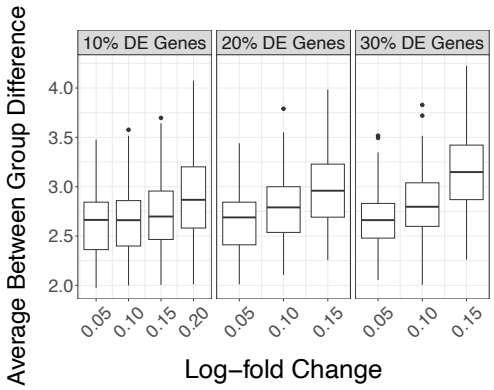
In our simulations we evaluate how well these two types of differences are detected by GloScope. We create datasets demonstrating either differentially expressed genes or differential cell-type composition. We see that the average differences between samples in different different phenotype groups, as measured by our GloScope representation, appropriately increase in response to both increased differences in global cell composition (Figure 5(a)) and increased differential gene expression (Figure 5(b)). 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 (Figure 5(c) 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 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 [Clarke, 1993]. 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 Methods 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.

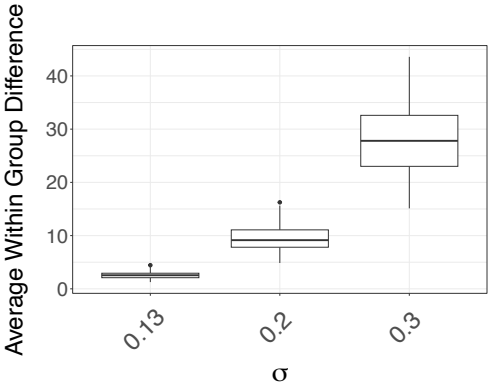
Using these power computations, we can see that changes in the sample variability



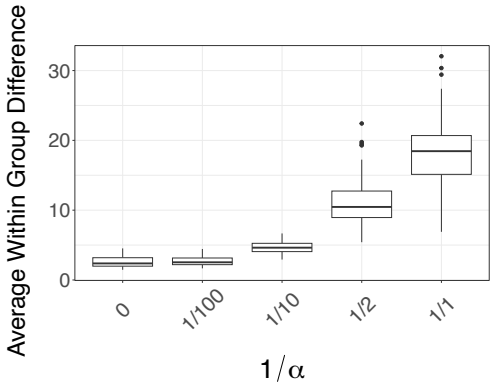
(a) Differential Cell Composition



(b) Differential Gene Expression



(c) Sample Variability in Gene Expression



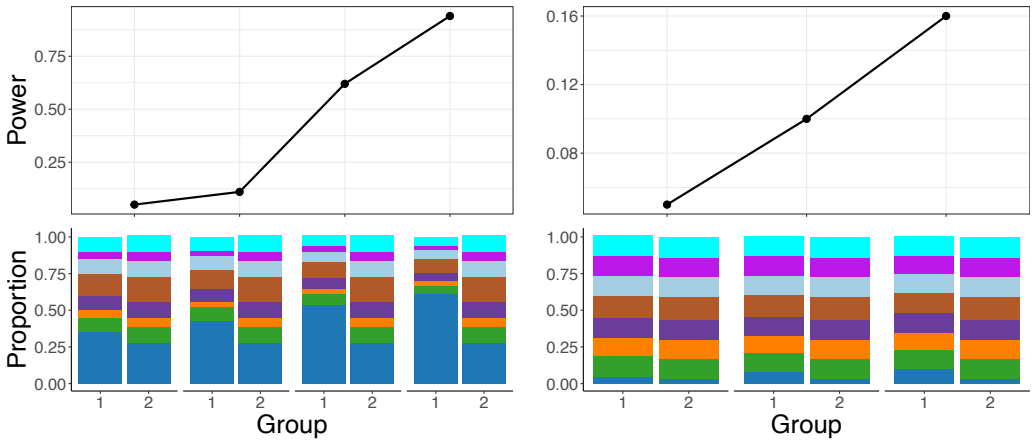
(d) Sample Variability in Cell Composition

Fig. 5: **GloScope captures simulated effects** Plots (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 boxplots 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 Supp Fig S7-S10

