Taylor & Francis
Taylor & Francis Group

Human Vaccines & Immunotherapeutics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/khvi20

Statistical and machine learning methods for immunoprofiling based on single-cell data

Jingxuan Zhang, Jia Li & Lin Lin

To cite this article: Jingxuan Zhang, Jia Li & Lin Lin (2023) Statistical and machine learning methods for immunoprofiling based on single-cell data, Human Vaccines & Immunotherapeutics, 19:2, 2234792, DOI: 10.1080/21645515.2023.2234792

To link to this article: https://doi.org/10.1080/21645515.2023.2234792

Published online: 24 Jul 2023.

HUMAN VACCINES & IMMUNOTHERAPEUTICS
2023, VOL. 19, NO. 2, 2234792
https://doi.org/10.1080/21645515.2023.2234792

Taylor & Francis
Taylor & Francis Group

REVIEW

OPEN ACCESS Check for updates

Statistical and machine learning methods for immunoprofiling based on single-cell data

Jingxuan Zhang[a], Jia Li[b], and Lin Lin[a]

[a]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA; [b]Department of Statistics, Pennsylvania State University, University Park, PA, USA

ABSTRACT

Immunoprofiling has become a crucial tool for understanding the complex interactions between the immune system and diseases or interventions, such as therapies and vaccinations. Immune response biomarkers are critical for understanding those relationships and potentially developing personalized intervention strategies. Single-cell data have emerged as a promising source for identifying immune response biomarkers. In this review, we discuss the current state-of-the-art methods for immunoprofiling, including those for reducing the dimensionality of high-dimensional single-cell data and methods for clustering, classification, and prediction. We also draw attention to recent developments in data integration.

ARTICLE HISTORY
Received 13 March 2023
Revised 30 June 2023
Accepted 4 July 2023

KEYWORDS
Immunoprofiling; machine learning; data integration; single-cell data

Introduction

Immunoprofiling refers to the measurement and analysis of the immune responses in individuals, with the aim of discovering and understanding the specific immune responses associated with a particular disease or condition. Research tasks in immunoprofiling are diverse. For example, it is important to identify the immune responses that offer protection against infection or disease, as well as the prediction of an individual's response to intervention. In many cases, clinical data collected during the course of intervention (such as treatment or vaccination) contain measurements of a group of individuals and their medical status and immune measurements before and/or after the interventions. As such, statistical analysis and machine learning methods are necessary to reveal the relationships between various quantities across different time points and in association with individuals' outcomes.

The insights gained from immunoprofiling can be utilized to guide the development of new treatments or vaccines and to evaluate the effectiveness of existing therapies. For example, in the context of infectious diseases such as HIV, to date, RV144 is the only HIV vaccine trial that shows a modest protective effect with the vaccine reduced the risk of HIV infection by 31.2%.[1] Immunoprofiling can help identify the immune responses associated with this reduced risk, which can guide the development of new

HIV vaccines by focusing on eliciting these responses in vaccinated individuals, thus potentially increasing the chances of a safer and more effective vaccine.[2–5] Similarly, immunotherapy is a rapidly growing field in the treatment of multiple cancer types.[6–11] The clinical benefits are impressive, with the ability to produce the greatest possibility of long-term survival. However, only a fraction of patients achieves this. Immunoprofiling can be used to evaluate the effectiveness of immunotherapy treatments by measuring changes in the immune response and determining whether a therapy is effective. It can also identify potential biomarkers that predict treatment response, allowing for personalized immunotherapeutic regimens for patients.

This review focuses on the single-cell technologies used in immunoprofiling. Those technologies provide assessments of gene and/or protein expressions to understand cellular immune responses, potentially bringing more insights and advancing the progress of immunotherapy and vaccine development. We list below the major technologies that generate single-cell data.

(1) *Modern flow cytometry*: Over 20 parameters are simultaneously measured on hundreds of thousands to millions of single cells.
(2) *Mass cytometry or Cytometry by time-of-flight (CyTOF)*: Over 40 parameters are measured at the single-cell level.[12]
(3) *Single-cell RNA sequencing (scRNA-seq)*: This technology can determine the transcriptome of individual cells.[13,14]
(4) *Single-cell multi-omics data*: This technology enables the simultaneous measurements of multiple biological layers in each individual cell, such as genome and transcriptome (G&T-seq),[15] gDNA-mRNA sequencing (DR-Seq),[16] DNA methylation and transcriptome (scM&T-seq),[17] epitope and transcriptome (CITE-seq),[18] and nucleosome, transcriptome, and methylation (scNMT-seq).[19]

While those advanced technologies enable researchers to achieve unprecedented resolution in tackling cellular heterogeneity, the generated high-dimensional and large-scale single-cell

CONTACT Lin Lin l.lin@duke.edu Department of Biostatistics and Bioinformatics, Duke University, 2424 Erwin Road, Suite 1102 Hock Plaza Box 2721, Durham, NC, USA.

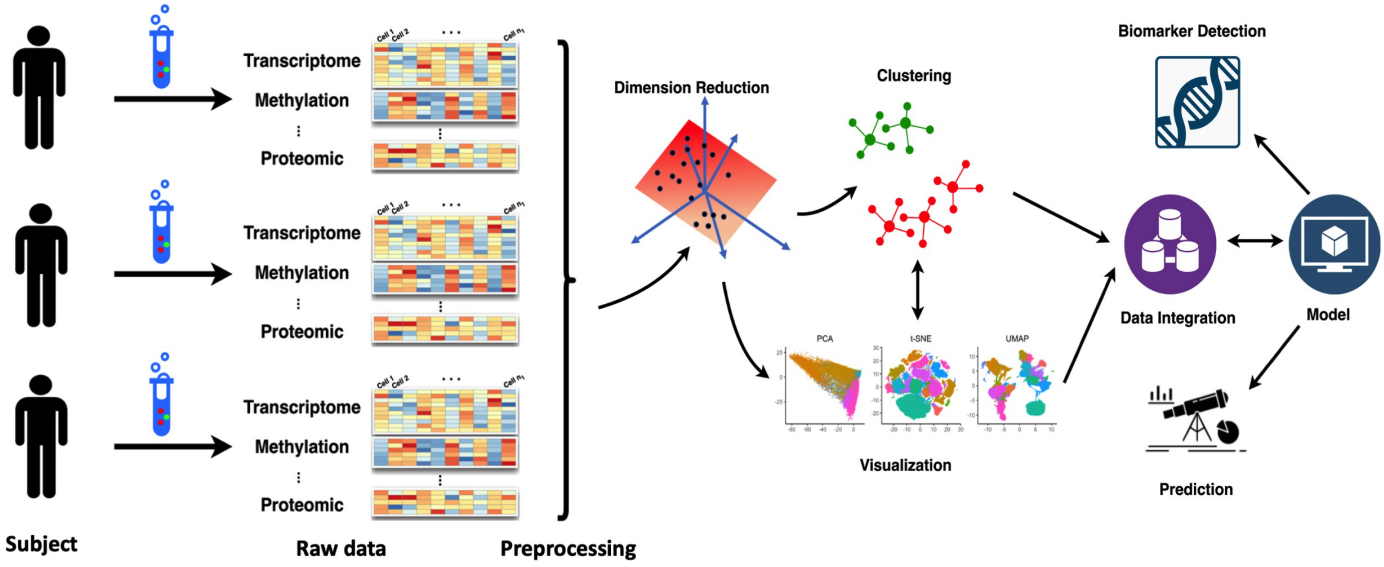

Figure 1. One example pipeline of single-cell data analysis.

data pose significant analysis challenges for capturing and understanding immune cellular heterogeneity and detecting rare immune cell subsets.

The objective of this paper is to provide an overview of the current state of machine learning techniques and statistical modeling approaches used for analyzing and modeling single-cell immunological data. The paper focuses on the biological areas where advanced analysis methods are popularly developed. We assume that the single-cell data have been preprocessed[20–22] and potential batch effects[23–33] have been properly removed. One example pipeline of single-cell data analysis, which is the focus of this review paper, is shown in Figure 1. The main analysis tasks shown in the figure are discussed in the following sections.

The rest of the paper is organized as follows. Section 2 contains a review of dimension reduction and visualization techniques for single-cell data. In Section 3, we provide an overview of clustering analysis. Subsequently, Section 4 is on prediction analysis methods, and Section 5 is on data integration techniques. In Section 6, we present a case study analysis and discuss practical challenges. Finally, we conclude with discussions in Section 7. Due to space constraints, we summarize a few popular methods in each section and refer to some existing review papers in case readers are interested in more details about these methods or the ideas of other relevant methods.

Single-cell data dimension reduction and visualization

Single-cell data visualization is an indispensable first step in comprehending high-dimensional single-cell data and allows the visual comparisons of cellular heterogeneity across different conditions. Dimension reduction is a key technique to enable the visualization of high-dimensional data in a low-dimensional space. It can also facilitate downstream analysis. Xiang et al.[34] provides a comprehensive review to compare different dimension reduction methods for scRNA-seq data. Those methods are also applicable to cytometry data. Traditional dimension reduction methods such as PCA (principal component analysis), t-SNE (t-distributed stochastic neighbor embedding),[35] and UMAP[36] are frequently used to obtain the low-dimensional representation of the single-cell data. PCA, being a linear method, identifies the linear combinations of the original variables with the highest variance. Thus, PCA has limitations in capturing complex nonlinear structures in high-dimensional data. On the other hand, t-SNE and UMAP are both nonlinear dimension reduction methods designed to reduce dimensionality while preserving the local structure, such as clusters. t-SNE uses t-distributed conditional probability to quantify pairwise similarities between samples. It aims to find a low-dimensional mapping to preserve the pairwise similarities from the high-dimensional data as well as possible. UMAP is a variant of t-SNE. Neither method is designed to preserve the global structure of the data and may produce false clusters in the low-dimensional representation.[37]

TriMap is a recent method proposed to preserve the global structure of high-dimensional data.[38] This method is based on the concept of triplets, which are used to compare data points and evaluate the relationships between them. Specifically, TriMap assesses whether "point i is closer (or more similar) to point j than to point k" and utilizes this triplet-based approach to encourage neighboring data points to remain close to each other in the low-dimensional space while making distant points to remain far apart. On the other hand, PaCMAP (Pairwise Controlled Manifold Approximation Projection) is another method designed to optimize both local and global structures.[37] PaCMAP achieves this by selecting a small subset of pairwise distances, which are then used to control the mapping of the high-dimensional data onto a lower-dimensional space. By doing so, PaCMAP ensures that the mapping preserves the relative distances between nearby points (local structure) while allowing for more flexibility in the mapping of more distant points (global structure).

The use of deep neural networks (DNNs) is another strategy for dimensionality reduction, given their ability to handle large-scale datasets. DNNs are composed of multiple layers of interconnected nodes, also known as neurons. Each neuron typically applies a linear transform on its multi-dimensional input and passes a one-dimensional output through a non-linear function such as a sigmoid function or a hard thresholding function, which is then fed as an input feature to one or more nodes at the next layer. The input of the first layer is the raw data, and the output of the last layer is the final prediction or classification result. An introduction to DNN is referred to Goodfellow et al.[39] The autoencoder (AE) is a type of DNN which comprises two main components: an encoder and a decoder. The encoder maps the high-dimensional input data, such as cells, to a low-dimensional representation, while the decoder maps the low-dimensional representation back to the original high-dimensional space. During training, the AE aims to minimize the discrepancy between the original input and the reconstructed output. In contrast to the discriminative nature of the AE, the variational autoencoder (VAE) is a generative model that learns a latent representation distribution rather than a specific vector (for each cell), as learned by the AE. By using the learned latent representation distribution, the VAE can generate examples of the latent representations of cells. Thus, compared to standard AEs, VAEs not only can reduce dimensionality but also quantify the uncertainty of the latent representation. Many existing packages use AE or VAE for dimension reduction and visualization. For example, scScope[40] uses AE to reduce dimensionality while simultaneously imputing the dropout events. To do so, scScope generates imputed input data based on the decoder output through an imputer layer. The imputed data is then sent back to the encoder, and an updated latent representation is learned in an end-to-end manner. scVI[41] uses a VAE to learn scRNA-seq latent representation by leveraging information from similar cells and genes to approximate the underlying distribution while accounting for batch effects. scDHA[42] operates in a hierarchical manner, where a non-negative kernel AE is used to filter out non-significant information and a VAE to map data onto a low-dimensional space. This approach enables the separation of noises from biological signals in single-cell data. Brendel et al.[43] provides a recent review paper to discuss the application of DNNs for scRNA-seq data analysis.

Cluster analysis for cell subsets identification

The primary objective of clustering single-cell data is to uncover and understand cellular heterogeneity and potentially reveal novel cell subsets. In the context of immunoprofiling, clustering facilitates the identification of different immune cell types and the recognition of distinctive immune cell patterns across conditions. The information derived from clustering analysis could potentially substantiate existing conjectures, inspire new hypotheses, and inform the design of further experimentation.

Cytometry data often involve large sample sizes, sometimes exceeding millions of cells per sample. The main challenge in this context lies in identifying cellular heterogeneity, particularly rare cell subsets, as the low probability clusters tend to be "concealed" by large background clusters in the data. On the other hand, for scRNA-seq data, the clustering task is complicated by the high dimensionality of the data, with the number of variables (genes) typically exceeding the number of cells. In addition, the high variability and drop-out rates, along with the low capture efficiency of scRNA-seq data, and their potential batch effects, can further complicate the clustering process. The single-cell multi-omics data raise new challenges for clustering analysis because different omics features are combined, resulting in more heterogeneous and higher dimensional data. The analysis of single-cell multi-omics data will be discussed specifically in Section 5. In this section, we review clustering methods developed to cluster cytometry and scRNA-seq data, including K-means, hierarchical clustering, mixture model, community detection, and DNN-based approaches. In Figure 2, different types of clustering methods are compared in terms of their ability to handle various levels of data structures and their computational complexity. Several review papers on clustering analysis are provided by Weber and Robinson[44]; Petegrosso et al.[45]; Krzak et al.[46]; Liu et al.[47]; Yu et al.[48]

K-means based clustering methods

K-means clustering is the most popular clustering approach, which iteratively applies Lloyd's algorithm[49] to find a prespecified number of K cluster centers (centroids) representing the mean of the data points in each cluster. The objective is to assign cells to groups (centroids) such that cells belonging to the same group are close, but those in different groups are far apart. The major advantage of

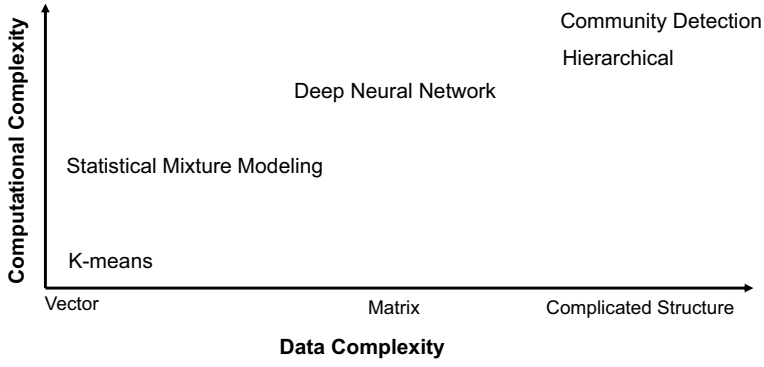

Figure 2. Overview and comparison between each type of clustering methods.

K-means clustering is that it scales linearly with the number of cells, so it is suitable for large datasets.

As K-means is locally optimal, it cannot guarantee the identification of the global minimum, i.e., the unique clustering solution. For scRNA-seq data, this limitation is overcome by repeated utilization of K-means with different initializations and deriving a consensus clustering result, as demonstrated in SC3.[50] Another issue always encountered in practice is how to determine the number of clusters *K*. SC3[50] addresses this issue by a method based on random matrix theory to determine the optimal *K*. RaceID[51] uses the Gap statistic to determine the number of clusters. With respect to the cytometry data, to address the challenge of finding the optimal number of clusters, FlowPeaks[52] performs a two-stage clustering approach. It first applies K-means with a large initial *K*, larger than the expected number of clusters in the data, and then applies a hierarchical clustering method based on the distance between the centroids of the *K* clusters to combine the nearest clusters into one cluster. Thus, FlowPeaks can effectively remove noise while ensuring that the final number of clusters is appropriate for the data.

K-means clustering also tends to bias toward identifying equal-sized clusters, potentially resulting in rare cell types being concealed within a larger group. To mitigate this issue, RaceID augments K-means with outlier detection to pinpoint rare cell types.[51] To reduce the bias of equal-size clustering, Flock employs a grid-based method to identify high-density regions and use them as initial centroids for K-means clustering. This reduces the bias of equal-size clustering.[53] The recently developed method, DisRFC (Dissimilarity Random Forest Clustering),[54] combines random forest and K-means to address some other limitations of K-means such as sensitivity to initialization and outliers along with the previously discussed issues (*e.g.*, the tendency to converge to local optima and the bias toward equal-size clusters).

Hierarchical clustering based methods

Hierarchical clustering is a method for building a hierarchy of clusters based on the connectivity between data points. This approach creates a dendrogram, a tree-like structure, by repeatedly merging the closest pairs of clusters based on some similarity measure, either by starting with all data points in a single cluster and then splitting the cluster into smaller clusters until each data point is in its own cluster (divisive clustering), or by starting with each data point in its own cluster and merging the clusters iteratively until a single cluster is obtained (agglomerative clustering).

Unlike K-means clustering, hierarchical clustering does not require the prior specification of the number of clusters. The number of clusters can be determined afterward through visual inspection or by using measures such as elbow methods, Silhouette analysis, or Gap statistic.

To address the challenges of high-dimensional data, pcaReduce[55] utilizes PCA to reduce the dimension before performing hierarchical clustering. SINCERA[56] employs hierarchical clustering with centered Pearson correlation and average linkage as default to identify cell clusters. Additionally, CIDR[57] integrates an implicit imputation process to alleviate the effect of dropouts, uses principal coordinate analysis to reduce the dimension, and performs hierarchical clustering on the first few principal coordinates. The number of clusters is determined based on the Calinski – Harabasz index.[58]

FlowGrid[59] is a computational framework for analyzing flow cytometry data. It combines density-based clustering and hierarchical clustering. Density-based clustering partitions the data into subsets, while hierarchical clustering organizes the subsets into a hierarchical structure, helping in identifying rare and low-density cell populations. FlowGrid also allows users to explore and visualize data at different levels of granularity and identify clusters that are biologically relevant. Recently, FlowGrid demonstrated fast clustering of very large scRNA-seq data.

Both Louvain and Leiden algorithms are community-detection-based clustering methods commonly used for clustering large-scale single-cell data. The goal of community detection is to discover the communities in networks. A community in a network is a group of nodes having dense connections within the group and sparse connections with other groups. The Louvain algorithm is an agglomerative clustering method.[60] It starts with each node (cell) in its own community and iteratively merges communities to maximize modularity. The modularity is a measure of the degree to which nodes in a community are more densely connected to each other than to nodes outside the community. The Louvain algorithm has been shown to be fast and scalable, making it a popular choice for clustering large single-cell datasets. The Leiden algorithm is an extension of the Louvain algorithm to overcome the resolution limit of the Louvain algorithm. The resolution limit is a problem where there is a minimum community size able to be resolved. Thus, it occurs when the algorithm merges small communities into larger ones, resulting in a loss of resolution. Both Louvain and Leiden algorithms are implemented in various software packages for single-cell data analysis, such as Seurat,[61] SCANPY,[62] Monocle,[63] and PARC.[64]

Statistical mixture modeling approach

Statistical mixture modeling is a widely adopted framework for model-based clustering. A finite mixture model, which has a density function represented by a convex combination of densities in a parametric family, offers the advantage of improving the goodness of fit to data samples by increasing the number of mixture components. Among them, the finite Gaussian Mixture Model (GMM) is popularly used for clustering. The simplest approach to clustering based on a mixture model is to assign each mixture component to an individual cluster. However, clusters can have arbitrary shapes, and the parametric distribution of each mixture component is often inadequate to capture the different shapes of the clusters. Various strategies have been proposed to merge multiple mixture components so that an individual cluster can be more properly modeled.[65–70] Mixture modeling has become an established paradigm for clustering cytometry data.[67,68–70,71,72,73] The mathematical formulation of a GMM is presented in Figure 3(a). We also show the density function given by an example GMM containing six components in Figure 3(b). This plot shows that Gaussian mixture components may overlap substantially, and thus the components are not equivalent to clusters. In fact, the shape of

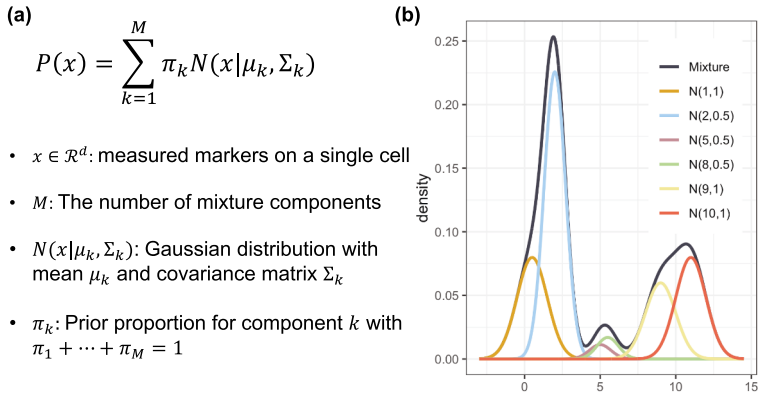

Figure 3. (a) Formula showing the density of a GMM. (b) The black curve shows the overall distribution of a GMM with 6 Gaussian distributions. Each Gaussian distribution is plotted in different colors.

the mixture density suggests that some components should be merged to form a cluster, an observation that has motivated some of the methods aforementioned.