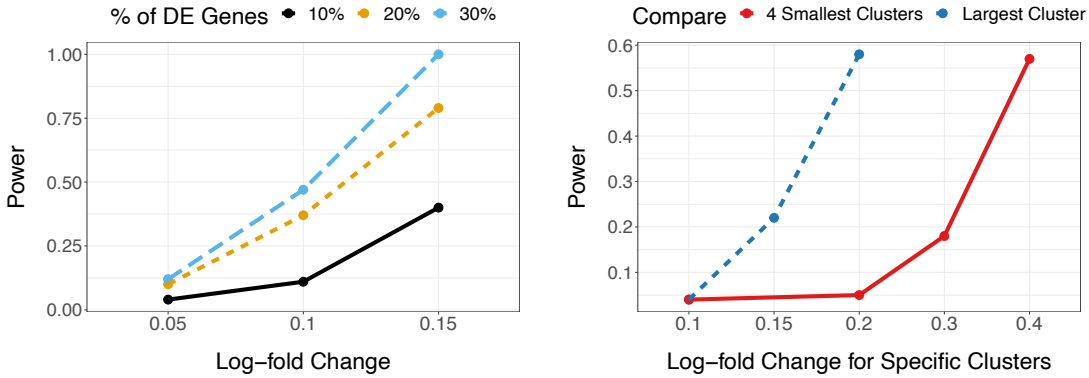
and sample size are reflected as expected in these power calculations: increasing all of these sources of variability naturally reduces the power (Supplementary Figure S1). These types of simulations, in conjunction with our GloScope representation, can be used to evaluate design choices at the sample-level, 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 (Figure 6(a)), and gene expression differences concentrated in small clusters are harder to detect than those found in large clusters (Figure 6(c)).

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 (Supplementary Figure S2). The latent variable representations given by scVI demonstrates much greater variability between samples than the those of PCA (Supplementary Figure S3), 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 (Supplementary Figure S4), 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 (Supp. Figure S4), but kNN is somewhat more powerful in the presence of cell-type composition changes (Supp. Figure S2). Applying both methods on a wide range of datasets (Supp. Figure S5-S6) 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 [Singh and Póczos, 2016], it is known



(a) Change in cell-type composition



(b) Change in proportion of genes DE, and average log-fold change

(c) Log-fold change differences concentrated in specific cell-subtypes

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 θ =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 $m = 5,000$ cells per sample; the sample level variability parameter σ is fixed at 0.13, and the sequencing depth $\lambda = 8.25$ (see Methods for details on these parameters). GloScope was calculated based on GMM density estimation with latent space representation via the first 10 dimensions of PCA.

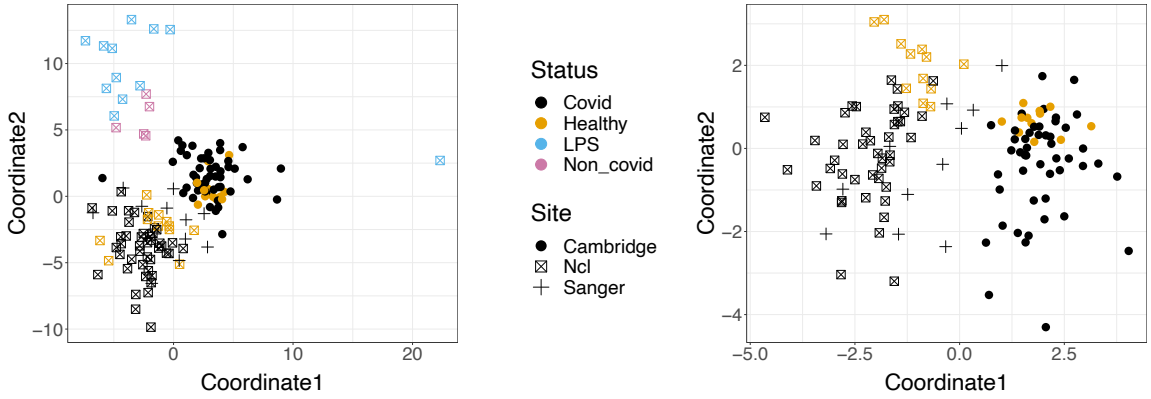
to exhibit downward finite sample bias due to underestimation of density in the tails of a distribution [Noshad et al., 2017, Wang et al., 2009, Zhao and Lai, 2020]. Due to these considerations, we relied on GMM estimates of density, though none of the results shown qualitatively change if kNN estimates are used instead.

2.4 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 [Stephenson et al., 2021] 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 7(a) shows these samples after applying MDS to the pairwise divergences calculated from the GloScope representation for the 143 samples of the study.

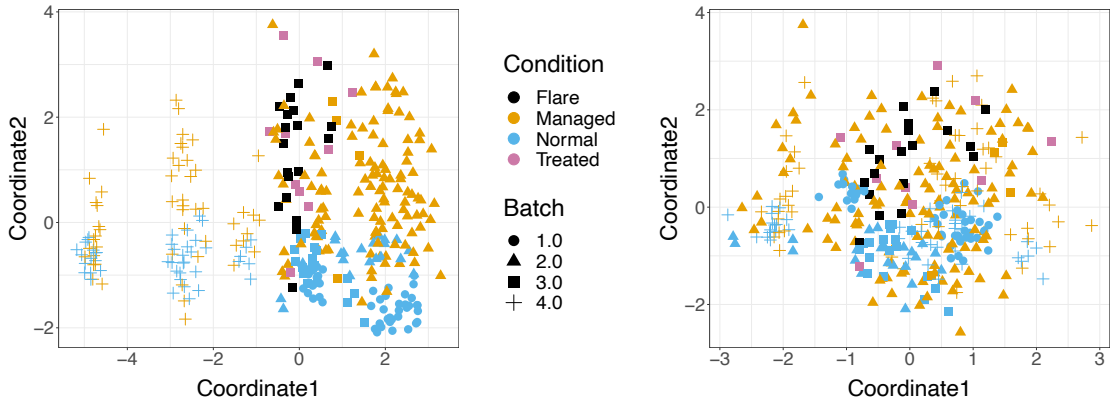
The visualization shows that both COVID patients and healthy donors are clearly separated from patients with other respiratory conditions (LPS and non-COVID). The other noticeable pattern is that the remaining patients do not show a strong separation between the COVID and Healthy phenotypes, but do appear 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 (Figure 7(b)). Exploration of the provided sample data from Stephenson et al. [2021] 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 (Supplementary Figure 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. [2021] indicates that samples from these different sites underwent



(a) MDS plot of all 143 samples.

(b) MDS plot of COVID and Healthy samples

Fig. 7: GloScope representation applied to samples sequenced in Stephenson et al. [2021]. 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 Supplementary Figure S12.



(a) Without batch correction (b) With batch correction via Harmony

Fig. 8: GloScope representation applied to a Systemic lupus erythematosus (SLE) dataset of 336 samples from 261 patients [Perez et al., 2022]. 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.

different sequencing steps such as cell isolation and library preparations (and the original analysis in Stephenson et al. [2021] corrected for potential batch effects by applying the batch correction method, Harmony [Korsunsky et al., 2019]).

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 [Perez et al., 2022]. 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 (Figure 8(a)). After application of Harmony to this data based on the batch, our GloScope representation shows much greater intermingling of the data from different batches (Figure 8(b)). 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.

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. These examples immediately demonstrates the power of our GloScope representation for exploratory data analysis. Current visualization strategies of scRNA-Seq data generally consist of applying tools such as UMAP or tSNE to create a two-dimensional visualization

of the individual cells, as in Supplementary Figure S11. However, batch effects are often due to variables that vary per sample or patient, such as the hospital of collection. Detecting such shared differences based on visualization of the individual cells would generally be quite difficult, particularly for large numbers of samples. Visualization of samples based on our GloScope representation immediately highlights the differences in these samples.

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 [Crowell et al., 2020, Zhang et al., 2022]; 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. We create such a PCA visualization of the pseudo-bulk of several of the datasets mentioned above (Supp Fig S13). 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 SLE data, the pseudobulk representation does not identify the strong batch effects seen in our GloScope representation.