Due to their high sparsity and high-dimensionality, scRNA-seq data are difficult to analyze. To overcome this challenge, ZIFA (Zero Inflated Factor Analysis) has been developed for data pre-processing before fitting the mixture model. ZIFA uses a factor analysis model for dimension reduction and accounts for the presence of zero inflation and technical noise in the data. Instead of reducing the dimension, HMMVB (Hidden Markov Model on Variable Blocks) is a mixture modeling framework that proposes to first partition variables into a sequence of subsets (variable blocks).[74] Such information is typically available for flow cytometry data from domain knowledge about the lineage, maturation, and activation of cells or can be computationally derived for scRNA-seq data. Hence, each additional variable block in the sequence corresponds to a lower-dimensional manifold for separating clusters at an increased level of granularity. Thus, it is effective at finding rare clusters for large-scale data and is computationally efficient. Under the framework of mixture modeling, variable selection methods have also been developed.[70,75,76] Through variable selection, we can not only reduce the dimension of data but also obtain more interpretable models.

DNN-based methods

DNN-based methods, particularly AEs and VAEs, have demonstrated great potential in clustering single-cell data. For example, scAIDE[77] is a robust and highly scalable framework for clustering scRNA-seq data. scAIDE first uses a customized AE to learn a good representation of data and then applies a random projection hashing-based K-means algorithm to detect rare cell subsets. Random projection hashing reduces computational complexity and makes scAIDE scalable for large datasets. However, the AE in scAIDE is a relatively simple architecture compared to other DNNs. Thus, scAIDE may not be the best choice for extracting an effective non-linear representation for complicated data. Similarly, scDMFK[78] uses VAE to learn the low-dimensional latent space and applies an adaptive fuzzy K-means algorithm with entropy regularization to perform probabilistic clustering. Noisy data points that do not clearly belong to any cluster are given less weights via entropy regularization penalty, a mechanism that reduces the effects of outliers in the data. Additionally, SAUCI[79] proposes to use novel regularizations imposed on AE architecture so that the learned representation is better for clustering, batch correction, denoising and imputation, and visualization. SAUCI not only clusters data but also provides solutions for several other problems commonly encountered in single-cell data analysis. One advantage of using AEs or VAEs for clustering is that they can learn a low-dimensional representation of high-dimensional single-cell data while incorporating prior knowledge or assumptions about the data. For example, incorporating batch information can improve clustering performance and reduce batch effects. By using this low-dimensional representation, clustering algorithms can be applied more efficiently and accurately. While AE and VAE have shown promise in clustering single-cell data, they have some limitations. AE may suffer from overfitting, resulting in poor generalization to new data. As for VAE, the generative modeling approach has difficulty treating small datasets due to the stochastic nature of its sampling process.

The various schools of approaches discussed above have their respective pros and cons. K-means is appealing for its simplicity. However, K-means is intrinsically related to clustering based on GMMs with several constraints imposed on the model parameters. In light of this, statistical mixture modeling is more generic and has a solid probabilistic foundation for estimation. The probabilistic perspective enables us to understand the limitations of any given model and may point to ways to overcome them. As a result, there is a rich literature on clustering by statistical mixture modeling, addressing various challenges encountered in high dimensional data. Agglomerative clustering, such as dendrograms, is flexible in the sense that only pairwise distances between instances are required. Users can tailor the definition of distance to embed prior knowledge. As long as the distance can be computed, the exact representation of each individual data point is not restricted. However, computing all the pairwise distances does not scale well with large data, and to manually define a distance can be difficult. DNN-based methods have several advantages compared with other

methods, such as their ability to handle and learn complex, high-dimensional data. However, DNNs require a large amount of training data, which are not always available. Furthermore, training DNNs is computationally expensive. It is also a challenge to configure and tune a typical DNN model because many hyperparameters are involved, such as the numbers of layers, neurons, and epochs, the learning rate, and the batch size.

Uncertainty assessment for clustering analysis

Clustering results obtained computationally are known to vary depending on different samples, algorithms used, or even initializations. To validate a clustering result, assessing the stability of the result seems to be the minimal we should do. While dimension reduction and visualization of the clustering results can provide a manual inspection, different dimension reductions can lead to different visualizations. Several clustering validation criteria have been proposed assuming that true cluster labels (cell types) are available, such as cluster stability, compactness, separation, and closeness to a given ground truth.[80] However, in most cases, true cell labels are not available, evaluating clustering stability is then regarded as the issue of assessing clustering uncertainty. The idea is to generate perturbed versions of the data by performing bootstrap sampling or adding noise and obtaining a collection of clustering results. A stability measure is defined as the average of pairwise distances between clustering results across different perturbed data. Various distances have been used for partitions, such as the Rand index, and the method that generates a more stable clustering result is preferred. Existing work on clustering stability primarily addresses stability at the level of overall clustering results. However, for studies in which cell clusters are considered new findings, assessing the uncertainty of individual clusters is more pertinent. Recently, Li et al.[81] and Zhang et al.[82] proposed aligning clusters across different perturbed data via soft matching solved by optimal transport. The main idea is illustrated in Figure 4. The cluster alignment enables quantification of the variation in the clustering result at the levels of both overall partitions and individual clusters. This method is useful in addressing the critical question of whether any cluster is an intrinsic or spurious pattern.

Predictive biomarkers detection

One major goal of immunoprofiling is to discern distinct immune response biomarkers capable of predicting an individual's response to a particular intervention, such as treatment or vaccination. Biomarkers that can predict treatment outcomes are essential for immunotherapy, which can guide treatment decisions, allowing for personalized immunotherapeutic regimens and ultimately improving patient outcomes. In the case of vaccination, the predictive biomarkers allow the identification of which individuals will respond to vaccines and which will not, thereby facilitating the design of more effective vaccines and their deployment to the public.

Single-cell data have emerged as a promising source for identifying immune response biomarkers. Since each individual will have at least one single-cell dataset, typically in the form of a matrix, there are two main types of approaches for predicting an individual's outcome based on single-cell data. In the first type, the idea is to transform every data matrix into a feature vector. Specifically, summary statistics derived from the matrix are used, such as cell subset proportions and summary measures on measured cell markers, including surface proteins, intracellular proteins, and gene expressions, all shown to be informative for predicting immune response. On the other hand, existing DNN-based methods can directly take the single-cell matrix data as an individual's feature matrix (Section 4.1). Because the matrix dimension needs to be fixed for a DNN, sampling based on the original data is often applied. In the second type of approaches, the single-cell data matrix is treated as distributional data (Section 4.2) since the order of the rows in the matrix has no particular meaning (each row corresponds to a cell). A basic way of forming the distributional data is to take the feature vector of each cell as an element in an unordered set and to assign an equal probability to the vector of each cell. A more sophisticated approach can involve statistical modeling of the single-cell data of every individual, *e.g.*, by GMM. Distributional data well preserve information in the data matrix but pose challenges for subsequent analysis.

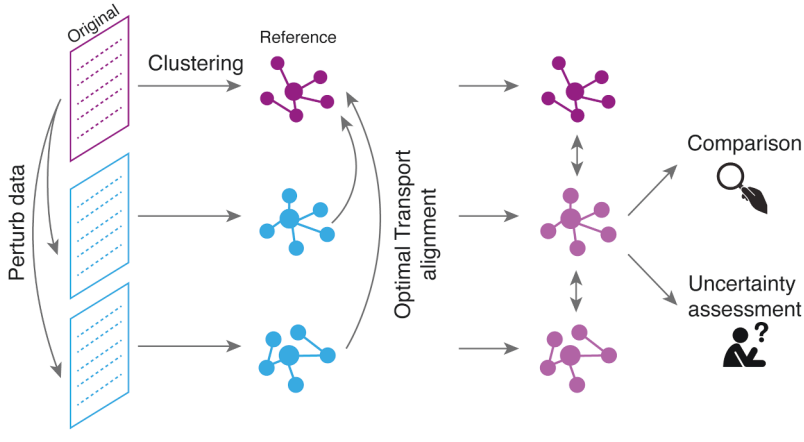

Figure 4. Uncertainty assessment for clustering results based on optimal transport.

Vector or matrix data

Summary statistics typically form a high-dimensional immune feature vector per individual, and furthermore, we would like to include other bulk data and individual-level covariates in the feature vector. It is thus necessary to reduce the dimension of the feature vectors before fitting any predictive model, for instance, by methods discussed in Section 2. Unfortunately, dimension reduction decreases (arguably removes) the interpretability of the original features, making it especially difficult to identify predictive biomarkers. Alternatively, univariate analysis, such as removing features with low variance or comparing immune features between two groups (*e.g.*, responders vs. non-responders) via statistical tests, can be an effective first step for variable screening before fitting any model and selecting predictive biomarkers.

Methods such as generalized linear regression, random forest,[83] and K-Nearest Neighbor (KNN) regression and classification[84] are commonly used as baseline models. For example, Babelomics[85] uses the medians of gene expression levels as features and applies random-forest-based classification to select predictive biomarkers, whereas a benchmark testing shows that a support-vector-machine-based method can obtain better results in practice on high-dimensional data.[86] Besides those default methods, by incorporating a linear mixed regression model with the individual-level random effect defined by the individual ID, Nowicka et al.[87] develops a new approach called HDCyto for the case of batch effect not removed in advance.

For regression-based methods, regularization is typically employed for data of high dimensions, *e.g.*, L_1 Lasso, L_2 ridge, and elastic net. Through regularization, variable selection/shrinkage is achieved, which prevents overfitting. Several measures of variable importance have been developed for random forest models, *e.g.*, Gini importance and permutation accuracy importance,[88] based on which variables can be selected. The interpretation of the selected variables varies depending on the choice of the importance measure. KNN regression and classification yield "black-box" prediction models that do not directly reveal which variables are important. A generic approach to select variables for black-box models is to apply step-wise greedy search (typically forward addition or backward deletion), which can be computationally intensive but nevertheless feasible when data are not too large or the training process is fast. Model selection criteria such as AIC (Akaike information criterion), BIC (Bayesian information criterion), and cross-validation accuracy are usually used to determine how many variables and which variables should be selected.

Although using summary statistics as features allows us to detect potential biomarkers conveniently, this approach has a notable drawback – other distributional characteristics of a biomarker exhibited across cells, such as multi-modality, skewness, and variance, are ignored. Recently, DNNs have gained popularity as powerful tools for predicting outcomes from high-dimensional single-cell matrix data. For example, both CyTOF DL[89] and CytoSet[90] are example DNNs that directly take the single-cell matrix data as input. Both methods leverage cell-invariant permutation functions or pooling layers in conjunction with classifier layers. However, these approaches can be sensitive to batch effects and become computationally intensive when the noise level is high. Additionally, the requirement of an equal number of cells across individuals may cause a loss of information since cells are usually re-sampled from the original data. In addition, as black-box prediction models, DNNs lack model interpretability and transparency, posing a barrier for the identification of predictive biomarkers.

Distributional data

One major difference between single-cell data and the most common data encountered in machine learning is that every instance (individual) in a study is a data cloud – an unordered set of feature vectors (cells). We call such instances distributional data in contrast to vector data. In the methods discussed above, various processes are applied first to convert every distributional instance into a vector, for example, one that contains proportions of cell subsets. One potential limitation of the existing paradigm is that useful information in the distribution for making predictions may be missed when condensing the distribution into summary statistics. A different approach has been proposed to treat distributional data directly when making predictions. The main idea is to build a pseudo density on the space of distributions. In particular, Qiao and Li[91] developed a pseudo-mixture model based on pairwise distances between instances, which is illustrated in Figure 5.