Furthermore, the pseudo-bulk strategy is based on summarizing across individual genes, usually raw counts. Many public datasets provide other normalized versions of the data (e.g. residuals); similarly many batch-correction methods, like Harmony [Korsunsky et al., 2019], provide a batch-corrected latent variable representation. None of these are obvious candidates for a pseudo-bulk approach. Our GloScope representation requires as input only a latent-variable representation of the data 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 visualize the data before and after batch correction with the Harmony algorithm (Figure 8), allowing us to confirm that the Harmony algorithm has removed much of the differences between batches.

3 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.

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-varible 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.

4 Conclusion

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.

5 Methods

5.1 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., Arandjelovic et al., 2005, Wang and Qi, 2015, Wang et al., 2015]

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 2,000-8,000 (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 [Pierson and Yau, 2015, Risso et al., 2018, Eraslan et al., 2019, Van den Berge et al., 2018, Jiang et al., 2022].

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 $\sigma : R^d \to R^g$. 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 Section 2.1, 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` [Risso et al., 2018] or `scVI` [Lopez et al., 2018], or techniques that simultaneously remove batch effects and estimate a latent space, such as `Harmony` [Korsunsky et al., 2019] or `fastMNN` [Haghverdi et al., 2018].

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 applies a Gaussian mixture model to the Z_{ic} to estimate h_i, the density associated with the distribution H_i, and then calculates $d(\hat{H}_i, \hat{H}_j)$ as our estimate of $d(F_i, F_j)$. 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, motivates our use of Gaussian mixture models (GMMs). We use the R package `mclust` [Scrucca et al., 2016] 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 [Wang et al., 2006, Boltz et al., 2009]. 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

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

where d is the dimension of the latent space [Wang et al., 2006, Boltz et al., 2009]. We implement this strategy using the `FNN` package to estimate the symmetrized KL divergence between sample i and sample j [Beygelzimer et al., 2022].

5.2 *Simulating scRNA-Seq data*

5.2.1 Simulation Model

To simulate population-level scRNA-Seq data with which we benchmark our methodology, we follow the model introduced by the `muscat` R package. We would note that this 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 `muscat` package 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 in `muscat` 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 in the `muscat` package 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 Crowell et al. [2020]); for simplicity, we allowed only the two most typical gene expression patterns, which are EE (equally expressed) and DE (differentialy 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 (Supp. Figure S14); 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.

Supplementary Figure S15 shows the effect of changing different parameters (σ and log-fold change), visualized using UMAP on an illustrative example.

5.2.2 Simulation Settings

In following the above strategy of selecting parameters, we randomly chosen 5 COVID samples from the COVID-19 PBMC dataset, [Stephenson et al., 2021]. 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 5,000, which is similar to the average cell per samples in several datasets (e.g. Stephenson et al. [2021], Melms et al. [2021], Pelka et al. [2021]). 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 Figures 5(d)).

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 $F_{i,c}$.

Supp Table S1- S6 provide the different parameter settings that were run and their resulting power and average ANOSIM values corresponding to the figures shown here.

5.2.3 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 [Clarke, 1993]. The test statistic is calculated as:

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

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 determine 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 $\alpha = 0.05$.

5.3 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 ensur-

ing 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 immune cell 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, are available in the following GitHub repository: `https://github.com/epurdom/GloScope_analysis` [Wang et al., 2023a].

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. [2018] 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. [2021] which provided only PCA embeddings that we used directly).

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 $2,000$ 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 $2,000$ 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 [Amezquita et al., 2020]. Each of the 2,000 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 GitHub repository: `https://github.com/epurdom/GloScope` [Wang et al., 2023b].

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. Based on 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. [2022]; this study processed some tissue samples in multiple processing cohorts, and we 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. [2018] had duplicate encodings, which we standardized, and one sample with a mislabeled phenotype. For multiple myeloma cells from Ledergor et al. [2018], we parsed concatenated strings into patient, observation period, and phenotype indicators. Two colorectal tumor samples from Pelka et al. [2021] were sequenced with two technologies, and we only considered the replicates using the newer technology.

We also chose to remove samples with less than 50 cells. This excluded 2 samples from Ledergor et al. [2018] (`AB3178` and `AB3195`) and 1 from Pelka et al. [2021] (`C119N`). We noted that one sample from Ledergor et al. [2018] (`AB3461`), had extreme divergences with other samples. We removed all cells from this sample for the results presented in this paper.

5.4 Pseudobulk data analysis

For datasets where raw count data are available, we performed pseudobulk analysis by summing each sample's cell entries across each genes using `muscat`'s *aggregateData* function [Crowell et al., 2020]. 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*.

5.5 Prediction of phenotype with GloScope representation

After using GloScope to obtain the symmetrized KL divergence matrices of COVID PBMC samples [Stephenson et al., 2021], we obtained their MDS embeddings with 10 dimensions. 60% of the data points were reserved for training and rest 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 valiadtion to tune the hyperparameters cost and γ [Cortes and Vapnik, 1995, Meyer et al., 2022]. Finally, we used the test sets to assess the prediction algorithm by counting the prediction accuracy rate.

Declarations

- Ethics approval and consent to participate

  Not applicable.

- Consent for publication

  Not applicable.

- Availability of data and materials

  Code for running these analyses, generating figures, and text files detailing dataset download source and specific processing choices are available in the following GitHub repository: `https://github.com/epurdom/GloScope_analysis` [Wang et al., 2023a]. The R package for GloScope is available in the GitHub repository: `https://github.`

com/epurdom/GloScope [Wang et al., 2023b], and will be submitted to Bioconductor shortly.

- Competing interests

  The authors declare no competing interests.

- Funding

  This work has been supported by NIH grant 1R01GM144493, NIH grant U19MH114830, NSF training grant DMS RTG 1745640, and a Chan Zuckerberg Initative Data Insights Award. EP is a Chan Zuckerberg Biohub investigator.

- 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.

- Acknowledgements

  Not applicable.

References

Helena L. Crowell, Charlotte Soneson, Pierre-Luc Germain, Daniela Calini, Ludovic Collin, Catarina Raposo, Dheeraj Malhotra, and Mark D. Robinson. Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature Communications*, 11(1), 2020. doi: 10.1038/s41467-020-19894-4.

Simone Tiberi, Helena L Crowell, Pantelis Samartsidis, Lukas M Weber, and Mark D Robinson. distinct: A novel approach to differential distribution analyses. 2020. doi: 10.1101/2020.11.24.394213.

Mengqi Zhang, Si Liu, Zhen Miao, Fang Han, Raphael Gottardo, and Wei Sun. Ideas: Individual level differential expression analysis for single-cell rna-seq data. *Genome Biology*, 23(1), 2022. doi: 10.1186/s13059-022-02605-1.

Carman Man-Chung Li, Hana Shapiro, Christina Tsiobikas, Laura M. Selfors, Huidong Chen, Jennifer Rosenbluth, Kaitlin Moore, Kushali P. Gupta, G. Kenneth Gray, Yaara Oren, and et al. Aging-associated alterations in mammary epithelia and stroma revealed by single-cell rna sequencing. *Cell Reports*, 33(13):108566, 2020a. doi: 10.1016/j.celrep.2020.108566.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018. doi: 10.1038/s41592-018-0229-2.

O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005. doi: 10.1109/cvpr.2005.151.

Trevor F. Cox and Michael A. A. Cox. *Multidimensional scaling*. Chapman and Hall, 2001.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3), 2008. doi: 10.1214/009053607000000677.

X. Wang, E. P. Xing, and D. J. Schaid. Kernel methods for large-scale genomic data analysis. *Briefings in Bioinformatics*, 16(2):183–192, 2014. doi: 10.1093/bib/bbu024.

Darya Y. Orlova, Noah Zimmerman, Stephen Meehan, Connor Meehan, Jeffrey Waters, Eliver E. Ghosn, Alexander Filatenkov, Gleb A. Kolyagin, Yael Gernez, Shanel Tsuda, and et al. Earth mover's distance (emd): A true metric for comparing biomarker expression levels in cell populations. *PLOS ONE*, 11(3), 2016. doi: 10.1371/journal.pone.0151859.

Johanna Wagner, Maria Anna Rapsomaniki, Stéphane Chevrier, Tobias Anzeneder, Claus Langwieder, August Dykgers, Martin Rees, Annette Ramaswamy, Simone Muenst, Savas Deniz Soysal, and et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell*, 177(5), 2019. doi: 10.1016/j.cell.2019.03.005.