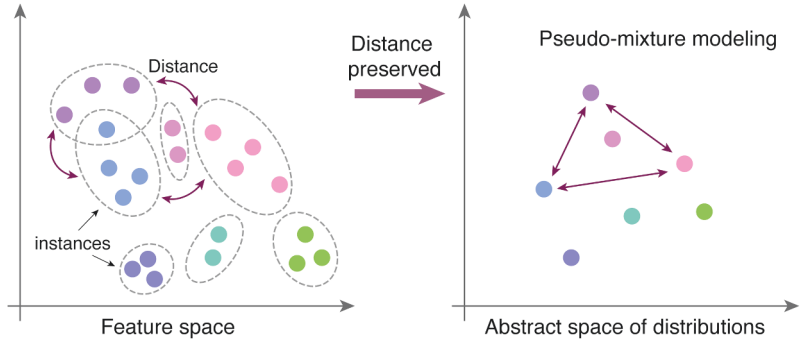

Figure 5. Pseudo-mixture modeling for distributional data. Left: Each individual/instance is represented by a distribution, indicated by a grey oval containing multiple feature vectors. The distribution is essentially a set of unordered vectors assigned with probabilities. Right: In the abstract space of distributions, each instance is a data point, and the distance between the distributions on the feature vector space is preserved in this abstract space.

In the method by Qiao and Li,[91] first, a distance between distributions is defined. One increasingly popular choice for a distance between distributions is the Wasserstein metric. For continuous distributions, the Wasserstein metric is usually estimated based on empirical distributions, that is, samples of distributions in which each sample point is assigned with a uniform weight. However, for high dimensional data, computing the Wasserstein distance based on samples suffers from the curse of dimensionality – an exponentially growing size of the sample in terms of the dimension is needed. Chen et al.[78] proposed to model high dimensional data using GMMs and then defined a semi-metric for GMMs, namely, Minimized Aggregated Wasserstein (MAW) distance. The MAW distance is computed by optimal transport between Gaussian components in two GMMs, where the Wasserstein metric between Gaussian distributions, provided by a closed formula, is taken as the baseline distance between the components. Once the distances between distributions are available, through a notation called hypothetical local mapping, the distances are used to estimate parameters in the pseudo-mixture model. Finally, the Bayes formula is applied to compute the posterior probabilities of classes based on the pseudo-mixture models estimated for each class. Although we have only used the pseudo-mixture model for classification applications, it is straightforward to extend the model to a regression setting.

One popular method for making predictions based on single-cell data relies on computing the proportions of cell types in a dataset. Suppose there are K cell types, a K-dimensional vector consisting of the proportions of cell types is obtained for every individual. Then any machine learning method for treating tabular data can be applied. This approach, however, requires that the cell type of each cluster in the single-cell dataset of any individual is known. If the clusters are generated by e.g., manual gating which is a default method for clustering flow cytometry data, the identities of clusters are given. On the other hand, if the clusters are generated by an algorithm and different clustering results are not aligned, we do not even know the correspondence between clusters of different individuals, not to mention the cell types of clusters. In such a scenario, the aforementioned approach cannot apply. In contrast, the pseudo-mixture-modeling approach works on the overall distributions instead of the clusters and thus does not require information on cell types. Another advantage of the pseudo-mixture-modeling approach is that it does not require a large collection of example cases to train. As will be shown by a case study in Section 6, the method yields competitive results for a collection of 24 individuals. DNN-based methods, on the contrary, require significantly larger data to train.

A disadvantage of the basic pseudo-mixture-modeling approach is that the classifier obtained will not directly point to biomarkers. Overcoming this disadvantage is an interesting future work. One approach we have envisioned is to first align components in the GMMs (one for each individual), for instance, by the method of Li et al.[81] We can then use a step-wise selection of components to find out which components are most useful for making prediction. These components can then serve as biomarkers. As the pseudo-mixture-modeling method is fast to train, step-wise selection of components is computationally efficient.

Data integration

There are two types of data integration problems: (1) Data integration with multiple views, which involves combining data (cells) from the same sample but collected using different experimental techniques or measuring multiple biological layers/modalities. Thus, the same set of cells are measured in different views. For example, integrating scRNA-seq data with single-cell epigenomic data to obtain a more comprehensive understanding of the regulatory mechanisms of gene expression in individual cells; (2) Data integration with multiple sources, which involves combining data from different samples or sources, such as different tissues, organisms, replicates, or platforms. In this case, the same type of measurement is made on different sets of cells. For example, integrating scRNA-seq data from multiple replicates to monitor the reproducibility of a biological experiment.

Data integration with multiple views

Recently, significant advances have been made in the field of single-cell isolation and barcoding technologies. This has provided researchers with a unique opportunity to simultaneously profile multiple views (omics) such as DNA, mRNA, and proteins at a single-cell resolution.[15–19,92] These innovative approaches offer a more comprehensive understanding of individual cells. The ultimate goal of integrating data from multiple views is to extract information from different modalities to enhance learning performance beyond what can be achieved with any single modality. This includes improvements in cell population characterization and regulatory networks construction. Nevertheless, multi-omics single-cell data presents unique challenges. The data is often sparse and heterogeneous among different omics feature spaces. In addition, different omics features can have large differences in dimensions. Different omics features do not necessarily carry equally important information toward a specific learning objective. To overcome these challenges, various integrative learning methods have been developed.

Most existing methods combine data from multiple views into one using weights, transformations, or simplification based on similarity or dimension reduction. Downstream analyses can then be performed based on the integrated data. Examples of such integration include Seurat V4,[93] which uses weighted nearest-neighbor to learn the weights of different views and generates a similarity graph of cells based on a weighted combination of views; CiteFuse[94] computes pairwise cell similarity matrices for each view and subsequently merges the similarity matrices into one using a similarity network fusion algorithm. Other methods such as MOFA[95] and MOFA+[96] both use factor analysis to project the high-dimensional data onto a common latent space and learn view-invariant information. MOFA+ scales to large datasets; MoClust[97] performs dimension reduction independently for each view using AEs and then employs contrastive learning to align the view-specific latent dimensions to form a fused representation of the data. In contrast, Cobolt,[98] scMM,[99] and scMVAE[100] use a multi-modal VAE to jointly model the multiple views and learn a joint embedding of the single-cell data.

These methods represent a diverse range of approaches to the integration of single-cell multi-omics data and offer different trade-offs in terms of accuracy, scalability, and interpretability.

Data integration with multiple sources

To achieve meaningful insights from single-cell data sets that are generated from multiple sources, it is imperative to address the issue of systematic variations, also known as batch effects, that can confound downstream analyses. In light of this, various methods have been developed to mitigate the impact of batch effects and align data sets to facilitate accurate comparisons and integrative analysis.

One such method is the mutual nearest neighbors (MNN) approach.[23] The MNN approach identifies cells in two data sets that share the nearest neighbors, and then utilizes the differences between these identified pairs to align one data set with the other. A similar strategy is implemented by the Seurat algorithm,[27] which first computes the MNN in a lower dimensional space and then performs data integration. Alternatively, other methods such as scVI[41] and scAlign[101] employ DNN embedding to align two data sets by seeking a common dimension-reduced space to encode the data sets. In a benchmark study conducted by Tran et al.[33] the performance of various methods, including MNN, Seurat, scVI, and scAlign, were compared in terms of batch-effect correction.

Once batch effects have been properly removed, downstream analysis can be conducted. A natural choice to integrate data from multiple sources is simply to combine them into a unified dataset for subsequent analysis. Another research direction is to consider the multi-source nature of the datasets. For example, Lin et al.[102] developed an analysis framework to combine clustering results acquired from multiple sources. In many cases, the clustering analysis is often performed on each dataset independently for reasons such as scalability and the need to identify rare cell subsets. Assuming clustering results are obtained from each source. The clusters must be labeled consistently across samples to carry out a meaningful integration and comparison among the cell clusters. To solve this problem, Lin et al.[102] proposed to use a GMM to summarize the clustering result of each data set, where the cluster-specific sample mean and sample covariance are used to estimate the mean vector and covariance matrix for each Gaussian component, respectively. The prior probability of every component is set to be the proportion of cells in the corresponding cluster. Given the set of GMMs, an integrated clustering result can be obtained based on the notion of Wasserstein barycenter. This framework allows for flexibility in the choice of batch-effect removal and clustering methods. The process is illustrated in Figure 6. Another method, LIGER,[103] uses integrative nonnegative matrix factorization to compute a low-dimensional representation across all the data sources. Clustering is then performed, and a search for shared clusters is conducted based on a shared-factor-neighborhood graph.

Case study and practical challenges

In this section, we apply multiple methods introduced in previous sections to analyze a dataset obtained from Chua et al.[104]

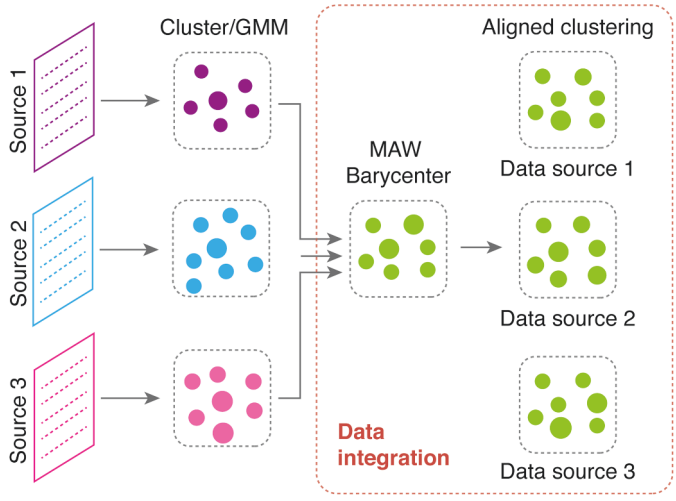

Figure 6. Data integration for clustering results obtained from different data sources by computing the MAW barycenter of GMMs.

This dataset contains scRNA-seq data obtained from 19 patients with COVID-19 and 5 COVID-negative donors. Specifically, all 19 patients and 5 donors were tested through the primary infection site: the nasopharyngeal area (NS), where the scRNA-seq is used to predict the outcome of COVID infection. Moreover, multiple samples from both the upper and lower airways were collected on two of the patients using bronchial protected specimen brushes (PSB) and bronchial lavages (BL). This dataset provides us a good example to demonstrate a data integration process for single-cell data. In addition, the gender and age of each individual are provided as meta-data. By applying Seurat using the R package Seurat to cluster the pooled NS cells from all 24 individuals, we show in Figure 7 the 22 identified cell subsets from a total of 160,528 cells using different visualization methods: PCA, t-SNE, and UMAP. It is clear that PCA, as a linear dimension reduction algorithm, is unable to discern nonlinear cell patterns. Moreover, the nonlinear dimension reduction methods yield remarkably different visualization. Low-dimensional visualization can be a convenient way to assess and identify potential biomarkers across conditions. For example, the two plots in the top row of Figure 8 show that the nonresident macrophage (nrMa) cell subset in one COVID-positive individual has a notably higher proportion than the same cell type in a selected COVID-negative individual. Similarly, by comparing samples obtained from different sites (NS, PSB and BL), neutrophils (Neu) is identified as the most dominant cell type in the NS sample, but this type of cells seldom appear in samples collected from PSB and BL.

Prediction

Next we perform prediction analysis to compare multiple methods for classifying an individual's COVID outcomes (positive versus negative) based on scRNA-seq data. Because of the small sample size, we use leave-one-out cross-validation (LOOCV) to evaluate prediction performance. A larger AUC (Area Under the Receiver Operating Characteristics Curve) for LOOCV indicates a more accurate prediction. We compare the following methods:

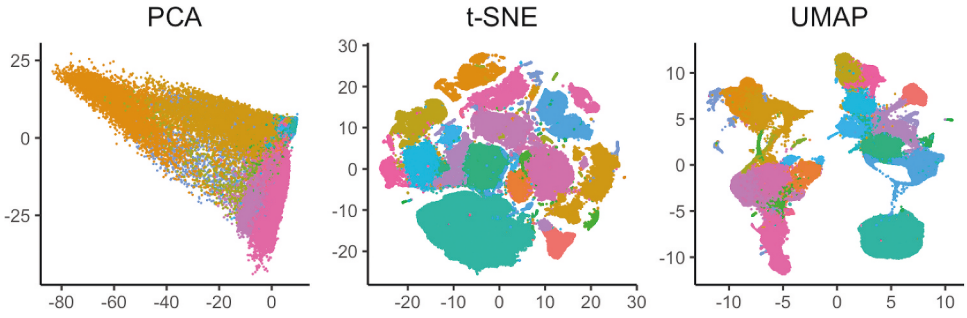

Figure 7. PCA, t-SNE, and UMAP visualizations with individual cells pooled across 24 individuals color-coded by their cell subsets membership. The same color indicates the same cell subset across three visualizations.

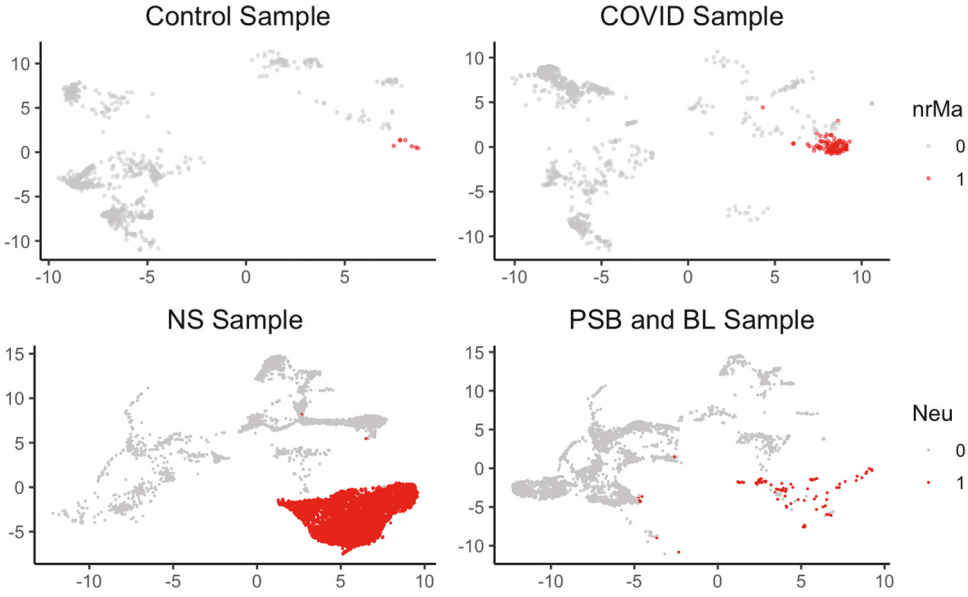

Figure 8. Top row: UMAP visualizations of cells obtained from a COVID-negative individual (control sample) and those from a COVID-positive patient (COVID sample). For ease of comparison, only nrMA cell subset is colored in red, and the remaining cells are in gray. Bottom row: UMAP visualizations of cells obtained from the NS sample of a selected individual and PSB and BL sample of the same individual. Only Neu cell subset is colored in red for ease of comparison.

(1) Logistic regression with lasso penalty using the proportions of the identified 22 cell subsets as features and the individual-level meta-data as covariates.
(2) Logistic regression using the mean expression levels of nine selected genes as features and the individual-level meta-data as covariates.
(3) Random forest with mean expression levels of the nine genes and the individual-level covariates as input features. The number of trees is set to 100 to ensure that every input sample is predicted at least a few times.
(4) KNN classification based on the mean expression levels of the nine genes and the individual-level covariates. The number of nearest neighbors K is 5, chosen according to the empirical rule of using the square root of the sample size.
(5) A DNN model: CyTOF DL combined with Adam optimization algorithm along with the learning rate of 0.0001.
(6) Distributional data classification based on the GMMs estimated from the UMAP representation derived from the gene expression levels of the cells of each individual.

For methods 1–4, we perform analysis in R, for method 5, we use Python, and for method 6, we use Matlab. Different software platforms are used based on the availability of the codes.

More specifically, when using the proportions of cell subsets as features, we transform the features before fitting the logistic regression model. This pre-processing is needed because the resulting features form the so-called compositional data, i.e., the data lie on a simplex since the sum of the proportions is 1. A common technique to handle compositional data is to apply the *centered log-ratio* (clr) transform[105] to map the feature vectors into an "unbounded" space. The transform clr uses the geometric mean of all features in one vector as a reference value and then takes the log of the ratio between each feature and the geometric mean. Because the sample size 24 is very small relative to the feature dimension of 22, we incorporate the lasso penalty into the logistic regression to prevent overfitting.

On the other hand, Chua et al.[104] suggested some sets of genes that would be affected by the infection of COVID. For example, preferential ACE2 protein localization on motile cilia has been confirmed to be highly related with a strong infectivity of ciliated cells by SARS-CoV-2 in vitro. Also, the nrMa cell subset that showed a highly inflammatory profile characterized

by the expression of the chemokine encoding genes CCL2 (encoding MCP1), CCL3 (encoding MIP1α), CCL20, CXCL1 and CXCL3 and the pro-inflammatory cytokines IL1B, IL8, IL18, and TNF, was detected to be highly expressed in patients with critical COVID-19. Therefore, we select these genes a priori and obtain their mean expression levels for each sample (individual) as the representative features to predict the COVID infection outcome. Moreover, we also utilize CyTOF DL. This method requires an equal number of cells in all samples as input. We thus randomly sample cells with replacement so that each sample dataset has 16,000 cells. Figure 9 (Left) shows the ROC curves obtained by methods 1–5. The DNN-based method performs poorly because of the small sample size (the number of individuals). In addition, the requirement of an equal number of cells in each sample may lead to distortion in the single-cell data because the number of cells from COVID-negative individuals ranges from a few hundred to a thousand cells, much smaller than 16,000. The random forest model and LASSO model perform similarly well by utilizing gene (marker) expression information and cell type proportions, respectively.

After performing feature selection, age is identified as a significant variable in both LASSO and random forest models. Figure 10 shows the important features selected by LASSO (Top) and the random forest model (Bottom). In the LASSO logistic regression model, λ is the complexity parameter used in the L1 norm penalty term, which controls the bias-variance trade-off and results in the selection of important variables. When λ approaches zero, the solution of LASSO will approach the ordinary least square (OLS) estimator in a generalized linear model. On the other hand, if λ becomes larger and larger, all the regression coefficients will shrink to zero eventually. More important variables will have coefficients that approach zero slower. In Figure 10 (Top), a trace plot is provided to show the change of the regression coefficient associated with each variable at the increase of λ (x-axis). The most important variables will have coefficients that become zero the latest. The trace plot shows that the proportions of Epithelial cell subtypes (*e.g.*, basal, secretory, ionocyte, and squamous cells) and immune cell subtypes (*e.g.*, mast cells) are the most significant variables to distinguish COVID and healthy individuals. On the other hand, in terms of mean expression levels, Model 2 selects CXCL1 and CXCL3 as important variables at the significance level of 0.05. These two genes are also ranked highly according to both accuracy and Gini index in the random forest model. In addition, CCL3 is considered to be the most important feature by both measures in the random forest model.

We also experiment with using the pseudo-mixture model to estimate the posterior probabilities of each individual's infection status. Again, LOOCV is used to evaluate the performance with AUCs shown in Figure 9 (Right). Specifically, we first use a GMM to summarize the clustering result of each individual, i.e., the mean vector and covariance matrix for each Gaussian component is represented by the cluster-specific sample mean and sample covariance, respectively. The prior probability of every component is set to be the proportion of cells in the corresponding cluster. Here we do not assume the cell type of any component in the GMM of an individual is known. In other words, the clusters found in the individuals

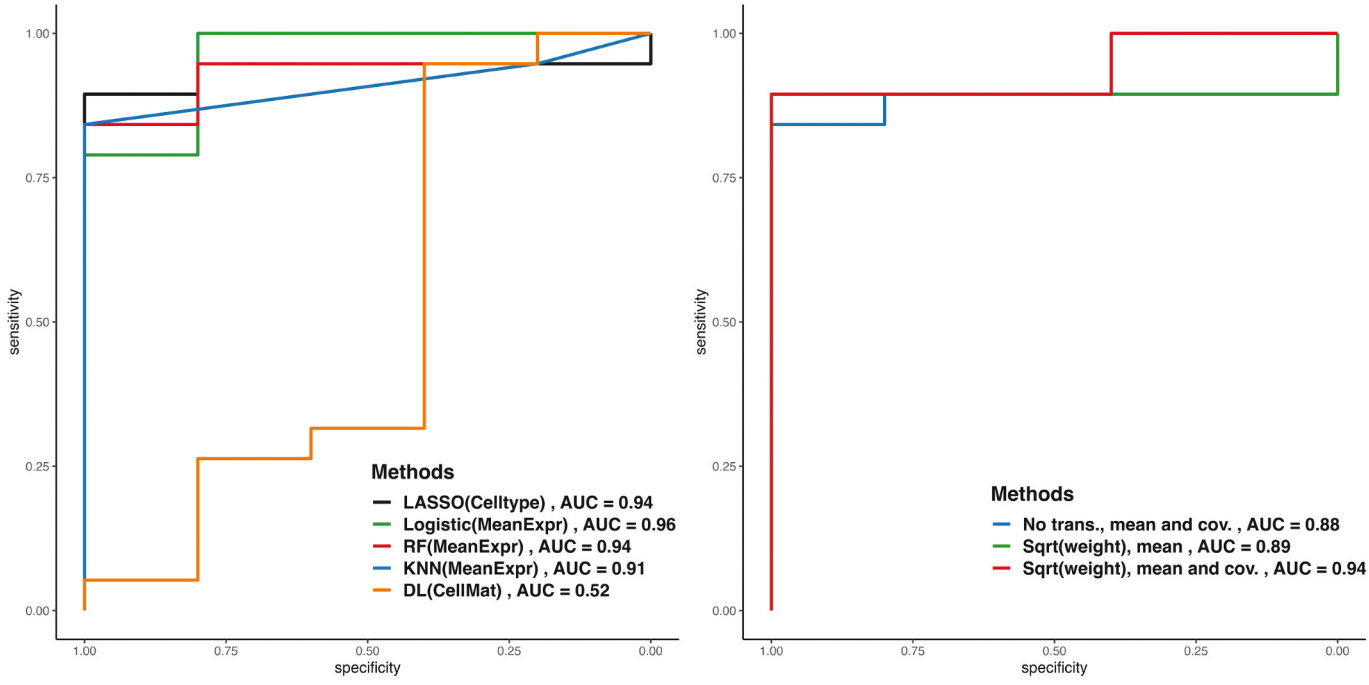

Figure 9. ROC curves. Left: Comparison of prediction performance among logistic regression with Lasso penalty using cell subsets proportions (Lasso(celltype), logistic regression, random forest and KNN with mean gene expressions (Logistic(meanecpr), RF(MeanExpr), KNN(MeanExpr)), and DNN method using the entire single-cell data (Dl(cellmat) in terms of LOOCV AUC. Right: Pseudo-mixture model for prediction based on MAW distances between GMMs. The cell types of the Gaussian components in the GMMs are assumed unknown. Three schemes are used. In the scheme indicated by the blue line, the Gaussian component weights are given by the proportion of data in the component and both the component mean and covariance matrix are considered when computing MAW. In the other two schemes, to emphasize rare components, a square-root transform (followed by normalization) is applied to the weights. The last two schemes differ by whether covariance matrices are considered when computing MAW.