Carman Man-Chung Li, Hana Shapiro, Christina Tsiobikas, Laura M. Selfors, Huidong Chen, Jennifer Rosenbluth, Kaitlin Moore, Kushali P. Gupta, G. Kenneth Gray, Yaara

Oren, and et al. Aging-associated alterations in mammary epithelia and stroma revealed by single-cell rna sequencing. *Cell Reports*, 33(13):108566, 2020b. doi: 10.1016/j.celrep.2020.108566.

William S. Chen, Nevena Zivanovic, David van Dijk, Guy Wolf, Bernd Bodenmiller, and Smita Krishnaswamy. Uncovering axes of variation among single-cell cancer specimens. *Nature Methods*, 17(3):302–310, 2020. doi: 10.1038/s41592-019-0689-z.

Kerstin Johnsson, Jonas Wallin, and Magnus Fontes. Bayesflow: Latent modeling of flow cytometry cell populations. *BMC Bioinformatics*, 17(1), 2016. doi: 10.1186/s12859-015-0862-z.

Robert V. Bruggner, Bernd Bodenmiller, David L. Dill, Robert J. Tibshirani, and Garry P. Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, 111(26), 2014. doi: 10.1073/pnas.1408792111.

Darya Y. Orlova, Stephen Meehan, David Parks, Wayne A. Moore, Connor Meehan, Qian Zhao, Eliver E. Ghosn, Leonore A. Herzenberg, and Guenther Walther. Qfmatch: Multidimensional flow and mass cytometry samples alignment. *Scientific Reports*, 8(1), 2018. doi: 10.1038/s41598-018-21444-4.

Zizhen Yao, Cindy T.J. van Velthoven, Thuc Nghi Nguyen, Jeff Goldy, Adriana E. Sedeno-Cortes, Fahimeh Baftizadeh, Darren Bertagnolli, Tamara Casper, Megan Chiang, Kirsten Crichton, and et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12), 2021. doi: 10.1016/j.cell.2021.04.021.

Jeffrey B. Cheng, Andrew J. Sedgewick, Alex I. Finnegan, Paymann Harirchian, Jerry Lee, Sunjong Kwon, Marlys S. Fassett, Justin Golovato, Matthew Gray, Ruby Ghadially, and et al. Transcriptional programming of normal and inflamed human epidermis at single-cell resolution. *Cell Reports*, 25(4):871–883, 2018. doi: 10.1016/j.celrep.2018.09.006.

Johannes C. Melms, Jana Biermann, Huachao Huang, Yiping Wang, Ajay Nair, Somnath Tagore, Igor Katsyv, André F. Rendeiro, Amit Dipak Amin, Denis Schapiro, and et al. A molecular single-cell lung atlas of lethal covid-19. *Nature*, 595(7865):114–119, 2021. doi: 10.1038/s41586-021-03569-1.

Karin Pelka, Matan Hofree, Jonathan H. Chen, Siranush Sarkizova, Joshua D. Pirl, Vjola Jorgji, Alborz Bejnood, Danielle Dionne, William H. Ge, Katherine H. Xu, and et al. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell*, 184(18), 2021. doi: 10.1016/j.cell.2021.08.003.

K. R. Clarke. *Non-parametric multivariate analyses of changes in community structure*, volume 18. 1993. doi: 10.1111/j.1442-9993.1993.tb00438.x.

Shashank Singh and Barnabás Póczos. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. *Advances in neural information processing systems*, 29, 2016.

Morteza Noshad, Kevin R Moon, Salimeh Yasaei Sekeh, and Alfred O Hero. Direct estimation of information divergence using nearest neighbor ratios. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 903–907. IEEE, 2017.

Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

Puning Zhao and Lifeng Lai. Minimax optimal estimation of kl divergence for continuous distributions. *IEEE Transactions on Information Theory*, 66(12):7787–7811, 2020.

Emily Stephenson, Gary Reynolds, Rachel A. Botting, Fernando J. Calero-Nieto, Michael D. Morgan, Zewen Kelvin Tuong, Karsten Bach, Waradon Sungnak, Kaylee B. Worlock, Masahiro Yoshida, and et al. Single-cell multi-omics analysis of the immune response in covid-19. *Nature Medicine*, 27(5):904–916, 2021. doi: 10.1038/s41591-021-01329-2.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, Soumya Raychaudhuri, and et al. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296, 2019. doi: 10.1038/s41592-019-0619-0.