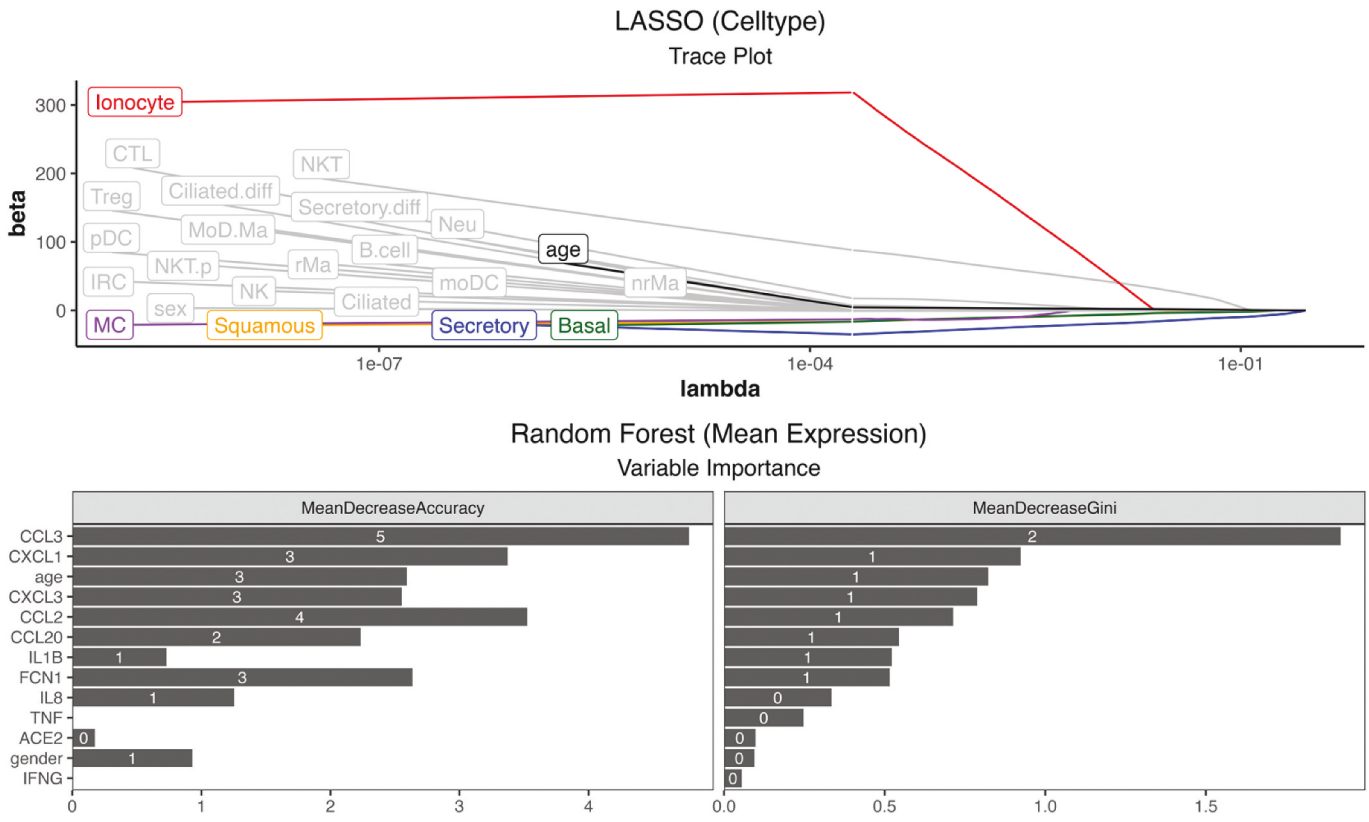

Figure 10. Top: The trace plot is generated from the LASSO logistic regression model with cell type proportions as covariates. The plot shows that ionocyte, basal, secretory, squamous and mast (MC) cells are selected as significant cell type markers. Bottom: Variable importance are ranked according to two measures: mean decrease accuracy and mean decrease in Gini index. It shows that CCL3 has the largest mean decrease in both accuracy and Gini coefficient. Thus, it is considered as the most important variable, and potentially a biomarker indicating COVID infection.

are not subject to a unified taxonomy of cell types (aka, not aligned). Technical details about this method are referred to Qiao and Li.[91] We test three schemes. In the basic scheme, we compute the pairwise MAW distance between the GMMs without altering the models. Considering that rare clusters/components are often highly informative, but their importance tends to be undermined in the calculation of MAW due to the low component weights, we adjust the GMMs by using the normalized square roots of the component weights as the new weights and then compute the MAW (the second scheme). Note that applying the square root transform will increase weights at the lower end. We find that with this adjustment, the AUC improved from 0.88 to 0.94. It is also interesting to investigate whether the shape of each Gaussian component, which is captured by the covariance matrix, matters. We thus compute the MAW distance for the GMMs without considering the covariance matrix or, equivalently, by assuming the covariance matrix is shared across the components. In this case, the MAW distance between GMMs is reduced to the Wasserstein metric between the discrete distributions over the component means. Again, the square root transform is applied to the component weights. The AUC obtained is 0.89, which is notably lower than that achieved with shape information. This result shows that useful information is lost if we represent each component using only its mean vector. The distribution of data in a component is valuable modulo the effect of its mean.

Data integration across three sites

We perform data integration across samples obtained from three sites: NS, PSB, and BL from two COVID-positive patients. Chua et al.[104] used Seurat to perform integration and clustered cells based on 3,000 genes that are identified as highly variable across the three sites. The authors then treated the clustering result as the ground truth and manually annotated the clusters.

We further perform integration using the MAW barycenter approach[102] and LIGER to integrate clustering results across single cells from the three sites within each patient. For the MAW approach, we construct a GMM in the UMAP latent space based on the clustering result of the sample from each site. Each cluster is treated as one Gaussian component in the GMM with its component mean and covariance matrix estimated from cells in this cluster. The proportion of cells in each cluster is taken as the prior probability of the corresponding Gaussian component. Then the MAW barycenter of the three GMMs is computed for each patient, yielding a "consensus" distribution across three sites. The MAW barycenter is then used to recluster all the cells from each patient. Therefore, the resulting cluster labels are naturally aligned across different samples. For LIGER integration, we use R package LIGER. We first normalize and scale all the count matrices to account for differences in sequencing depth and efficiency between cells, processed by the built-in functions normalize and scaleNotCenter. Since the clustering performance of LIGER depends strongly on the tuning parameter "resolution," we manually set it to match the number of clusters provided by

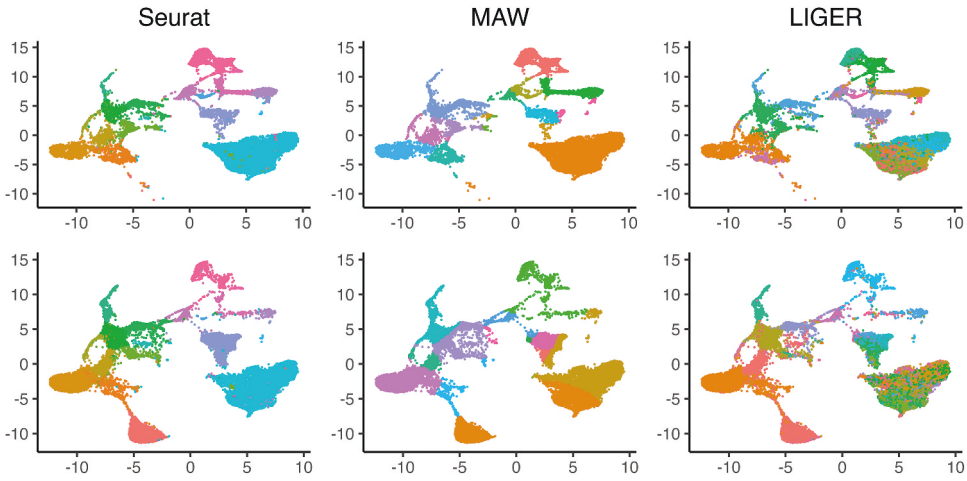

Figure 11. UMAP visualizations of Seurat, MAW and LIGER for integrating multi-source scRNA-seq data from three different testing sites. Each row corresponds to a particular COVID-positive individual.

Seurat. The comparison of results obtained by MAW and LIGER for two patients are shown in Figure 11 (one row for each patient). The ground truth of cell types, referred to as "Seurat," is visualized in the first column of Figure 11. Based on the plots in the figure, MAW yields clusters more similar to those in "Seurat." We also numerically assess the similarity of clustering results to the ground truth by computing the adjusted Rand index (ARI) and Meila's variation of information (VI). A higher value of ARI and a lower value of VI correspond to a higher level of agreement with the ground truth. For MAW, the ARIs of the two patients are (0.93, 0.57), and the VIs are (0.84, 2.29). For LIGER, the ARIs are (0.57, 0.45), and the VIs are (1.82, 2.31).

Discussion

Our study utilized various methods to analyze a scRNA-seq dataset obtained from Chua et al.,[104] with the aim of identifying key cell populations and genes associated with COVID-19 infection in the nasopharyngeal area. Our findings indicated that the nonresident macrophages (nrMa) were largely represented in COVID-positive individuals. Furthermore, we observed significant differences in cell type distribution across samples collected from three different sites, with neutrophils (Neu) being the most dominant cell type in the NS sample, but seldom appearing in samples collected from PSB and BL. In addition, we identified the chemokine encoding gene CCL3, encoding MIP1α, as the most predictive biomarker for the infection outcome selected by both important measures. This finding may be biologically explained that nonresident macrophages with overexpression of pro-inflammatory mediators like CCL3 could lead to an increase in monocyte recruitment and differentiation while finally resulting in a critical inflammation response. Notably, we found age-related differences in the immune response, with the elderly showing a stronger immune response compared to the youth. Our results were largely consistent with previous study. However, our study provided additional insights into the immune response to COVID-19, especially with regard to the role of nonresident macrophages and age-related differences in the immune response. While our study provides valuable insights, it also has some limitations. Firstly, the sample size is relatively small, which may limit the generalizability of our findings. Secondly, we focused on the nasopharyngeal area, and our findings may not apply to other infection sites. Future studies with larger sample sizes and including data from multiple infection sites will be needed to confirm and expand upon our findings.

Although we applied both clustering and classification methods to study the COVID-19 patients, the methods are generally applicable to other problems in the areas such as vaccines and immunotherapy. In summary, we have handled two fundamental types of data: tabular data and distributional data. The tabular data appear as more traditional data, like clinical data. The recently emerged single-cell data belongs to the distributional data type. In future work, there are potential extensions in methodology development. For example, there is ample room to improve existing methods so that the advantages of approaches developed separately for tabular and distributional data can be combined. Specifically, approaches for tabular data can exploit the algebraic structure of the data, but when the raw data are distributional, certain pre-processing must be carried out to convert them into tabular representations. What is the best way to convert distributional data into tabular data remains an open question. On the other hand, when we treat distributional data, we can take into consideration the information missing from the tabular data. However, as these approaches rely on the distance between the distributions, they are not as flexible as matrix-based operations applied to tabular data. One possible direction is to encode the information captured by the comparison of the distributions as tabular features.

In conclusion, immunoprofiling is a rapidly growing field with enormous potential for improving our understanding of the immune system. The statistical and machine learning methods reviewed in this paper demonstrate the power and utility of analyzing complex immunological data by cutting-edge technologies. The methods discussed, including flow cytometry, mass cytometry, scRNA-seq, and various algorithms, have enabled researchers to explore with unprecedented effectiveness the diversity and dynamics of immune cell populations and their interactions. Additionally, the application of machine learning

techniques has facilitated the development of predictive models for patient outcomes and treatment responses, further advancing precision medicine approaches. Despite these successes, challenges remain, including standardization of protocols, interpretation of complex data, and the need for robust validation and replication studies. Nevertheless, statistical and machine learning methods hold a great promise for further advancing our understanding of the immune system and for improving the prediction of patient outcomes in a clinical setting.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research of J. Zhang and L. Lin is supported by the Duke University Center for AIDS Research (CFAR), an NIH funded program (5P30 AI064518). J. Li's research is supported by NSF DMS-2013905.

ORCID

Lin Lin http://orcid.org/0000-0002-7464-1172

References

1. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, Premsri N, Namwat C, de Souza M, Adams E, et al. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med. 2009 Dec 3;361(23):2209–20. doi:10.1056/NEJMoa0908492.
2. Alter G, Barouch D. Immune correlate-guided HIV vaccine design. Cell Host Microbe. 2018;24(1):25–33. doi:10.1016/j.chom.2018.06.012.
3. Lopez Angel CJ, Tomaras GD, Spindler KR. Bringing the path toward an HIV-1 vaccine into focus. PLoS Pathog. 2020;16(9):e1008663. doi:10.1371/journal.ppat.1008663.
4. Pantaleo G, Koup RA. Correlates of immune protection in HIV-1 infection: what we know, what we don't know, what we should know. Nat Med. 2004 Aug 01;10(8):806–10. doi:10.1038/nm0804-806.
5. Lin L, Finak G, Ushey K, Seshadri C, Hawn TR, Frahm N, Scriba TJ, Mahomed H, Hanekom W, Bart P-A, et al. COMPASS identifies T-cell subsets correlated with clinical outcomes. Nat Biotechnol. 2015 June;33(6):610–6. doi:10.1038/nbt.3187.
6. Waldman AD, Fritz JM, Lenardo MJ. A guide to cancer immunotherapy: from T cell basic science to clinical practice. Nat Rev Immunol. 2020 Nov;20(11):651–68. doi:10.1038/s41577-020-0306-5.
7. Farkona S, Diamandis EP, Blasutig IM. Cancer immunotherapy: the beginning of the end of cancer? BMC Med. 2016 May 5;14(1):73. doi:10.1186/s12916-016-0623-5.
8. Chau I, Doki Y, Ajani JA, Xu J, Wyrwicz L, Motoyama S, Ogata T, Kawakami H, Hsu C-H, Adenis A, et al. Nivolumab (NIVO) plus ipilimumab (IPI) or NIVO plus chemotherapy (chemo) versus chemo as first-line (1L) treatment for advanced esophageal squamous cell carcinoma (ESCC): first results of the CheckMate 648 study. J Clin Oncol. 2021;39(18_suppl):LBA4001–LBA4001. doi:10.1200/JCO.2021.39.15_suppl.LBA4001.
9. Janjigian YY, Shitara K, Moehler M, Garrido M, Salman P, Shen L, Wyrwicz L, Yamaguchi K, Skoczylas T, Bragagnoli AC, et al. First-line nivolumab plus chemotherapy versus chemotherapy alone for advanced gastric, gastro-oesophageal junction, and oesophageal adenocarcinoma (CheckMate 649): a randomised, open-label, phase 3 trial. Lancet. 2021 July 3;398(10294):27–40. doi:10.1016/S0140-6736(21)00797-2.
10. Patel MA, Kratz JD, Lubner SJ, Loconte NK, Uboha NV. Esophagogastric cancers: integrating immunotherapy therapy into current practice. J Clin Oncol. 2022 Aug 20;40(24):2751–62. doi:10.1200/JCO.21.02500.
11. Chaft JE, Oezkan F, Kris MG, Bunn PA, Wistuba II, Kwiatkowski DJ, Owen DH, Tang Y, Johnson BE, Lee JM, et al. Neoadjuvant atezolizumab for resectable non-small cell lung cancer: an open-label, single-arm phase II trial. Nat Med. 2022 Oct;28(10):2155–61. doi:10.1038/s41591-022-01962-5.
12. Saeys Y, Van Gassen S, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nat Rev Immunol. 2016 July 01;16(7):449–62. doi:10.1038/nri.2016.56.
13. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell. 2015 May 21;58(4):610–20. doi:10.1016/j.molcel.2015.04.005.
14. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet. 2013 Sept;14(9):618–30. doi:10.1038/nrg3542.
15. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015 June;12(6):519–22. doi:10.1038/nmeth.3370.
16. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. Nat Biotechnol. 2015 Mar;33(3):285–9. doi:10.1038/nbt.3129.
17. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood SA, Ponting CP, Voet T, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods. 2016 Mar;13(3):229–32. doi:10.1038/nmeth.3728.
18. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017 Sept;14(9):865–8. doi:10.1038/nmeth.4380.
19. Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, Alda-Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun. 2018 Feb 22;9(1):781. doi:10.1038/s41467-018-03149-4.
20. White S, Quinn J, Enzor J, Staats J, Mosier SM, Almarode J, Denny TN, Weinhold KJ, Ferrari G, Chan C, et al. FlowKit: a python toolkit for integrated manual and automated cytometry analysis workflows. Front Immunol. 2021;12:768541. doi:10.3389/fimmu.2021.768541.
21. Crowell H, Chevrier S, Jacobs A, Sivapatham S, Bodenmiller B, Robinson MD. An R-based reproducible and user-friendly preprocessing pipeline for CyTOF data [version 2; peer review: 2 approved]. F1000Res. 2022;9(1263):1263. doi:10.12688/f1000research.26073.2.
22. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019 June 19;15(6):e8746. doi:10.15252/msb.20188746.
23. 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 June;36(5):421–7. doi:10.1038/nbt.4091.
24. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018 Jan 18;9(1):284. doi:10.1038/s41467-017-02554-5.
25. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell. 2019 June 13;177(7):1873–87.e17. doi:10.1016/j.cell.2019.05.006.
26. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh P-R, Raychaudhuri S, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019 Dec;16(12):1289–96. doi:10.1038/s41592-019-0619-0.

27. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. Cell. 2019 June 13;177(7):1888–902.e21. doi:10.1016/j.cell.2019.05.031.
28. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nat Methods. 2019 Aug;16(8):715–21. doi:10.1038/s41592-019-0494-8.
29. Schuyler RP, Jackson C, Garcia-Perez JE, Baxter RM, Ogolla S, Rochford R, Ghosh D, Rudra P, Hsieh EW. Minimizing batch effects in mass cytometry data. Front Immunol. 2019;10:2367. doi:10.3389/fimmu.2019.02367.
30. Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, Montgomery R, Kluger Y. Removal of batch effects using distribution-matching residual networks. Bioinformatics. 2017 Aug 15;33(16):2539–46. doi:10.1093/bioinformatics/btx196.
31. Van Gassen S, Gaudilliere B, Angst MS, Saeys Y, Aghaeepour N. CytoNorm: a normalization algorithm for cytometry data. Cytometry A. 2020 Mar;97(3):268–78. doi:10.1002/cyto.a.23904.
32. Rebhahn JA, Quataert SA, Sharma G, Mosmann TR. SwiftReg cluster registration automatically reduces flow cytometry data variability including batch effects. Commun Biol. 2020 May 7;3(1):218. doi:10.1038/s42003-020-0938-9.
33. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, Chen J. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020 Jan 16;21(1):12. doi:10.1186/s13059-019-1850-9.
34. Xiang R, Wang W, Yang L, Wang S, Xu C, Chen X. A comparison for dimensionality reduction methods of single-cell RNA-seq data. Front Genet. 2021;12:646936. doi:10.3389/fgene.2021.646936.
35. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(11):2579–2605.
36. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw. 2018;3(29):861. doi:10.21105/joss.00861.
37. Wang Y, Huang H, Rudin C, Shaposhnik Y. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. J Mach Learn Res. 2021;22(1):9129–201.
38. Amid E, Warmuth MK. TriMap: large-scale dimensionality reduction using triplets. arXiv e-prints. 2019:arXiv:1910.00204.
39. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016. https://mitpress.mit.edu/9780262035613/deep-learning/.
40. Deng Y, Bao F, Dai Q, Wu LF, Altschuler SJ. Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. Nat Methods. 2019 Apr;16(4):311–4. doi:10.1038/s41592-019-0353-7.
41. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018 Dec;15(12):1053–8. doi:10.1038/s41592-018-0229-2.
42. Tran D, Nguyen H, Tran B, La Vecchia C, Luu HN, Nguyen T. Fast and precise single-cell data analysis using a hierarchical autoencoder. Nat Commun. 2021 Feb 15;12(1):1029. doi:10.1038/s41467-021-21312-2.
43. Brendel M, Su C, Bai Z, Zhang H, Elemento O, Wang F. Application of deep learning on single-cell RNA sequencing data analysis: a review. Genom Proteom Bioinform. 2022 Oct;20(5):814–35. doi:10.1016/j.gpb.2022.11.011.
44. Weber LM, Robinson MD. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytometry A. 2016 Dec;89(12):1084–96. doi:10.1002/cyto.a.23030.
45. Petegrosso R, Li Z, Kuang R. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. Brief Bioinform. 2020 July 15;21(4):1209–23. doi:10.1093/bib/bbz063.
46. Krzak M, Raykov Y, Boukouvalas A, Cutillo L, Angelini C. Benchmark and parameter sensitivity analysis of single-cell RNA sequencing clustering methods. Front Genet. 2019;10:1253. doi:10.3389/fgene.2019.01253.
47. Liu X, Song W, Wong BY, Zhang T, Yu S, Lin GN, Ding X. A comparison framework and guideline of clustering methods for mass cytometry data. Genome Biol. 2019 Dec 23;20(1):297. doi:10.1186/s13059-019-1917-7.
48. Yu L, Cao Y, Yang JYH, Yang P. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. Genome Biol. 2022 Feb 8;23(1):49. doi:10.1186/s13059-022-02622-0.
49. Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory. 1982;28(2):129–37. doi:10.1109/TIT.1982.1056489.
50. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017 May;14(5):483–6. doi:10.1038/nmeth.4236.
51. Grun D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015 Sept 10;525(7568):251–5. doi:10.1038/nature14966.
52. Ge Y, Sealfon SC. flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. Bioinformatics. 2012 Aug 1;28(15):2052–8. doi:10.1093/bioinformatics/bts300.
53. Qian Y, Wei C, Eun-Hyung Lee F, Campbell J, Halliley J, Lee JA, Cai J, Kong YM, Sadat E, Thomson E, et al. Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. Cytometry B Clin Cytom. 2010;78(Suppl 1):S69–82. doi:10.1002/cyto.b.20554.
54. Bicego M. DisRFC: a dissimilarity-based random forest clustering approach. Pattern Recogn. 2023;133(C):9. doi:10.1016/j.patcog.2022.109036.
55. Zurauskiene J, Yau C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. BMC Bioinform. 2016 Mar 22;17(1):140. doi:10.1186/s12859-016-0984-y.
56. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. PLoS Comput Biol. 2015 Nov;11(11):e1004575. doi:10.1371/journal.pcbi.1004575.
57. Lin P, Troup M, Ho JWK. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017 Mar 28;18(1):59. doi:10.1186/s13059-017-1188-0.
58. Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat Theory Methods. 1974;3(1):1–27. doi:10.1080/03610927408827101.
59. Fang X, Ho JWK, Vitek O. FlowGrid enables fast clustering of very large single-cell RNA-seq data. Bioinformatics. 2021 Dec 22;38(1):282–3. doi:10.1093/bioinformatics/btab521.
60. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008 Oct 09;2008(10):P10008. doi:10.1088/1742-5468/2008/10/P10008.
61. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015 May;33(5):495–502. doi:10.1038/nbt.3192.
62. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018 Feb 6;19(1):15. doi:10.1186/s13059-017-1382-0.
63. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014 Apr;32(4):381–6. doi:10.1038/nbt.2859.
64. Stassen SV, Siu DMD, Lee KCM, Ho JWK, So HKH, Tsia KK. PARC: ultrafast and accurate clustering of phenotypic data of millions of single cells. Bioinformatics. 2020 May 1;36(9):2778–86. doi:10.1093/bioinformatics/btaa042.
65. Hennig C. Methods for merging Gaussian mixture components. Adv Data Anal Classif. 2010 Apr 01;4(1):3–34. doi:10.1007/s11634-010-0058-3.
66. Li J. Clustering based on a multilayer mixture model. J Comput Graph Stat. 2005 Sept 01;14(3):547–68. doi:10.1198/106186005X59586.

67. Pyne S, Hu X, Wang K, Rossin E, Lin T-I, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, et al. Automated high-dimensional flow cytometric data analysis. Proc Natl Acad Sci U S A. 2009 May 26;106(21):8519–24. doi:10.1073/pnas.0903028106.
68. Finak G, Bashashati A, Brinkman R, Gottardo R. Merging mixture components for cell population identification in flow cytometry. Adv Bioinformatics. 2009;2009:1–12. doi:10.1155/2009/247646.
69. Melnykov V. Merging mixture components for clustering through pairwise overlap. J Comput Graph Stat. 2016;25(1):66–90. doi:10.1080/10618600.2014.978007.
70. Lin L, Chan C, West M. Discriminative variable subsets in Bayesian classification with mixture models, with application in flow cytometry studies. Biostatistics. 2015;17(1):40–53. doi:10.1093/biostatistics/kxv021.
71. Boedigheimer MJ, Ferbas J. Mixture modeling approach to flow cytometry data. Cytometry A. 2008 May;73(5):421–9. doi:10.1002/cyto.a.20553.
72. Chan C, Feng F, Ottinger J, Foster D, West M, Kepler TB. Statistical mixture modeling for cell subtype identification in flow cytometry. Cytometry A. 2008 Aug;73(8):693–701. doi:10.1002/cyto.a.20583.
73. Lo K, Brinkman RR, Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. Cytometry A. 2008 Apr;73(4):321–32. doi:10.1002/cyto.a.20531.
74. Lin L, Li J. Clustering with hidden Markov model on variable blocks. J Mach Learn Res. 2017;18:3913–61.
75. Lee H, Li J. Variable selection for clustering by separability based on ridgelines. J Comput Graph Stat. 2012 Apr 01;21(2):315–36. doi:10.1080/10618600.2012.679226.
76. Seo B, Lin L, Li J. Block-wise variable selection for clustering via latent states of mixture models. J Comput Graph Stat. 2022 Jan 02;31(1):138–50. doi:10.1080/10618600.2021.1982724.
77. Xie K, Huang Y, Zeng F, Liu Z, Chen T. scAIDE: clustering of large-scale single-cell RNA-seq data reveals putative and rare cell types. NAR Genom Bioinform. 2020 Dec;2(4):lqaa082. doi:10.1093/nargab/lqaa082.
78. Chen Y, Ye J, Li J. Aggregated Wasserstein distance and state registration for hidden Markov models. IEEE Trans Pattern Anal Mach Intell. 2020 Sept;42(9):2133–47. doi:10.1109/TPAMI.2019.2908635.
79. Amodio M, van Dijk D, Srinivasan K, Chen WS, Mohsen H, Moon KR, Campbell A, Zhao Y, Wang X, Venkataswamy M, et al. Exploring single-cell data with deep multitasking neural networks. Nat Methods. 2019 Nov;16(11):1139–45. doi:10.1038/s41592-019-0576-7.
80. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. Bioinformatics. 2005 Aug 1;21(15):3201–12. doi:10.1093/bioinformatics/bti517.
81. Li J, Seo B, Lin L. Optimal transport, mean partition, and uncertainty assessment in cluster analysis. Stat Anal Data Min. 2019 May 14;12(5):359–77. doi:10.1002/sam.11418.
82. Zhang L, Lin L, Li J, Wren J. CPS analysis: self-contained validation of biomedical data clustering. Bioinformatics. 2020 June 1;36(11):3516–21. doi:10.1093/bioinformatics/btaa165.
83. Breiman L. Random forests. Mach Learn. 2001 Oct 01;45(1):5–32. doi:10.1023/A:1010933404324.
84. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21–7. doi:10.1109/TIT.1967.1053964.
85. Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, Tárraga J, Pascual-Montano A, Nogales-Cadenas R, Santoyo J, et al. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res. 2010 July;38(Web Server issue):W210–3. doi:10.1093/nar/gkq388.
86. Hu P, Zhang W, Xin H, Deng G. Single cell isolation and analysis. Front Cell Dev Biol. 2016;4:116. doi:10.3389/fcell.2016.00116.
87. Nowicka M, Krieg C, Crowell HL, Hartmann FJ, Guglietta S, Becher B, Levesque MP, Robinson MD. CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. F1000Res. 2017;6:748. doi:10.12688/f1000research.11622.1.
88. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics. 2007 Jan 25;8:25. doi:10.1186/1471-2105-8-25.
89. Hu Z, Tang A, Singh J, Bhattacharya S, Butte AJ. A robust and interpretable end-to-end deep learning model for cytometry data. Proc Natl Acad Sci U S A. 2020 Sept 1;117(35):21373–80. doi:10.1073/pnas.2003026117.
90. Yi H, Stanley N. CytoSet: predicting clinical outcomes via set-modeling of cytometry data. Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. Gainesville (FL): Association for Computing Machinery; 2021. p. Article 47.
91. Qiao M, Li J. Distance-based mixture modeling for classification via hypothetical local mapping. Stat Anal Data Min. 2016;9(1):43–57. doi:10.1002/sam.11285.
92. Lee J, Hyeon DY, Hwang D. Single-cell multiomics: technologies and data analysis methods. Exp Mol Med. 2020 Sept;52(9):1428–42. doi:10.1038/s12276-020-0420-2.
93. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 June 24;184(13):3573–87 e29. doi:10.1016/j.cell.2021.04.048.
94. Kim HJ, Lin Y, Geddes TA, Yang JYH, Yang P. CiteFuse enables multi-modal analysis of CITE-seq data. Bioinformatics. 2020 Aug 15;36(14):4137–43. doi:10.1093/bioinformatics/btaa282.
95. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018 June 20;14(6):e8124. doi:10.15252/msb.20178124.
96. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020 May 11;21(1):111. doi:10.1186/s13059-020-02015-1.
97. Yuan M, Chen L, Deng M, Mathelier A. Clustering single-cell multi-omics data with MoClust. Bioinformatics. 2023 Jan 1;39(1). doi:10.1093/bioinformatics/btac736.
98. Gong B, Zhou Y, Purdom E. Cobolt: integrative analysis of multimodal single-cell sequencing data. Genome Biol. 2021 Dec 28;22(1):351. doi:10.1186/s13059-021-02556-z.
99. Minoura K, Abe K, Nam H, Nishikawa H, Shimamura T. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. Cell Rep Methods. 2021 Sept 27;1(5):100071. doi:10.1016/j.crmeth.2021.100071.
100. Zuo C, Chen L. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. Brief Bioinform. 2021 July 20;22(4). doi:10.1093/bib/bbaa287.
101. Johansen N, Quon G. scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. Genome Biol. 2019 Aug 14;20(1):166. doi:10.1186/s13059-019-1766-4.
102. Lin L, Shi W, Ye J, Li J. Multisource single-cell data integration by MAW barycenter for Gaussian mixture models. Biometrics. 2022 Feb 27;79(2):866–77. doi:10.1111/biom.13630.
103. Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, Welch JD. Jointly defining cell types from multiple single-cell datasets using LIGER. Nat Protoc. 2020 Nov;15(11):3632–62. doi:10.1038/s41596-020-0391-8.
104. Chua RL, Lukassen S, Trump S, Hennig BP, Wendisch D, Pott F, Debnath O, Thürmann L, Kurth F, Völker MT, et al. COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. Nat Biotechnol. 2020 Aug;38(8):970–9. doi:10.1038/s41587-020-0602-4.
105. Aitchison J. The statistical analysis of compositional data. J R Stat Soc Series B Stat Methodol. 1982;44(2):139–77. doi:10.1111/j.2517-6161.1982.tb01195.x.