Richard K. Perez, M. Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C. Hartoularos, Sasha Targ, Yang Sun, Anton Ogorodnikov, Raymund Bueno, Andrew Lu, and et al. Single-cell rna-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, 376(6589), 2022. doi: 10.1126/science.abf1970.

Guobao Wang and Jinyi Qi. Pet image reconstruction using kernel method. *IEEE Transactions on Medical Imaging*, 34(1):61–71, 2015. doi: 10.1109/TMI.2014.2343916.

Shitong Wang, Yizhang Jiang, Fu-Lai Chung, and Pengjiang Qian. Feedforward kernel neural networks, generalized least learning machine, and its deep learning with application to image classification. *Applied Soft Computing*, 37:125–141, 2015. doi: 10.1016/j.asoc.2015.07.040.

Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1), 2015. doi: 10.1186/s13059-015-0805-z.

Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9:284, 2018. URL https://doi.org/10.1038/s41467-017-02554-5.

Gökcen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S. Mueller, and Fabian J. Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1), 2019. doi: 10.1038/s41467-018-07931-2.

Koen Van den Berge, Fanny Perraudeau, Charlotte Soneson, Michael I. Love, Davide Risso, Jean-Philippe Vert, Mark D. Robinson, Sandrine Dudoit, and Lieven Clement. Observation weights unlock bulk rna-seq tools for zero inflation and single-cell applications. *Genome Biology*, 19(1), 2018. doi: 10.1186/s13059-018-1406-4.

Ruochen Jiang, Tianyi Sun, Dongyuan Song, and Jingyi Jessica Li. Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome biology*, 23(1):1–24, 2022.

Laleh Haghverdi, Aaron T Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018. doi: 10.1038/nbt.4091.

Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317, 2016. URL https://doi.org/10.32614/RJ-2016-021.

Qing Wang, Sanjeev Kulkarni, and Sergio Verdu. A nearest-neighbor approach to estimating divergence between continuous random vectors. *2006 IEEE International Symposium on Information Theory*, 2006. doi: 10.1109/isit.2006.261842.

S. Boltz, E. Debreuve, and M. Barlaud. High-dimensional statistical measure for region-of-interest tracking. *IEEE Transactions on Image Processing*, 18(6):1266–1283, 2009. doi: 10.1109/tip.2009.2015158.

Alina Beygelzimer, Sham Kakadet, John Langford, Sunil Arya, David Mount, and Shengqiao Li. *FNN: Fast Nearest Neighbor Search Algorithms and Applications*, 2022. URL https://CRAN.R-project.org/package=FNN. R package version 1.1.3.1.

Hao Wang, William Torous, and Elizabeth Purdom. Gloscope_analysis, 2023a. URL https://github.com/epurdom/GloScope_analysis.

Guy Ledergor, Assaf Weiner, Mor Zada, Shuang-Yin Wang, Yael C. Cohen, Moshe E. Gatt, Nimrod Snir, Hila Magen, Maya Koren-Michowitz, Katrin Herzog-Tzarfati, and et al. Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. *Nature Medicine*, 24(12):1867–1876, 2018. doi: 10.1038/s41591-018-0269-2.

Robert A Amezquita, Aaron TL Lun, Etienne Becht, Vince J Carey, Lindsay N Carpp, Ludwig Geistlinger, Federico Marini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, et al. Orchestrating single-cell analysis with bioconductor. *Nature methods*, 17(2): 137–145, 2020.

Hao Wang, William Torous, Boying Gong, and Elizabeth Purdom. Gloscope, 2023b. URL https://github.com/epurdom/GloScope.

Livnat Jerby-Arnon, Parin Shah, Michael S. Cuoco, Christopher Rodman, Mei-Ju Su, Johannes C. Melms, Rachel Leeson, Abhay Kanodia, Shaolin Mei, Jia-Ren Lin, and et al. A cancer cell program promotes t cell exclusion and resistance to checkpoint blockade. *Cell*, 175(4), 2018. doi: 10.1016/j.cell.2018.09.006.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2022. URL https://CRAN.R-project.org/package=e1071. R package version 1.7-11.